INFORMATION TO USERS

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

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

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

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

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

**ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

Transmission of Multi-layered MPEG-4 Video over ATM/ABR Networks

by

Ramy M. Guirguis, B.Sc, M.Sc

A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements of the degree of

Doctor of Philosophy

Ottawa-Carleton Institute for Electrical Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University

February, 2001

© Copyright

2001, Ramy M. Guirguis

**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-60953-7

Canada

The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis

Transmission of Multi-Layered MPEG-4 Video Over ATM/ABR Networks

Submitted by Ramy M. Guirguis, M.Sc.
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Chair, Department of Systems and Computer Engineering

Dean S. Mahmoud, Thesis Supervisor

Dr. J. Yan, External Examiner

Carleton University
February, 2001

Abstract

Large-scale deployment of multimedia video applications is still being hindered by a major obstacle, which is the quality versus utilization efficiency trade-off. Currently, to accommodate the bursty nature of video traffic, good quality video is guaranteed by reserving more resources than actually required. This leads to a high cost for the end user. On the other hand, if less resources are reserved to achieve a higher resource utilization efficiency and lower cost for the end user, the result may be degradation of the quality of the transmitted video. The transmission of MPEG-4 video over Available Bit Rate (ABR) service provides an opportunity to solve the above noted problem through the interaction between the application and the network where both the application and the network can dynamically re-negotiate the required bandwidth.

The research presented in this thesis proposes a novel application and network-aware source architecture that is capable of bridging the gap between MPEG-4 application requirements and ATM/ABR network offerings. The proposed architecture combines: resource reservation and adaptation techniques, traffic shaping and rate shaping mechanisms, congestion control and congestion recovery mechanisms, and finally combines network and application awareness. A set of source algorithms are designed and presented in this research including: an Intelligent VOP Dropping algorithm (IVDA) which controls the transmission of enhanced layer Video Object Plans (VOPs), Quality Level Selection algorithm (QLSA), and a Rate Request Algorithm (RRA). The adaptive nature of the source architecture and algorithms alleviate the need for accurate traffic descriptors of the video traffic, which is a major problem when video is transmitted over resource reservation services such as Variable Bit Rate (VBR) service.

The proposed architecture and algorithms have been evaluated under various network conditions using an ATM/ABR network simulator and various video traces with different characteristics encoded into two layers according to the MPEG-4 standard.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Samy Mahmoud for his technical guidance and his continuous advice that saved me from reaching a dead end in several issues related to this research. I would also like to thank him for his continuous support and encouragement throughout the thesis.

I would like to thank Prof. John Lambadaris and Prof. Mike Devetsikiotis for the valuable technical discussions we shared, and for the strong technical background that I learned from the graduate courses that they taught.

The research in this thesis has greatly benefited from the technical discussions and industrial experience provided to me by Dr. Osama Aboul Magd and Dr. Sameh Rabie from Nortel Networks, Ottawa. I would like to express my deepest appreciation to both of them.

I would like to thank Mr. Simon Winder from Microsoft for his help in running and debugging the early releases of the MPEG-4 coder which was used to obtain the MPEG-4 traces used throughout this thesis.

This research was made possible through the financial support of the Telecommunication Research Institute of Ontario (TRIO) and the Natural Sciences and Engineering Research Council of Canada (NSERC) scholarship.

Last but not least, I would like to thank my wife Christiane and my young daughter Mariam, for their support and understanding and their sacrifices through the years of this research.

Table of Contents

Chapter 1: Introduction	1
1.1 Thesis Motivation	1
1.2 Problem Definition	3
1.3 Thesis Contributions	7
1.4 Thesis Outline	12
Chapter 2: MPEG-4 Standard Overview	14
2.1 History of MPEG standards	14
2.2 MPEG-4 Standard	15
2.3 MPEG-4 Delivery and QoS	18
2.3.1 Delivery Multimedia Integration Framework (DMIF)	18
2.3.2 MPEG-4 Quality of Service (QoS)	20
2.4 MPEG-4 Natural Video Coding	27
2.4.1 MPEG-4 Video Structure	29
2.4.2 Object-Based Encoding	32
2.4.3 MPEG-4 Scalability	36
2.5 MPEG-4 Video Decoding Process	38
2.6 Visual Profiles	42
2.7 Variable Bit Rate versus Constant Bit Rate Encoding	43
Chapter 3: MPEG-4 Video Transmission over ATM (Problem Definition)	45
3.1 Introduction	45
3.2 QoS Requirements for Video Applications	45

3.3 MPEG-4 Video Transmission over ATM Networks	48
3.3.1 Quality of Service Guarantee in ATM Networks	49
3.3.2 Video Transmission over ATM/CBR Service	53
3.3.3 Video Transmission over ATM/ VBR Service	55
3.4 MPEG-4 Video Transmission over ABR	58
3.5 Review of Previous Research on Video Transmission over ABR	62
3.5.1 AT&T Research Lab	62
3.5.2 Ohio State University	63
3.5.3 University of California at Irvine	64
3.5.4 Korea Advanced Institute of Science and Technology	65
3.6 Problem Formulation	66
3.6.1 Network Domain Problems	69
3.6.2 End-System Domain Problems	70
3.6.3 Adaptive Multi-layered MPEG-4 Encoder	74
Chapter 4: Performance Evaluation of the ATM/ABR Service	81
4.1 Introduction	81
4.2 ABR Engineering Considerations	84
4.2.1 Network Environment	84
4.2.2 ATM Switch Capabilities	86
4.2.3 Control-Loop Options	86
4.3 ABR Switch Design	89
4.4 ABR Performance Evaluation	94

4.4.1 Performance Metrics	94
4.4.2 Simulation Model	96
4.4.3 Simulation Results	98
4.5 ABR Network Delay	108
4.6 ABR Support for Video Applications	109
4.7 Conclusions	111
Chapter 5: Transmission of Uncontrolled Multi-layered MPEG-4 Video	113
5.1 Introduction	113
5.2 MPEG-4 Source Modeling	117
5.3 MPEG-4 Video Test Sequences	118
5.4 Transmission of Smoothed Uncontrolled MPEG-4 Video	130
5.4.1 Infinite Smoothing Buffer	133
5.4.2 Finite Smoothing Buffer	134
5.4.3 Setting of MCR	138
5.5 Conclusion	141
Chapter 6: Proposed Source Architectural Model	142
6.1 Introduction	142
6.2 Proposed Source Architectural Model	143
6.3 VOP Discard Algorithms	147
6.3.1 Design Objectives	147
6.3.2 Adaptive Early VOP Discard (AEVD)	149
6.3.3 Modified Adaptive Early VOP Discard (M-AEVD)	151

6.3.4 Intelligent VOP Discard Algorithm (IVDA)	153
6.3.5 Performance Study of the VOP Discard Algorithms	159
6.4 Quality Level Selection Algorithm (QLSA)	169
6.5 Rate Request Algorithm (RRA)	175
6.6 Adaptive Traffic Shaping Algorithm (ATSA)	181
6.7 On-line Measurements	182
6.7.1 Short-Cycle Measurements	182
6.7.2 Long-Cycle Measurements	183
6.7.3 Exponential Weighted Moving Average Measurements	184
6.8 Source State Diagram	187
6.8.1 Cell Transmit Opportunity (Cell Tx opp.)	187
6.8.2 Backward-RM Cell Arrival (B-RM Rx)	187
6.8.3 VOP Arrival (VOP Rx)	187
6.8.4 Short-Cycle Timer (Short Cycle)	188
6.8.5 Long-Cycle Timer (Long Cycle)	188
Chapter 7: Performance Evaluation	190
7.1 Introduction	190
7.2 Performance Metrics and Simulation Model	190
7.3 Simulation Results and Analysis	193
7.3.1 Performance of the Proposed Architecture & Algorithms	193
7.3.2 Behavior of Rate Request Algorithm	200
7.3.3 Impact of VOP Discard Algorithms	203

7.3.4 Impact of QLSA	207
7.3.5 Impact of Traffic Load	211
7.3.6 Impact of Background Traffic	217
7.3.7 Impact of Round Trip Delay	222
7.3.8 Impact of MCR Setting	222
7.4 Conclusions	223
Chapter 8: Conclusions and Future Work	224
8.1 Conclusions	224
8.2 Summary of Thesis Contributions	225
8.3 Suggestions for Future Research	229
References	230
Appendix A : MPEG-4 Encoder	241
A.1. Software Implementation of the MPEG-4 standard	241
A.2. MPEG-4 Controlling Parameters	244
Appendix B : ABR Service Parameters	249
B.1. RM Cell Structure	249
B.2. ABR Service Parameters	251

List of Figures

Figure 1-1: End-to-End QoS	5
Figure 2-1: Major Components of an MPEG-4 Terminal at the Receiver End	17
Figure 2-2: MPEG-4 System Layer Model [MPEG4-2606]	19
Figure 2-3: Transformation of the Media QoS [MPEG4-2606]	22
Figure 2-4: QoS Monitoring Modes [MPEG4-2606]	26
Figure 2-5: Classification of MPEG-4 Image and Video Coding Algorithms	28
Figure 2-6: Hierarchy of the Data Structure Classes in MPEG-4 Video [Ebrahimi97]	31
Figure 2-7: Basic Block Diagram of the MPEG-4 Video Encoder [Ebrahimi97]	35
Figure 2-8: Example of Sprite Coding of a Video Sequence[MPEG4-2459]	35
Figure 2-9: Simplified MPEG-4 Video Decoding Process [MPEG4-2202]	39
Figure 2-10: High Level Decoder Structure for Generalized Scalability	41
Figure 2-11: Distortion Level vs. Coding Rate	44
Figure 3-1: Cell Transfer Delay Probability Density Function	50
Figure 3-2: Video Transmission over CBR Service	54
Figure 3-3: Video Transmission over VBR Service	57
Figure 3-4: ABR Control-Loop	60
Figure 3-5: End-to-End QoS	68
Figure 3-6: Simple Source Model	71
Figure 3-7: Source Rate versus Service Rate	71
Figure 3-8: Adaptive Multi-layered Source Model	75
Figure 4-1: ABR Service Application	83

Figure 4-2: ABR Control-Loop Options	88
Figure 4-3: ERSA+ Switch Algorithm	90
Figure 4-4: Simulation Model	97
Figure 4-5: Simulation Results of ER Mode	102
Figure 4-6: Simulation Results of VS/VD Mode	103
Figure 4-7: Queue Size in Congested State	107
Figure 5-1: VOP size and SNR vs. Quantization Index (Q)	116
Figure 5-2: VOP Size and Occurrence Histogram for the Movie Sequence	124
Figure 5-3: VOP Size and Occurrence Histogram for the CNN Sequence	125
Figure 5-4: VOP Size and Occurrence Histogram for the Soccer Sequence	126
Figure 5-5: VOP Size and Occurrence Histogram for the BBC Sequence	127
Figure 5-6: SNR for the MPEG-4 Test Sequences	129
Figure 5-7: Simulation Model	132
Figure 5-8: $Q(t)$ and $D(t)$ at $r = 0.75$ for Infinite Smoothing Buffer	136
Figure 5-9: $D(t)$ and $CL(t)$ at $r = 0.8$ and $D_{max} = 150ms$	137
Figure 5-10: MCR vs. CLR and Average Delay (Max Delay=150ms)	140
Figure 6-1: Source Interface	146
Figure 6-2: Source Architectural Model	146
Figure 6-3: Buffer Model in AEVD	150
Figure 6-4: AEVD Algorithm	150
Figure 6-5: M-AEVD Algorithm	152
Figure 6-6: Intelligent VOP Dropping Algorithm	154
Figure 6-7: Buffer Behavior under IVDA	156

Figure 6-8: IVDA Algorithm	158
Figure 6-9: Comparison of VOP Discard Algorithms (movie trace)	164
Figure 6-10: Comparison of VOP Discard Algorithms (soccer trace)	165
Figure 6-11: Comparison of Various Dropping Techniques (BBC trace)	166
Figure 6-12: Comparison of VOP Discard Algorithms (CNN trace)	167
Figure 6-13: QLSA Interface	173
Figure 6-14: QLSA Algorithm	174
Figure 6-15: Requested Rate Function Options	178
Figure 6-16: Rate Request Algorithm (RRA) Interface	178
Figure 6-17: Rate Request Algorithm	180
Figure 6-18: Aging Property of EWMA	186
Figure 6-19: Source State Diagram of the source	189
Figure 7-1: Simulation Model	192
Figure 7-2: Video Quality of the Proposed Architecture	196
Figure 7-3: CDF of Source Delay	197
Figure 7-4: Data Source & Link1 Utilization	198
Figure 7-5: Behavior of RRA with an Uncongested Network	202
Figure 7-6: Video Quality of Proposed Architecture without IVDA	205
Figure 7-7: Video Quality of Proposed Architecture with M-AEVD	206
Figure 7-8: Buffer-Based Quality Level Control Mechanism	209
Figure 7-9: Video Quality of Proposed Architecture with Buffer-Based QL Control ..	210
Figure 7-10: Video Quality of Proposed Architecture with IVDA (Congested Network)	213

Figure 7-11: Video Quality of Proposed Architecture with M-AEVD	
(Congested Network)	215
Figure 7-12: Video Quality of Proposed Architecture with Buffer-Based Control Mechanism (Congested Network)	216
Figure 7-13: Video Quality of Proposed Architecture with Background Traffic	
(Soccer and BBC traces)	220
Figure 7-14: Video Quality of Proposed Architecture with Background Traffic	
(CNN and movie traces)	221
Figure A.1: Microsoft MPEG-4 Coder	243
Figure A.2: Microsoft MPEG-4 Decoder	243

List of Tables

Table 2-1: ISO/IEC 14496-6 (MPEG-4 DMIF) Defined QoS Metrics	23
Table 2-2: Temporal Scalability Types	37
Table 3-1: Traffic Descriptor & Qos Guarantee for CBR, rt-VBR and ABR Services ..	52
Table 4-1: ABR Parameters	96
Table 4-2: Summary of Results for the Steady State	99
Table 4-3: Summary of Results for the Transient State	101
Table 4-4: Summary of Results for the Congested State (RTD=200ms)	106
Table 5-1: Classification of the Test Sequences	119
Table 5-2: MPEG-4 Coding Parameters for the Test Sequences	120
Table 5-3: Movie Sequence Statistics	121
Table 5-4: CNN Sequence Statistics	122
Table 5-5: Soccer Sequence Statistics	122
Table 5-6: BBC Sequence Statistics	123
Table 5-7: Mean SNR for Various Test Sequences	128
Table 5-8: MCR vs.CLR and Davg for Dmax=150ms	139
Table 6-1: Performance Results at CLR=0 (movie trace)	168
Table 6-2: Performance Results at CLR=0 (soccer trace)	168
Table 6-3: Performance Results at CLR=0 (BBC trace)	168
Table 6-4: Performance Results at CLR=0 (CNN trace)	168
Table 6-5: Components Invoked with Various Events	188
Table A-1: MPEG-4 Codec Parameters	244

Table B-1: Fields and their Position in the RM-cell 250

Table B-2: ABR Service Parameters 251

List of Abbreviations

ABR	Available Bit Rate
ACR	Allowed Cell Rate
AR	Available Rate
ASIU	ABR Service Interface Unit
ATM	Asynchronous Transfer Mode
ATSA	Adaptive Traffic Shaping Algorithm
ATVR	Average Transmitted VOP Rate
AU	Access Unit
AVED	Adaptive Early VOP Discard
B-RM cell	Backward Resource Management cell
B-VOP	Bidirectionally-coded VOP
CAC	Call Admission Control
CBR	Constant Bit Rate
CCR	Current Cell Rate
CDV	Cell Delay Variation
CDVT	Cell Delay Variation Tolerance
CI	Congestion Indication
CLP	Cell Loss Priority
CLR	Cell Loss Ratio
CRT	Congestion Response Time

CTD	Cell Transfer Delay
DAI	DMIF Application Interface
dB	DeciBel
DCT	Discrete Cosine Transform
DMIF	Delivery Multimedia Integration Framework
DNI	DMIF Network Interface
EFCI	Explicit Forward Congestion Indication
ER	Explicit Rate
ERISM	Explicit Rate Indication Scheme for MPEG
ERSA	Explicit Rate Switch Algorithm
ES	Elementary Stream
EWMA	Exponential Weighted Moving Average
F-RM cell	Forward Resource Management Cell
GCRA	Generic Cell Rate Algorithm
GFR	Guarantee Frame Rate
GOP	Group of Picture
I-VOP	Intra-coded VOP
ICR	Initial Cell Rate
IVDA	Intelligent VOP Discard Algorithm
KAIST	Korea Advanced Institute of Science and Technology
LU	Link Utilization
M-AVED	Modified Adaptive Early Discard

MBS	Maximum Burst Size
MCR	Minimum Cell Rate
ME	Matching Efficiency
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MQS	Maximum Queue Size
NI	No Increase
Nrm	Maximum number of cells a source may send for each F-RM cell
nrt-VBR	non-real time Variable Bit Rate
P-VOP	Predictive-coded VOP
PCR	Peak Cell Rate
PMR	Peak to Mean Ratio
PSNR	Peak Signal-to-Noise Ratio
QL	Quality Level
QLSA	Quality Level Selection Algorithm
QoS	Quality of Service
RC	Rate Controller
RDF	Rate Decrease Factor
RED	Random Early Detection
RIF	Rate Increase Factor
RM cell	Resource Management cell
RRA	Rate Request Algorithm

rt-VBR	real time Variable Bit Rate
RTD	Round Trip Delay
SAVE	Smoothed Adaptive Video Over Explicit Rate Networks
SCR	Sustainable Cell Rate
SL	Synchronization Layer
SNR	Signal to Noise Ratio
SR	Source Rate
SVR	Source VOP Rate
TAU	Traffic Anticipation Unit
TES	Transform Expand Sample
TQoS	Transport QoS
TransMux	Transport Multiplexing
TRT	Transient Response Time
UBR	Unspecified Bit Rate
UNI	User Network interface
VBR	Variable Bit Rate
VC	Virtual Connection
VLBV	Very Low Bit Rate Video
VO	Video Object
VoD	Video on Demand
VOP	Video Object Plane
VPN	Virtual Private Network

VS	Video Session
VS/VD	Virtual Source / Virtual Destination

List of Symbols

α	Smoothing factor for calculating the EWMA of cell interarrival time in ERSA+
ait	Average cell internal arrival time
α_{ME}	Smoothing factor used in calculating EWMA of matching efficiency
α_{SR}	Smoothing factor used in calculating EWMA of source rate
β	Source buffer size
B_j	Buffer occupancy just before arrival of j^{th} VOP
Burst(τ)	Maximum amount of traffic that can be sent by the source during any interval τ (Burstiness constraint)
c	A multiplier that indicates the speed of convergence of ER algorithm
C_{ij}	Control signal for the i^{th} quality level and the j^{th} VOP
CL(t)	Cell loss at time t
D(t)	Source buffer delay at time t
D_{avg}	Average delay in the source buffer
$d_i(n)$	Luminance value of the i^{th} pixel of the reconstructed n^{th} VOP
D_j	Source buffer delay for the j^{th} VOP
D_{max}	Maximum delay at the source buffer
$D_{network}$	Network delay
Δt	Time between the arrival of ATM cells
k	Aggressiveness parameter of RRA
l	AVED threshold

λ_j	Encoder rate for the j^{th} VOP
L_j	Flag for the j^{th} VOP type (1 for basic layer, 0 for enhanced layer)
M	Total number of B-RM cells
μ	Source buffer average service rate
μ'_j	Source buffer average Service rate for j^{th} VOP
μ_i	Source buffer service rate as calculated from i^{th} B-RM cell
N	Total number of VOPs in the video session
N_{cells}	Max number of cells in the queue in order not to violate the CDV
N_p	Total number of pixels in the VOP
$O_i(n)$	Luminance value of the i^{th} pixel of the original n^{th} VOP
Q	Quantization Index
$Q(t)$	Source queue size at time t
Q_{max}	Maximum queue size at the source buffer
ρ	Average utilization
$S(t_1, t_2)$	Amount of traffic sent by the source during interval $[t_1, t_2]$
T	Total time of the video session
τ	Time interval
t_i	Time between i^{th} and $(i+1)^{\text{th}}$ B-RM cell
VS_{ij}	VOP size for the j^{th} VOP with quality level i
VS_j	VOP size for the j^{th} VOP for a single quality stream

Chapter 1

Introduction

1.1 Thesis Motivation

Multimedia applications play an important role in the rapidly evolving era of connectivity and interactivity. Major research efforts started in the mid-nineties and are still in progress to define compression technologies, related tools, and communication networks for the efficient transmission of multimedia applications.

Nonetheless, large-scale deployment of multimedia video applications is still hindered by a major obstacle - the quality versus utilization efficiency trade-off, which can be translated into a trade-off between quality and cost as perceived by the user. Currently, to accommodate the bursty nature of video traffic, good quality video is guaranteed by reserving more resources than actually required. This leads to a high cost for the end user. On the other hand, if less resources are reserved to achieve higher resource utilization efficiency and lower cost for the end user, the result may be degradation of the quality of the transmitted video.

At the beginning of this research work in 1996, the standardization and research efforts on two independent fronts, namely the networking and video compression fronts, seemed to converge towards providing an effective solution for the quality versus utilization efficiency problem.

On the networking front, great success was achieved through the ATM Forum in defining the Explicit Rate (ER) algorithm for the Available Bit Rate (ABR) service for the second generation of ATM switches. Although the ABR service was originally designed for data applications and not for real-time applications, it remains the only service that offers a closed-loop feedback control mechanism. This mechanism has the potential to provide high bandwidth utilization based on the interaction between the user and the network.

On the video compression front, the MPEG working group, directs its efforts towards defining a new MPEG standard, namely MPEG-4 which is expected to become the enabling compression technology for multimedia transmission. This emerging standard promises a new way of interaction between the user, the coder, and the network while providing the required flexibility and interactivity for future multimedia applications.

Since the ATM/ABR networks have the capability to dynamically negotiate their resource requirements and to differentiate between the applications, it is possible to consider them as application-aware networks. In a similar manner, we can consider applications that involve MPEG-4 as network-aware applications since they will be able to monitor network conditions and dynamically re-negotiate the required Quality of Service (QoS).

This dynamic re-negotiation capability between the network and the application has been the primary motivation for this research since it offers the promise of solving the

major obstacle to large-scale deployment of video applications.

The transmission of MPEG-4 video over ATM/ABR networks provides several advantages. First, the dynamic negotiation capability enables it to provide a guarantee for minimum video quality, and to efficiently use available resources.

Moreover, in resource reservation networks (e.g. ATM/VBR) it is essential to define accurately the traffic descriptors in order to reserve the appropriate level of resources. However, in ATM/ABR networks, that combine both resource reservation and resource adaptation techniques, it is possible to alleviate the need for accurate traffic descriptors at the connection setup time since the application can dynamically re-negotiate the bandwidth requirements during the connection active time span.

The expected benefits described above provided the motivation to study the feasibility of transmitting MPEG-4 video over the ATM/ABR networks and to design a mechanism that controls their interactions.

1.2 Problem Definition

The objective of this research is to design an architectural model, and a set of algorithms that provide an efficient and economic way of transmitting video applications over ATM network without sacrificing the video quality. In order to achieve this objective, this research used ATM/ABR service for MPEG-4 video transmission. The MPEG-4 video provides efficient compression at low bit rates, which is one step toward achieving the two

goals of better quality, and more economic cost than any other standard. ATM/ABR service fulfills the requirement of a higher network utilization at a more economic cost than either CBR or VBR service. In addition, ATM/ABR service partially support the QoS requirements for video applications. However, this match between MPEG-4 and ABR is not straight-forward and poses several challenges as has been pointed out earlier in this chapter.

Providing an acceptable video quality can be translated into supporting the required QoS in terms of loss and delay. In order to achieve this, the problem should be divided into source, network and destination sub-problems while keeping in mind the overall performance. Figure 1-1 describes the end-to-end QoS and its various components. It is assumed that the application can tolerate an end-to-end delay of $D_{\text{end-to-end}}$ and end-to-end loss of $\text{Loss}_{\text{end-to-end}}$. Note that the loss can be caused by buffer overflow at the sender, the network, and/or the receiver. In addition, any packets that violate the delay constraint will be considered lost. For a successful transmission, the end system should negotiate with the network acceptable delay and loss levels and allocate extra quota for its own buffer while ensuring not to exceed the overall end-to-end delay and loss values $D_{\text{end-to-end}}$ and $\text{Loss}_{\text{end-to-end}}$.

Figure 1-1: End-to-End QoS

In the following subsections, the problems associated with the transmission of MPEG-4 video over ATM/ABR networks is mentioned in the context of the network and the end-system domains.

Network Domain Issues

When this research started in 1996, the ATM/ABR service was not mature enough, and several ER algorithms and related features were proposed. Additionally, given that the ATM/ABR service was originally designed to support data and delay-insensitive traffic, it provided no guarantee for bounds on delay or delay jitter in the network. Although several studies were underway to examine how to control delay in ATM/ABR networks, it was essential that this research validate the performance of the ATM/ABR networks, and determine the important set of features that need to be implemented in order to be able to support video application.

End-system Domain Issues

From the perspective of the end-system (source), the major challenge is characterizing the highly variable MPEG-4 coder output rate and matching it with the available ATM/ABR network bandwidth, which also varies as network loading conditions continue to change.

Another challenge is imposed by the video coding requirement, whereby the design of the source must provide a guarantee of the maximum delay that an ATM cell

encounters in the source buffer. ATM/ABR networks provide the application with a mechanism to dynamically re-negotiate bandwidth requirements and the MPEG-4 coder with a mechanism that controls its quality level. However, a proper mapping algorithm between the bandwidth and the quality level needs to be defined. This is complicated by the fact that the source does not know the rate associated with the improvement or degradation of the quality level.

1.3 Thesis Contributions

The major contribution of the research presented in this thesis is in the design and performance evaluation of a novel source architecture, and a set of supporting algorithms, which are both network-aware and application-aware, and which interact and interface with the ATM/ABR networks and the MPEG-4 coder. The main objective of the novel architecture and supporting algorithms is to resolve the end user domain issues thereby ensuring the efficient transmission of MPEG-4 video over ATM/ABR networks while guaranteeing a minimum level of video transmission quality of service. In addition, to alleviate the need for the user to accurately define the traffic descriptor parameters, tuning of the parameters of these algorithms must be less sensitive to the traffic characteristics. The following is a summary of the major and minor contributions of the research reported in this thesis dissertation:

- ***Contributions of the designed novel source architecture***

The novel source architecture, proposed and studied under a wide range of network conditions in this thesis, introduces and incorporates in its designed model several new ideas [Guirguis00a] [Guirguis00b]. The most significant ideas can be summarized as follows:

- The source relies on information from the network, the application and its own internal components in order to perform the required rate matching between the rate offered by the network and the MPEG-4 encoder rate combining adaptive traffic shaping, quality control and congestion recovery techniques.
- Keeping track of the cyclic behavior of the Video Object Plane (VOP) encoding sequence and using this information to provide protection to IVOPs.
- Maintaining on-line measurements of the required performance metrics and the capability to adapt the decision accordingly.

Another novelty of the source architecture is the integration and the interaction between its various components.

- ***Contributions of the proposed set of novel source algorithms***

During this research several algorithms were designed and their performance evaluated [Guirguis00a] [Guirguis00b] to support the newly designed source architectural model in attaining its main objective in the efficient transmission of MPEG-4 video over ATM/ABR networks. The performance of the algorithms was studied along with the

source to confirm the attainment of the thesis objectives. The design of each of the following novel algorithms adds to the contributions of this thesis and can be briefly summarized as follows:

- A Rate Request algorithm (RRA), which depends in its design on a simple Exponential Weighted Moving Average (EWMA) of the source rate to predict the future demand.
 - An Intelligent VOP dropping algorithm (IVDA), which is designed to consider the combination of network conditions, coder state, buffer occupancy, and delay constraints in its decision.
 - A Quality Level Selection algorithm (QLSA), which is designed to rely on on-line measurements, network conditions and coder state in its decision.
- ***Contributions of MPEG-4 video traffic and transmission studies***

This research has contributed several important studies and results to the network research community. Most significantly, this research resulted in the first published studies on the transmission of smoothed uncontrolled MPEG-4 video using real MPEG-4 traces at low-bit rates for the finite and infinite smoothing buffer cases [Guirguis99a]. The trade-offs between MCR, delay, and utilization for various MPEG-4 video traffic were also presented.

In addition to the above major contributions, the research presented in this thesis adds significant contributions to previous research in terms of the used input traffic, of the

expansion of the studied set of performance metrics, of the underlying ABR network, and of its compliance to the existing standards. These contributions can be summarized as follows:

- ***Input traffic:***

With respect to previous research, this thesis was the first to study and use real MPEG-4 traces coded in the low bit-rate range (64Kbps to 1Mbps) with layered coding. Due to the novelty of MPEG-4, all previous research focused on MPEG-1 or MPEG-2 video with bit-rates over 1Mbps. In addition, the focus of previous research was mainly on single layer video. The exception is the research by [Vickers98b], where multi-layered video is studied, but only for the sole purpose of supporting multicast which differs from its usage in this thesis.

- ***Performance Evaluations, metrics and ABR networks***

The performance of the ABR networks, the proposed architecture, and the source algorithms was evaluated and analyzed under a wide range of network conditions and environments. The performance studies and evaluations presented in this research include a wider consideration in terms of the observed performance metrics such as SNR, loss, delay, utilization, VOP rate [Guirguis98] [Guirguis99b]. Additionally, this study used second generation ABR switches which implement the ER algorithm. These contributions differentiate this thesis from previous research. For example in [Vickers97b] [Vickers98b] [Kanakia96], only binary feedback ABR networks were evaluated as opposed to the ER

networks. [Ahn97] [Vandalore98a] [Vandalore99b], which presented the study of ER network, had the delay as the single performance parameter that was considered without evaluating the video quality in terms of SNR and loss. In addition to the extensive studies of the performance and engineering of ABR networks, this thesis presents the first studies highlighting all the required features in the ER switches and in the source in order to support the transmission of video.

- ***Compliance to existing standards***

Taking the implementation perspective into consideration, one realizes that it is much easier to support application through the deployment of a solution that is based on source innovation rather than on switch innovation. Therefore, this thesis, unlike most of the previous research, conforms to the existing standards for both ATM/ABR and MPEG-4, and enforces its compliance throughout all the design contributions that are introduced and being proposed at the source, while maintaining a standard interface to ABR and MPEG-4. This compliance renders the future implementation of the ideas and contributions presented in this thesis more practical in terms of deployment. Most of the previous research depends on non-standard ABR networks. For example, research by [Ahn97] proposed a solution that is based on the definition of a new field in the RM cells, similarly in studies by [Vickers97b] [Vickers98b], the solution relies on the introduction of a new ATM service category called quasi VBR and source adaptive video service.

1.4 Thesis Outline

Chapter 2 provides a review of the MPEG-4 standard in general, of its video component in specific and of the current activities in the standardization body with respect to MPEG-4 delivery and QoS.

Chapter 3 presents the problem definition, challenges, and considerations for both the network and end-system (source) domains. Even though there has not been any published research on the transmission of MPEG-4 over ABR yet, Chapter 3 summarizes the most significant research literature on video transmission over ABR.

Chapter 4 focuses on the network perspective of the problem and presents the evaluations of second generation ATM/ABR switches. Although, there are several ABR implementations, they are not all equal since many ABR features that are attractive to video transmission are left optional in the ATM TM4.0 standard. Therefore, Chapter 4 provides a highlight of these features and analyses the delay requirements needed for video applications.

Changing the focus to the end-system, Chapter 5 investigates the traffic characteristics of MPEG-4 video and the performance of the transmission of MPEG-4 over ABR in cases of finite and infinite smoothing buffers.

Chapter 6 introduces the source architectural model that is proposed by this research. It also introduces the proposed source algorithms including the Rate Request

Algorithm (RRA), Intelligent VOP Discard Algorithm (IVDA) and the Quality Level Selection Algorithm (QLSA).

Chapter 7 presents the performance evaluation and analysis of the proposed source architecture and algorithms for various network environments.

Chapter 8 presents the main conclusion as well as a list of the major contribution of this thesis. Recommendations for future research are also included.

Chapter 2

MPEG-4 Standard Overview

This chapter provides a summarized overview of the MPEG-4 standard based on the complete specifications published in the MPEG-4 standard documents produced by the MPEG-4 working group (ISO/IEC JTC1/SC29/WG11) [MPEG4-2201], [MPEG4-2202], and [MPEG4-2606]. This chapter provides the required MPEG-4 background review needed for this research particularly of the MPEG-4 video codec capabilities and MPEG-4 video transmission. It should be noted that the information provided here is based on the Committee Draft (CD) documentation. The MPEG-4, whose formal ISO/IEC designations is ISO/IEC 14496 became an international standard in 1999.

2.1 History of MPEG standards

MPEG-4 is the latest ISO/IEC standard being developed by the Moving Picture Experts Group, also known as MPEG. MPEG is a very large group of experts dedicated to the development of standards for digital audio and video. In the early nineties, MPEG developed MPEG-1, the first MPEG standard, which was approved in 1992. The MPEG-1 standard was targeted to the storage and retrieval of video and audio on compact disc at bit rates of 1.5 Mbps. Later in 1994, MPEG delivered the MPEG-2 standard, developed for digital television which later became the technology of choice for High Definition TV (HDTV). MPEG-4 is the third and latest standard developed by the MPEG group, after the great successes of MPEG-1 and MPEG-2. MPEG-4 ver1 became an international standard in late 1999 (ISO/IEC 14496).

2.2 MPEG-4 Standard

The MPEG-4 standard is mainly targeted for multimedia applications. The explosion of the web clearly demonstrated the need for a new standard that can support interaction between the user and the audiovisual applications. In order to fulfill this need, MPEG-4 offers a new kind of interactivity, which integrates objects of different nature and allows universal access to multimedia information by taking into account the wide variety of networks. The MPEG-4 standard addresses the coded representation of both natural and synthetic (computer-generated) audio and visual objects. MPEG-4 systems were developed to provide the necessary elements in order to be able to specify how such objects can be composed to form complete scenes. The MPEG-4 standard is poised to become the enabling technology for multimedia communications and will therefore contribute to the solution of the problems that are hindering multimedia communications [Chiariglione97].

The MPEG-4 standard defines the audiovisual scene as a coded representation of audiovisual objects that have certain relations in space and in time. Distinct from other MPEG standards, MPEG-4 is the only coding standard that organizes the scene as a composition of arbitrarily shaped objects rather than rectangular frames. This *content-based coding* feature of MPEG-4 supports the integration of objects from different sources (natural and synthetic). Furthermore, MPEG-4 allows the user to interact with the objects without the need for transcoding as shown in Figure 2-1 on page 17, which illustrates the major components of an MPEG-4 terminal at the receiving end.

In addition to the *content-based* feature, MPEG-4 provides *scalable coding*, where scalable objects can be accessed at different bit rates depending on the available bandwidth. Scalability is a very important feature as it enables MPEG-4 applications to be provisioned in different network environments (mobile, internet, ATM, PSTN,... etc.). Based on the capabilities of the receiver and the network conditions, the appropriate layers are transmitted [MPEG4-2457]. MPEG-4 supports two types of scalability, namely spatial and temporal scalability.

MPEG-4 provides more efficient coding than existing standards. Natural audio is coded at bit rates ranging from 2Kbps up to and above 64Kbps, while the natural video coding ranges from 5Kbps (very low bit-rate coding) up to 600Mbps [MPEG4-2456]. Subjective testing shows that MPEG-4 provides 30%-50% bitrate savings over MPEG-1 for the same quality level for bit rates below 1 Mbps [MPEG4-2773].

Moreover, in order to support mobile and wireless environments, which are characterized by their high error rate, MPEG-4 provides tools and algorithms for error robustness and resilience through re-synchronization, data recovery and error concealment.

Figure 2-1: Major Components of an MPEG-4 Terminal at the Receiver End [MPEG4-2459]

2.3 MPEG-4 Delivery and QoS

The MPEG-4 working group realized the need to take into consideration the continuous changes and technological advances taking place at the application and the network domain. By such, MPEG-4 aims to support the evolution of multimedia applications without overloading it with the details and different technologies of the underlying network. Consequently, the MPEG-4 standard defines the concept of abstracting the application from the delivering network technology. The multimedia applications define a set of QoS parameters that the network should satisfy regardless of the deployed underlying network technology [MPEG4-2606]. MPEG-4 can be carried directly by digital streams at the physical layer, or over ATM or IP. The following section describes how this is achieved by MPEG-4.

2.3.1 Delivery Multimedia Integration Framework (DMIF)

Part 6 of the MPEG-4 standard [MPEG4-2606] defines a Delivery Multimedia Integration Framework (DMIF), which addresses the operation of multimedia applications over interactive networks. The DMIF architecture is such that applications, which rely on DMIF for communication, do not need to be concerned with the underlying communications method. The implementation of DMIF takes care of the network details presenting the application only with a simple standardized interface. In other words, DMIF's main objective is to hide the details of the transport network from the application and to assume responsibility for all the signaling and transport inter-operability between end-systems.

Figure 2-2: MPEG-4 System Layer Model [MPEG4-2606]

As shown in Figure 2-2 on page 19, the MPEG-4 codec produces Elementary Streams (ES), which are sent to the Synchronization Layer (SL). The synchronization layer sends the SL-Packetized streams to the DMIF layer, which is composed of a two-layer multiplexer. The first multiplexing layer, the DMIF Application Interface (DAI), allows the grouping of Elementary Streams (ES) with similar QoS requirements to reduce the number of network connections. The second multiplexing layer, the DMIF Network Interface (DNI), also defined as the TransMux (Transport Multiplexing) Layer, offers the transport services matching the requested QoS. Only the interface to the DMIF layer is specified by MPEG-4. Any suitable existing transport protocol stack such as RTP/UDP/IP, AAL5/ATM, MPEG-2 Transport Stream over a suitable link may become a specific TransMux instance. The choice of transport protocol is left to the end user/service provider and hence provides the flexibility for MPEG-4 to be used over a wide variety of network operating environments.

2.3.2 MPEG-4 Quality of Service (QoS)

MPEG-4 demands a generic set of QoS parameters for the different MPEG-4 media. However, the exact mapping between the MPEG-4 QoS parameters and the network QoS parameters is not specified in the standard and is left to the network providers.

Figure 2-3 on page 22 shows the different QoS involved when an MPEG-4 application requests the establishment of channels (with specific QoS and bandwidths) from the DMIF. The major functionality of the DMIF layer is to fill the gap between the application

QoS (referred to in the DMIF standard as transport QoS) and the network QoS. The application QoS, denotes the QoS required by the application. The network QoS, is the QoS offered by the network using the QoS parameters as defined by the network infrastructure. The DMIF takes the request from the application and computes the network QoS based on the deployed network technology.

The DMIF can also aggregate multiple Elementary Streams, carry them over a single TransMux, and compute the aggregate transport QoS, also known as the stream QoS. Furthermore, the DMIF then maps the stream QoS requirements for a particular TransMux into a specific network QoS.

Although, the QoS descriptor must be fulfilled by the transport network for a connection to be established, it is possible that during the connection lifetime in particular network operating environments, changes in network conditions may cause a variation in the QoS that actually are provided. Under such circumstances, the DMIF shall notify the application of the perceived degradation, so that the appropriate actions may be taken.

The QoS Descriptor shall be able to carry a set of QoS metrics. The set of metrics, that are used by the standard specifications to calculate the needed resources in order to meet the desired QoS, are summarized in Table 2-1 on page 23. Note that whereas the LOSS_PROB measures the individual Access Unit (AU) loss, the MAX_GAP_LOSS measures the whole stretch of AU that are missed.

Figure 2-3: Transformation of the Media QoS [MPEG4-2606]

Table 2-1: ISO/IEC 14496-6 (MPEG-4 DMIF) Defined QoS Metrics [Huard98]

QoS Metric	Description
MAX_Delay	Absolute maximum end-to-end delay for the stream
AVG_Delay	Average end-to-end delay for the stream
PREF_MAX_DELAY	Preferred maximum end-to-end delay for the stream
De-jitter Buffer	Buffer reserved for the removal of the transport jitter from the stream
LOSS_PROB	Maximum allowed probability of loss of access units
MAX_GAP_LOSS	Maximum allowable number of consecutive loss of access units
MAX_AU_SIZE	Maximum size of access unit
AVG_AU_SIZE	Average size of access unit
MAX_AU_RATE	Maximum arrival rate of access unit
PRIORITY	Priority of the stream

An example of a QoS mapping when delivering MPEG-4 over ATM/VBR networks is the mapping from Max_AU_Size, Max_AU_Rate and AVG_AU_Size to the PCR and SCR parameters of the VBR service as shown by the calculations of EQ 3.1 and EQ 3.2.

$$\text{PCR} = (\text{Max_AU_Size} \times \text{Max_AU_Rate}) / (48 \times 8) \quad (\text{EQ 3.1})$$

$$\text{SCR} = (\text{AVG_AU_Size} \times \text{Max_AU_Rate}) / (48 \times 8) \quad (\text{EQ 3.2})$$

The DMIF version 2.0 standard [MPEG4-2459] addresses the QoS support in different networks. The DMIF is capable of providing this support in both proactive and reactive networks. In proactive networks, the resources are allocated during the connection set-up, whereas in the case of reactive networks the resources are allocated during the lifetime of a connection based on resources availability. In order to support both types of networks, DMIF defines three types of DMIF Application Interface (DAI) primitives:

Request of QoS

These primitives specify the required QoS, either qualitatively or quantitatively, in terms of delay, loss and required bandwidth.

Monitoring of QoS

These primitives monitor the QoS offered by the network and provide a feedback mechanism from the DMIF layer informing the application layer about the underlying network resources. There are three modes defined for QoS monitoring: periodic monitoring,

QoS violation detection and application request as shown in Figure 2-4 on page 26.

Renegotiation of QoS

These primitives are proposed to be added in the future version of the DMIF to allow for the change of the Channel Descriptor of an already established channel in a DMIF session [MPEG4-3202].

It should be mentioned that several of the DMIF activities are still underway. However, what is of interest to this research is the concept of introducing a mechanism for interaction between the MPEG-4 application and the underlying network.

Figure 2-4: QoS Monitoring Modes [MPEG4-2606]

2.4 MPEG-4 Natural Video Coding

As mentioned earlier in this chapter, MPEG-4 allows the coding of rectangular as well as arbitrarily-shaped video objects of a scene. Since it supports both scalable and non-scalable coding, MPEG-4 allows handling of normal scalabilities as well as object scalabilities. Figure 2-5 on page 28 shows the basic classification of the bit rates and functionalities currently provided by the MPEG-4 visual standard for natural images and video. At the lowest end, a Very Low Bit rate Video Core (VLBV), provides algorithms and tools for applications operating at bits rates between 5Kbps and 64Kbps. However, MPEG-4 also supports higher rates up to 600Mbps [Koenen97].

Figure 2-5: Classification of MPEG-4 Image and Video Coding Algorithms [MPEG4-2459]

2.4.1 MPEG-4 Video Structure

MPEG-4 coded video data consists of an ordered set of video bit-streams known as layers. If there is only one layer, the coded video data is referred to as a non-scalable video bit-stream. Multi-layers are referred to as scalable video bit-streams [MPEG4-2202].

A video object, as defined in the standard, is an entity that a user is allowed to access and manipulate [MPEG4-2202]. The instances of video objects at a given time are called Video Object Planes (VOPs). The encoding process generates a coded representation of a VOP as well as the composition information that is needed for the display. At the decoder level, the user may interact with and modify the composition process as needed. Natural video objects consist of a sequence of two-dimensional representations, referred to as VOPs, which include representation of its shape, its motion and its texture.

The following hierarchy data structure is defined for MPEG-4 video objects [Ebrahimi97].

Video Session

A Video Session (VS) is defined as an entity embedding the following three classes: Video Object, Video Object Layer and Video Object Plane. A complete video sequence may be formed of one or more video sessions.

Video Object

A Video Object (VO) is a class defining specific objects in a scene. The video

object can be of arbitrary or rectangular shape.

Video Object Layer

The Video Object can be composed of one or more layers called Video Object Layer (VOL). The various layers can be used to enhance the temporal or spatial resolution of a VO.

Video Object Plane

A Video Object Plane (VOP) is an occurrence of a VO at a given time. Note that two different VOPs may belong to the same object at two different instances of time.

In summary, as shown in Figure 2-6 on page 31, a Video Session may consist of one or more Video Objects, each of which in turn can consist of one or more Video Object Layers, where each layer is composed of an ordered sequence of snapshots in time in the form of Video Object Planes.

Figure 2-6: Hierarchy of the Data Structure Classes in MPEG-4 Video [Ebrahimi97]

2.4.2 Object-Based Encoding

The object-based encoding of MPEG-4 is described in details in [MPEG4-2202] In summary, the encoding process makes use of a number of techniques to obtain high compression rates. The algorithm starts by applying Block-Based Motion Compression to reduce temporal redundancy. Motion compensation is used both for prediction of the current VOP from a previous VOP, as well as for interpolative prediction from past and future VOPs. Discrete Cosine Transform (DCT) is then used to further compress the prediction error to remove spatial correlation before it is quantized. The Quantization is an irreversible process, that discards the less important information. The last stage of the encoding uses variable length codes to encode the shape information, motion vectors and the quantized DCT information.

MPEG-4 defines three basic VOP types: Intra-coded VOPs (I-VOPs), Predictive-coded VOPs (P-VOPs) and Bidirectionally-coded VOPs (B-VOPs). I-VOPs are coded without references to other VOPs and provide access points to the coded sequence where decoding can begin, and hence are only coded with a moderate compression ratio. On the other hand, P-VOPs are coded more efficiently using motion compensated prediction from past Intra and Predictive-coded VOPs and are generally used as a reference for future prediction. However, B-VOPs provide the highest compression ratio but require both past and future reference VOPs for motion compensation. The organization of the three VOP types in a sequence is very flexible and has to be specified through an encoder parameter.

For natural video scenes, VOPs are generated by segmenting the scene according to some semantic meaning. The shape information is referred to as the binary alpha plane, which is coded on a macroblock basis by a coder that makes use of the context information, motion compensation and arithmetic coding.

Figure 2-7 on page 35 shows the basic block diagram of an MPEG-4 video encoder. The basic coding structure involves shape coding for arbitrarily shaped VOs and motion compensation as well as DCT-based texture coding using standard 8x8 DCT or shape adaptive DCT. An important advantage of the content-based coding approach is that the compression efficiency can be significantly improved for some video sequences by using the appropriate and dedicated object-based motion prediction tools for each object in a scene.

Segmentation

The MPEG-4 standard offers several content-based functionalities, requiring that the description of the scene be in terms of Video Objects (VOs). Due to the flexibility to access the bit-stream and the easy manipulation of the video information, the separate coding of the video objects enables user interaction with multimedia applications. The encoder performs some pre-processing in order to automatically identify the objects appearing in the sequence. Hence, segmentation of the scene is a key issue to efficiently apply the MPEG-4 coding scheme.

In other words, segmentation is the process of partitioning an image or video

sequence into regions extracted according to a given criterion or semantic meaning. Different methods have been proposed for segmenting the video sequences, based on either spatial homogeneity (a motion coherence criterion), or on joint processing of spatio-temporal information. The details of the segmentation techniques are out of the scope of this research but can be found in [Colonna97a] [Mech97].

Sprite

The static sprite technology enables efficient encoding of video objects, whose content along a video sequence is not expected to vary in time. It is well suited to represent backgrounds of scenes. A static sprite is a frame containing spatial information for a single object, obtained by gathering information for that object throughout the sequence in which it appears.

MPEG-4 syntax defines a dedicated coding mode to obtain VOPs from static sprites, the so called Sprite VOPs. Sprite VOPs can be used for background objects over which classical objects are superimposed.

Figure 2-7: Basic Block Diagram of the MPEG-4 Video Encoder [Ebrahimi97]

Figure 2-8: Example of Sprite Coding of a Video Sequence [MPEG4-2459]

2.4.3 MPEG-4 Scalability

The scalability feature of MPEG-4 can prove to be of importance to several applications. Some applications that can take advantage of scalability over a single layer video are internet video, wireless video, and multi-quality video services as will be demonstrated by the research presented in this thesis.

A simple solution to achieve scalable video is the simulcast technique, that is based on the transmission and storage of multiple independently coded reproductions of video. A more efficient alternative is scalable video coding, by which the bandwidth is allocated to a given reproduction of the video that can be partially re-utilized in coding the next reproduction of the video. Scalability provides resilience to transmission errors as the more important data of the basic layer can be sent over a channel with better error performance, whereas the less critical enhancement layer can be sent over a channel with poor error performance. MPEG-4 supports two basic types of scalability: *temporal scalability* and *spatial scalability*. In addition, a combination of these two basic scalability types is referred to as *hybrid scalability*. In the current standard, temporal and spatial scalability define two layers of video: the base and the enhanced (Enh) layer. Hybrid scalability defines up to four layers.

Temporal Scalability

Temporal scalability involves partitioning VOPs into layers, where the basic layer is coded by itself to provide the basic temporal rate, and the enhanced layer is coded using

temporal prediction with respect to the lower layer. Both layers have the same spatial resolution, but when decoded and temporally multiplexed, yield full temporal resolution i.e. increase the frame rate of the video. Temporal scalability is useful in bandwidth limited network applications, where adaptation to frequent changes in the bandwidth is necessary. One good example of such networks is the ATM/ABR networks where the bandwidth can be frequently changed. Object-based temporal scalability can also be employed to allow graceful control of picture quality by controlling the temporal rate of each video object under the constraint of a given bit budget.

The MPEG-4 standard defines three coding sequence variants for temporal scalability, which are listed in Table 2-2 on page 37.

Table 2-2: Temporal Scalability Types

Temporal Scalability Type	Coding Sequence			
0	Enh	P	P	
	Basic	I	P	P
1	Enh	BB	BB	
	Basic	I	P	P
2	Enh	P	B	B
	Basic	I	B	P

Spatial Scalability

Spatial scalability involves generating two spatial resolution video layers from a single video source, such that the basic layer is coded by itself to provide the basic spatial resolution, and the enhanced layer employs the spatially interpolated basic layer to carry the full spatial resolution of the input video source. The MPEG-4 standard defines two types of spatial scalability. In the first type, the enhanced layer is coded as “PPPPP.....”. In the second type, the enhanced layer is coded as “PBBBBBB....”.

Hybrid Scalability

There are a number of applications, where neither the temporal scalability nor the spatial scalability may offer the necessary flexibility and control. This may necessitate the use of temporal and spatial scalability simultaneously, which is known as *hybrid scalability*.

2.5 MPEG-4 Video Decoding Process

The MPEG-4 video decoder is mainly composed of three building blocks: the shape decoder, motion decoder and texture decoder (shown in Figure 2-9 on page 39). The reconstructed VOP is obtained by combining the decoded shape, texture, and motion information.

Figure 2-9: Simplified MPEG-4 Video Decoding Process [MPEG4-2202]

Generalized Scalable Decoding

In the case of temporal scalability, both rectangular as well as arbitrarily-shaped VOPs are supported. However, in the case of spatial scalability, only rectangular VOPs are supported.

Figure 2-10 on page 41 depicts a high level decoder structure for generalized scalability. When spatial scalability is performed, Mid-Processor1 performs spatial up or down-sampling of input. The scalability post-processor performs any needed operations, such as spatial up or down-sampling of the decoded basic layer for display purposes resulting at output-0, whereas the enhancement layer without any resolution conversion may be obtained from output-1.

When temporal scalability is to be performed, the decoding of basic and enhanced layer bit-streams occurs at the corresponding basic and enhanced layer decoders shown in Figure 2-10 on page 41. In such a case, Mid-Processor1 does not perform any spatial resolution conversion. The post-processor simply outputs the basic layer VOPs without any conversion, but temporally multiplexes the basic and enhanced layer VOPs to produce a higher temporal resolution enhancement layer.

Figure 2-10: High Level Decoder Structure for Generalized Scalability [MPEG4-2202]

2.6 Visual Profiles

The profiles for the coding of natural, synthetic, as well as synthetic and natural hybrid visual content are described in [MPEG4-2458]. The following are the profiles for natural video content:

- The Simple Visual Profile provides efficient, error resilient coding of rectangular video objects, and is suitable for applications on mobile networks.
- The Simple Scalable Visual Profile adds to the Simple Visual Profile the support for the coding of temporal and spatial scalable objects. It is useful for applications, which provide services at more than one level of quality due to bit rate or decoder resource limitations.
- The Core Visual Profile adds to the Simple Visual Profile the support for the coding of arbitrarily-shaped and temporally-scalable objects. It is useful for applications, such as those providing relatively simple content-interactivity (internet multimedia applications).
- The Main Visual Profile, which adds to the Core Visual Profile the support for the coding of interlaced, semi-transparent, and sprite objects. It is useful for interactive and entertainment-quality broadcast as well as DVD applications.

2.7 Variable Bit Rate versus Constant Bit Rate Encoding

The output bit rate of the MPEG-4 encoder, similar to many other coders, can be selected to be either Constant Bit Rate (CBR) or Variable Bit Rate (VBR) streams. Either can be chosen depending on the transport media and the required quality. However, given the nature of the coding process, which involves compression, each VOP may have a different compression ratio based on its type (I,P or B), and based on the spatial and temporal redundancy, leading to different VOP sizes and hence variable rates. If it is required to maintain a constant quality level for the encoded video this will yield variable and bursty bit rates. On the other hand, to maintain a constant traffic rate (CBR), the encoder has to vary the quantization level from VOP to VOP, which results in a varying video quality.

Figure 2-11 on page 44 shows the relationship between the distortion and the required coding rate. With respect to transmission, CBR encoding is simpler and needs an uncomplicated admission control scheme in the network, however the quality of the transmitted video will be constantly varying. Various research studies indicate the benefits of VBR coding in terms of better quality and the expected multiplexing gain when transporting VBR video [Krunz99] [Luo97]. Still, VBR video introduces several challenges with respect to QoS guarantees for such bursty traffic and which have been a hot subject for many studies in recent years. Note that this thesis focuses on VBR video coding and whenever we refer to video encoding stream it will implicitly mean VBR encoding.

Figure 2-11: Distortion Level vs. Coding Rate

Chapter 3

MPEG-4 Video Transmission over ATM (Problem Definition)

3.1 Introduction

Large bandwidth requirements for multimedia applications can be considered the major factor hindering its wide spread. This thesis focuses on the study of MPEG-4 video transmission as the transmission of the video component is very demanding in terms of bandwidth requirements. Moreover, MPEG-4 with its low bit-rate encoding capabilities is expected to remedy this major obstacle, as will be demonstrated in the studies presented in Chapter 5. However, bandwidth requirements do not constitute the only challenge for video transmission as most multimedia applications require stringent QoS guarantees, which require the support of the underlying network. This chapter, focuses on the QoS requirements for the MPEG-4 video and its transmission over ATM networks in general and on the ABR service in particular. A summary of the previous research carried out in this area is presented. Finally, this chapter concludes with a formulation of the problem that is being solved throughout this thesis.

3.2 QoS Requirements for Video Applications

There are four basic QoS metrics that need to be satisfied with certain levels of guarantee for all video applications namely: bandwidth, delay, delay jitter and loss. However, the exact value attached to each of these metrics depends on the application type.

These QoS metrics are defined as follows:

Bandwidth

Bandwidth metric determines the minimum bandwidth required from the network to support multimedia applications, and is given by the mean bit rate of the source. However, due to the fluctuations and bursty nature of multimedia traffic, the actual bandwidth that is needed is typically higher than the mean rate, and depends on several factors including the sensitivity to loss and delay. Bandwidth requirements vary greatly from one application to another. For example, video conference applications with low frame rates and low motion activities require much less bandwidth than Video on Demand (VoD) application. The network should guarantee a certain amount of bandwidth for video applications in order to satisfy the required loss and delay bounds. The bandwidth requirements can range from as low as 24Kbps to as high as 40Mbps.

Delay

Delay is defined as the time needed to deliver the VOP from the sending application to the receiving application. Interactive communication has stringent requirements in terms of maximum delay. Acceptable values for real-time video applications are normally less than 450 ms [Krunz99]. However, non-real time applications can tolerate higher values.

Delay Jitter

Delay jitter is an essential QoS metric that needs to be controlled in most video applications. Delay jitter occurs due to the arrival of packets/cells after widely varying transfer delays. In order to overcome this delay variation, the receiver should buffer the packets and send them to the application at constant intervals of time. This mechanism is called De-Jitter, however, it increases the end-to-end latency.

Loss

Packet or cell loss can occur at both the end-system and in the core network mainly due to buffer overflow. Generally, these losses have great impact on the quality of the video. However, different applications can tolerate different loss ratios. Typically, the loss ratio ranges between 10^{-2} to 10^{-6} [Krunz99].

In spite of the significant performance impact of the above mentioned video QoS metrics, in the context of end-system applications, what is actually perceived is the delay and loss (assuming the receiver has enough buffer to De-Jitter the delay). In other words, bandwidth requirements should be guaranteed in order to guarantee loss and delay. In order to define the QoS requirements for video applications, they are classified as real-time when video transmission is live from a camera. or non real-time when videos that have been pre-recorded and stored. Real-time communications systems are targeted towards applications which encompass a two-way human interaction, or towards one-way applications that impose strict one way delay constraints. Video conferencing is a prime

example of a two-way real-time system. Real-time video applications are faced with two challenges; namely the stringent delay requirements, and the unavailability of accurate traffic characteristics of the video being transmitted. However, they normally tolerate loss and require less bandwidth than non real-time video applications.

3.3 MPEG-4 Video Transmission over ATM Networks

Multimedia applications require the support of the underlying networks to guarantee its stringent QoS requirements. However, the internet cannot, at the present time offer the required QoS for multimedia applications. ATM networks promises an efficient way for the transmission of voice, video and data over the same network while guaranteeing the QoS for each type of traffic. In this section, the various alternatives for the transmission of MPEG-4 video over ATM networks were reviewed, focusing on the low-bit rate range as the most significant contribution added by the MPEG-4 standard.

The ATM Forum defines six types of services namely: Constant Bit Rate (CBR), non-real time Variable Bit Rate (nrt-VBR), real-time Variable Bit Rate (rt-VBR), Available Bit Rate (ABR), Unspecified Bit Rate (UBR), and Guarantee Frame Rate (GFR) services. Each of these services is characterized by a different level of QoS guarantee and is targeted towards different applications. Video applications, in general, are expected to be carried by either CBR or rt-VBR service. In this section, the advantages and disadvantages of using each of these services for video transmission are demonstrated. Section 3.4 highlights the more challenging alternative for video transmission using the ABR service.

3.3.1 Quality of Service Guarantee in ATM Networks

The ATM networks have a well-defined set of QoS parameters described in the ATM Forum Traffic Management 4.0 [ATMF96]. These QoS parameters may be negotiated between the end systems and the network. Based on the service type, part or all of these QoS parameters are supported. The QoS parameters that are negotiated are the delay and loss parameters.

The two negotiated delay parameters are expressed in terms of Cell Transfer Delay (CTD), which is defined as the elapsed time between a cell exit event at the source User Network Interface (UNI) and the corresponding cell entry event at the destination UNI for a particular connection. The CTD is best explained through the typical cell delay probability density model as shown in Figure 3-1 on page 50.

The following two delay parameters are negotiated in ATM networks:

Maximum Cell Transfer Delay (maxCTD)

maxCTD for a connection is the $(1-\alpha)$ quantile of CTD distribution function

$$\text{i.e. if maxCTD} = \tau \quad \text{then} \quad \text{Prob}\{\text{CTD} > \tau\} < \alpha. \quad (\text{EQ 3.1})$$

Peak-to-Peak Cell Delay Variation (peak-to-peak CDV)

The peak-to-peak CDV is the difference between $(1-\alpha)$ quantile of the CTD and the fixed delay as in EQ 3.2.

$$\text{peak-to-peak CDV} = \text{maxCTD} - \text{Fixed Delay} \quad (\text{EQ 3.2})$$

Figure 3-1: Cell Transfer Delay Probability Density Function

If the fixed delay is negligible then peak-to-peak CDV will be equal to maxCTD. The definition of peak-to-peak CDV can be translated as the connection queue size. Assuming a FIFO scheduler then upon the arrival of a cell at its connection queue, the number of cells already in the connection queue should not exceed N with a probability higher than $(1-\alpha)$. N is calculated according to EQ 3.3.

$$N_{\text{cells}} = \frac{\text{maxCTD} - \text{Fixed Delay}}{\text{Cell Service Time}} \quad (\text{EQ 3.3})$$

With respect to loss, the ATM Forum defines the Cell Loss Ratio (CLR) parameter, as the ratio between lost cells and the total transmitted cells. Both the network and the end system negotiate the CLR that should be guaranteed by the network over the life-time of the connection.

It should be noted, that QoS guarantees in ATM networks are satisfied only if the user traffic follows the agreed-upon contract. Traffic characterization is not always a trivial task. Consequently, this shifts the QoS guarantee that is to be offered by ATM networks to a traffic characterization problem. Table 3-1 on page 52 outlines the input traffic, QoS parameters and the types of guarantees offered for the three ATM services of interest.

Table 3-1: Traffic Descriptor and Qos Guarantee for CBR, rt-VBR and ABR Services

Traffic Descriptor Parameter	CBR	rt-VBR	ABR
Peak Cell Rate (PCR)	Specified. It is always guaranteed.	Specified. This rate is guaranteed for a maximum duration of Maximum Burst Size.	Specified. It is offered only if the network has enough bandwidth.
Sustainable Cell Rate (SCR)	Not Applicable	Specified. This rate should always be available for the source.	Not Applicable.
Maximum Burst Size (MBS)	Not Applicable	Should be specified to control the maximum duration that the source can send at Peak Cell Rate.	Not Applicable
Minimum Cell Rate (MCR)	Not Applicable	Not Applicable.	Specified. This rate is always guaranteed to be available.
CDV Tolerance (CDVT)	Specified	Specified	Specified
QoS Parameters			
MaxCDT	Guaranteed	Guaranteed	Not guaranteed
Peak-to-Peak CDV	Guaranteed	Guaranteed	Not guaranteed
CLR	Guaranteed	Guaranteed	Guaranteed
Feedback	Not Applicable	Not Applicable	Closed loop

3.3.2 Video Transmission over ATM/CBR Service

In the case of the ATM/CBR service, the connection requests a static amount of bandwidth to be continuously available during the life-time of the connection. This bandwidth is determined by the Peak Cell Rate (PCR). The CBR service is considered a resource reservation service, which means that traffic characteristics should be known in advance, and the network should possess adequate knowledge about how its resources are committed. Hence, based on such information, the network can either accept or reject a connection. The CBR service is intended to support real-time applications with stringent delay constraints. As long as the cells conform to the traffic contract, the CBR service guarantees the agreed upon peak_to_peak CDV, max CTD and CLR.

Figure 3-2 on page 54 describes the video transmission over the CBR service, where the MPEG-4 internal buffer and its rate control are used to produce a constant bit-rate that matches the rate offered by the CBR service. However, it should be mentioned that video traffic is bursty in nature due to the used compression techniques, and that using a CBR channel causes a quality degradation for scenes with large information. CBR traffic is easy to manage; however, previous research [Dalgic97] has shown that using CBR service for video transmission can result in either unacceptable level of image degradation, or inefficient use of bandwidth.

Figure 3-2: Video Transmission over CBR Service

3.3.3 Video Transmission over ATM/ VBR Service

The ATM/VBR service differs from the CBR service in the negotiation of the Sustainable Cell Rate (SCR), and the Maximum Burst Size (MBS) in addition to PCR. The source is expected to send traffic with a mean rate equal to SCR. However the source can occasionally send a burst of size equal to MBS cells at PCR. Any cells sent with a rate higher than PCR or with a burst larger than MBS, will be either dropped or tagged as least priority cells. With the VBR service, the network will reserve a bandwidth equal to a value between SCR and PCR depending on the Call Admission Control (CAC) algorithm that is used.

As video traffic is bursty in nature (variable bit rate), it becomes more natural to use the VBR service for video transmission. In this case, the rate control which has been used for fitting an encoded bit-stream to a fixed capacity channel is therefore disregarded. The expected gain is the ability to have a constant perceptual quality of the reconstructed video at the receiver side. Additionally, as video sources generate bursts at different times, this results in multiplexing gain when several video sources are aggregated at the link level. Hence, video transmission over VBR service promises an efficient use of bandwidth without sacrificing quality. However, statistical multiplexing may lead to congestion, and as a consequence, cells may be discarded. In addition, VBR traffic is difficult to manage and successful transmission of video depends on the accurate estimation of the traffic descriptors (PCR, SCR and MBS) which again is not a trivial task. Figure 3-3 on page 57 depicts the transmission of video over VBR service using the token bucket to ensure the

conformance of source cells. As shown, the video encoder produces a VBR traffic which is enqueued in the smoothing buffer. The smoothing buffer is served with a rate less than, or equal to, PCR. The token buffer is drained at a constant rate equal to at least SCR. Each cell sent from the smoothing buffer adds a token into the token buffer. The cell is considered a conforming cell and is allowed to enter the network if the token buffer is not full. If the token buffer is full, the cell is considered to be non-conforming and is either dropped or tagged. The VBR service guarantees delivery of conforming cells by reserving bandwidth equal to SCR and a buffer space equal to MBS at each node. Several studies have been carried out to study VBR video transmission over ATM/VBR service [Luo97], [Pancha94], [Hamidi97], and [Hsu97].

Figure 3-3: Video Transmission over VBR Service

3.4 MPEG-4 Video Transmission over ABR - Motivations & Challenges

From the review presented in the Section 3.3, it becomes evident that in terms of QoS guarantees, the rt-VBR service seems to be the most suitable service for the transmission of real-time video. Nevertheless, two important points need to be considered. First, to obtain the guarantee for the rt-VBR case, the traffic contract parameters should be accurately defined. For a user to define the traffic contract parameters, the user needs to be aware of the traffic that *will* be transported over the network. Numerous research studies in the areas of video traffic characteristics and traffic modeling [Ismail95], [Izquierdo98] and [Rose97] are in progress. These studies are trying to accurately determine the video traffic descriptors with the ultimate objective of designing efficient admission control schemes so that the QoS can be guaranteed. These studies have been successful for pre-recorded video, such as Video-On-Demand (VoD) applications, where the traffic can be known in advance. However, when it comes to real-time video this task becomes more difficult as there is no beforehand information on the forthcoming traffic. Secondly, in the case of VBR as with all the other ATM services, except ABR, the bandwidth reservation is done once during the connection setup. This may lead the network to drop some user cells even if the link is under-utilized. Practically, video transmission over VBR typically leads to either overestimating traffic parameters, and hence reserving more bandwidth than required and sacrificing utilization to guarantee the required QoS for video, or to achieve a high utilization while risking QoS violation. These reasons motivate researchers to search

for alternatives other than *rt_VBR* for video transmission. One alternative is the use of the ABR service. In this section, the motivations and challenges involved in this approach are discussed, after a brief overview of how the ABR service works, and specifically its flow control mechanism which is of specific interest to this study. More details on the ABR service specifications are presented in Chapter 4.

As shown in Figure 3-4 on page 60 and as described in more details in the ATM forum Traffic Management Version 4.0, the ABR flow control algorithm can be summarized as follows:

- ABR flow control occurs between a sending end-system (source) and a receiving end-system (destination).
- A source generates Forward-RM-cell (F-RM), which are turned around by the destination and sent back to the source as Backward-RM-cells (B-RM).
- Backward RM-cells carry the feedback information provided by the switch and/or destination back to the source.
- The switch may directly insert feedback information into RM-cells as they pass in the forward or backward direction. It may also indirectly inform the source about congestion by setting the Explicit Forward Congestion Indicator (EFCI) of the data cell header for the forward-flow information cells. In this case, the destination could update Backward-RM-cell. The switch may also generate Backward-RM- cells.
- Source, destination and the switch should follow the respective behavior as specified by the ATM Forum.

Figure 3-4: ABR Control-Loop

Several studies were conducted to investigate the feasibility of video transmission over ABR as will be reviewed in Section 3.5. In summary, the published studies confirm the feasibility and the expected advantage for the transmission of real-time video over ABR, and are summarized as follows:

- ABR service is the only ATM service, which provides a closed -loop feedback. The ABR feedback control can be used to adapt the coder rate based on the available bandwidth as conveyed by the feedback from the network.
- ABR service through RM-cells presents a way for a dynamic renegotiation of the bandwidth distinct from the other ATM services, that have their bandwidth requirements negotiated only at the connection setup with no specified standard process for renegotiation.
- The ABR traffic contract is less demanding than the VBR traffic contract. There is no demand for a detailed information of the future traffic to be able to specify the required ABR traffic parameters.
- The minimum acceptable quality level can be guaranteed through MCR.
- ABR service can be expected to provide an economic alternative for transporting video as it is expected that the carriers will price the ABR service more attractively than either CBR or VBR services since ABR allows high utilization of the network bandwidth.

Yet, the major disadvantage of the ABR service is the lack of guarantee for the delay and the delay jitter. In addition, the effect of the ability of the source to adapt to

changes in network conditions, as conveyed by the RM-cells, needs to be investigated for the case of video applications. Finally, a set of intelligent algorithms and techniques needs to be implemented at the source to match the source rate with the rate offered by the network while maintaining the required QoS.

3.5 Review of Previous Research on Video Transmission over ABR

Due to the novelty of MPEG-4, to my knowledge, there has not been any published papers on the MPEG-4 video transmission over ATM Networks. However, there are several published papers covering video transmission over ATM in a general sense. Between 1995 and 1999 several studies investigated MPEG-1 transmission over ATM/ABR. In this section, the most significant studies in this area are summarized. It is worth mentioning, that the major difference in the nature of the problem between video transmission in general and MPEG-4 in particular is two-fold. First, the MPEG-4 uses different algorithms and tools than previous MPEG standards and hence the traffic characteristics are different. Second, the bandwidth requirement is different, MPEG-1 and MPEG-2 need bandwidth in excess of 1Mbps for 240*360 pixels and 30 fps. However, MPEG-4 bandwidth requirements may be as low as 90Kbps for 240*360 pixels at 30 VOP/s.

3.5.1 AT&T Research Lab

The idea of transporting video over ABR service was first introduced in 1995 by Kanakia and others from the AT&T research lab [Kanakia95]. This research demonstrated

the benefits of modulating the source rate of the video encoder based on the network feedback. The major benefits are the efficient use of the transmission capacity and the graceful degradation of video quality in cases of congestion. This research also disproves the belief that feedback-based control is not useful for real-time video transmission. The experiments conducted in this research used real video traces generated by an MPEG-1 encoder. The main performance metrics used in this study were SNR and the perceptual quality of the image. The rate of the coded video sources varied between 2Mbps to 8Mbps. In this research, it was assumed that the video encoder had a rate control mechanism that could be used to match the target rate computed by the feedback mechanism. A continuation of this research proposed an algorithm for Smoothed Adaptive Video Over Explicit Rate Networks (SAVE) [Duffield98], [Lakshman97], [Kanakia96].

3.5.2 Ohio State University

Raj Jain of Ohio State University and his group conducted several studies on the ABR service, and the Explicit Rate (ER) switch design [Kaly00] [Kaly97b] [Kaly98]. Although they did not publish any results of video traffic studies, they studied several features that are important to video transmission. In [Jain96], Jain introduced the concept of Real-Time ABR service to the ATM Forum, where he suggests a new service called *rt_ABR* that provides delay guarantees. He continued investigating controlling the delay within the current ABR service, and he demonstrated that any ER switch algorithm with a fast response time and a queue control can guarantee a small delay in the ER switch

[Vandalore98a] [Vandalore99b]. In 1999, Vandalore published a paper [Vandalore99a] on the support of multimedia applications over ABR service explaining the expected benefits. Yet, no results were published. It should be mentioned that the research carried out by the Ohio State University group lies completely in the network domain and no published studies have been performed on the end system domain.

3.5.3 University of California at Irvine

Several published papers have resulted from Vickers' Ph.D. thesis [Vickers98b]. This thesis research handled both the point-to-point case and the multipoint case [Vickers97a] [Vickers98a]. Of more interest is the point-to-point part of the research as the multipoint case is out of the scope of this research. In [Vickers97b], an ATM service architecture referred to as quasi-VBR was proposed for the transmission of adaptively encoded live video. The quasi-VBR service is based on a binary feedback mechanism of the EPRCA rate control algorithm proposed in [ATMF96] as an example of an implementation for the ABR service. In this research, the author used actual MPEG-1 video sequences. The proposed mechanism does not rely heavily on prior accurate traffic parameter estimation as the minimum guaranteed rate was set to 0.6Mbps, and the maximum rate to 50Mbps for all the sequences in spite of their different characteristics. The encoder adjusts its rate by controlling the quantization step based on the buffer occupancy. This research left many open areas for further studies, including the usage of an ER mechanism instead of a binary feedback. The study focuses on utilization and PSNR. However, no

results were reported with respect to delays which is a very important metric for video transmission. It should be noted that the proposed scheme requires changes to the current standard as well as a new ABR switch implementation.

3.5.4 Korea Advanced Institute of Science and Technology (KAIST)

In this study, the transmission of MPEG video over ABR was investigated. The contribution of this research lies mainly in the network domain where an Explicit Rate algorithm, "Explicit Rate Indication Scheme for MPEG" (ERISM), is proposed which allows for non zero MCR and maintains a small queue length in the switch. This research relies on having adaptive encoders. The MPEG-1 encoder generates two streams of MPEG frames, one is compressed under a maximum compression ratio (between 2-4 Mbps) and the other stream is encoded with a quantization level determined by a Traffic Anticipation Unit (TAU). An ABR Service Interface Unit (ASIU) is defined which transmits the MPEG-1 frames at the rate allocated by the network. If the generated frame size needs bandwidth more than the allocated then ASIU transmits the frame with the maximum compression ratio.

One major drawback of this scheme is that it cannot provide a graceful degradation in the quality of service. In the event that the MPEG-1 encoder generates frame sizes larger than the cell rate allocated by the network then the frame with the minimum quality is transmitted. This event can occur frequently in real-time traffic due to inaccurate estimation by the TAU or due to congestion in the network. Another major drawback is its

incompatibility with the current standard as it requires the definition of a new field referred to as the Available Rate (AR) in the RM-cell. The research focuses on delay as the performance metric, yet no results on the quality of the transmitted video were reported. This approach requires extensive processing time due to the prediction model and the continuous generation of two encoded MPEG frames. The details of this research can be found in [Ahn97].

3.6 Problem Formulation

The objective of this research is to design an architectural model, and a set of algorithms that provide an efficient and economic way of transmitting video applications over ATM network without sacrificing the video quality. In order to achieve this objective, this research used ATM/ABR service for MPEG-4 video transmission. The MPEG-4 video provides efficient compression at low bit rates, which is one step toward achieving the two goals of better quality, and more economic cost than any other standard. ATM/ABR service fulfills the requirement of a higher network utilization at a more economic cost than either CBR or VBR service. In addition, ATM/ABR service partially support the QoS requirements for video applications. However, this match between MPEG-4 and ABR is not straight-forward and poses several challenges as has been pointed out earlier in this chapter.

As mentioned in Section 3.2, providing an acceptable video quality can be translated into supporting the required QoS in terms of loss and delay. In order to achieve this,

the problem should be divided into source, network and destination sub-problems while keeping in mind the overall performance. Figure 3-5 on page 68 describes the end-to-end QoS and its various components. It is assumed that the application can tolerate an end-to-end delay of $D_{\text{end-to-end}}$ and end-to-end loss of $\text{Loss}_{\text{end-to-end}}$. Note that the loss can be caused by buffer overflow at the sender, the network, and/or the receiver. In addition, any packets that violate the delay constraint will be considered lost. For a successful transmission, the end system should negotiate with the network acceptable delay and loss levels and allocate extra quota for its own buffer while ensuring not to exceed the overall end-to-end delay and loss values $D_{\text{end-to-end}}$ and $\text{Loss}_{\text{end-to-end}}$.

In the following subsections, the problems associated with the transmission of MPEG-4 video over ATM/ABR networks is formulated in the context of the network and the end-system domains.

Figure 3-5: End-to-End QoS

3.6.1 Network Domain Problems

In order to achieve this research objectives, the network should not exceed its quota from delay and loss. For the ABR service, the loss target is easily achieved given that the ABR service was originally designed for loss-sensitive applications. The ABR uses a CLP0 for all its cells which leads to a low value of CLR. The challenging part is guaranteeing the delay since it is not supported by the current ABR standard. In addition, due to the nature of the ABR service, its closed-loop feedback feature, and its non-trivial engineering, there was a need to study and investigate the engineering and performance of the ABR service in general. In summary, from the network's perspective regarding the ABR service, the following questions need to be answered:

- What are the important features that should be supported in an ABR switch?
- Can the Virtual Source / Virtual Destination (VS/VD) option in the ABR service be used to provide a faster response time?
- What is the performance of the ABR service in general in terms of utilization, queueing delay, congestion response time, and transient response time?
- Most important, can ABR support a bounded delay?

The answers to the above questions and a complete analysis to the performance of the ABR service are resolved in Chapter 4.

3.6.2 End-System Domain Problems

The end-system domain generally includes both the source and destination. However, in this research, a source architectural model is proposed where all of the intelligence is placed at the source. No changes to the ARB service, as defined by the current standard framework, are proposed in this research rendering the proposed design easier to implement. On the other hand, the destination is only considered in terms of the encountered delay in the De-jitter buffer. Hence, the source plays a very important role in fulfilling the required QoS. As a matter of fact, the network can not guarantee any sort of QoS unless the source sends the traffic according to the predefined traffic contract. Due to the bursty nature of video traffic and the unknown behavior of video traffic, designing a source architecture and implementing the appropriate traffic management techniques to ensure the required QoS is the most challenging part of this research.

Isolating the part of the original problem that concerns the source, Figure 3-6 on page 71 presents a simple isolated source model of the original problem that was depicted earlier in Figure 3-5 on page 68. The problem can be described as having an MPEG-4 coder which produces VOPs with variable size (VS_j) at fixed intervals ($1/VOP_Rate$), where j denotes the arrival time of VOP number j . The variation of the VOPs size leads to a variable input traffic rate entering the smoothing buffer. Rate λ_j is defined as the arrival rate due to the arrival of j^{th} VOP. λ_j is calculated as in EQ 3.4.

$$\lambda_j = \frac{VS_j}{VOP_time} \quad (EQ\ 3.4)$$

Figure 3-6: Simple Source Model

Figure 3-7: Source Rate versus Service Rate

On the other hand, the buffer is served with variable service rate μ_i which varies depending on the network feedback through RM-cells. The frequency of this variation is controlled by the N_{rm} parameter negotiated during the call setup of the ABR connection. Figure 3-7 on page 71 depicts the source rate and the channel rate with respect to time. The source rate is changed every $1/VOP_Rate$ second, whereas the channel rate changes with the arrival of B-RM cells.

In this model, the objective is to provide the maximum possible video quality. This objective can be translated into two sub-objectives. The first is to encode the video at the maximum possible quality which results in the maximum possible VOP size (VS) and hence the maximum possible rate λ i.e to maximize the utilization of available bandwidth. The second objective is to preserve the required QoS. The QoS for the source portion is set such that CLR equals to zero and the smoothing delay for j^{th} VOP d_j should be less than the precalculated value of D_{max} based on the application and the other delay component in the overall path. Consequently, the problem can be defined as follows:

$$\max_{\bar{\lambda}} \left(\rho = \frac{\sum_{j=1}^N \lambda_j}{N \times \sum_{i=1}^M \frac{\mu_i \times t_i}{T}} \right) \quad (EQ 3.5)$$

Subject to:

$$d_j \leq D_{max} \Rightarrow \lambda_j \leq [\mu'_j \times D_{max} - B_j] \times SVR \quad \forall j \quad (\text{EQ 3.6})$$

Buffer evolution is given by:

$$B_j = \max \left\{ \left[B_{j-1} + \frac{\lambda_{j-1}}{SVR} - \frac{\mu'_{j-1}}{SVR} \right], 0 \right\} \quad (\text{EQ 3.7})$$

where:

λ_j	Source Rate of the j^{th} VOP
B_j	Buffer occupancy just before arrival of j^{th} VOP
μ_i	Service rate as calculated from i^{th} B-RM cell
μ'_j	Average service rate for j^{th} VOP
D_j	Source Buffer delay for the j^{th} VOP
VS_j	VOP size for the j^{th} VOP
N	Total number of VOPs
M	Total number of B-RM cells
D_{max}	Maximum allowable VOP delay in the source buffer
T	Total time of the video session
t_i	Time between i^{th} and $(i+1)^{\text{th}}$ B-RM cell
SVR	Source VOP_Rate (number of VOP per sec.)
$ATVR$	Average Transmitted VOP_Rate
ρ	Average utilization

3.6.3 Problem Formulation for Adaptive Multi-layered MPEG-4 Encoder

In order to achieve the previously mentioned objectives, coordination and interaction between the encoder and the network is needed. Hence, the simple model of Figure 3-6 on page 71 cannot satisfy this study's requirements. In this section, the adaptive and multi-layered capabilities of the MPEG-4 encoder is used to build a model of the problem. Adaptive means the capability of the encoder to encode the VOPs at different quality levels based on a certain control signal. Multi-layered implies the capability to encode the VOPs into two or more layers; namely basic and enhanced layers. Basic layer VOPs should be always transmitted; however, the enhanced layers can be discarded as described earlier in Chapter 2. Therefore, another control signal is needed to control the transmission and the discard of enhanced layer VOPs. Figure 3-8 on page 75 depicts the new model for the MPEG-4 encoder that will be the base model for this research. It was assumed that the MPEG-4 encoder can encode each VOP at three different quality levels: low, medium and high. In addition, the encoder produces two types of layers: basic layer and an enhanced layer.

It is worth noting that this problem formulation is for illustrative purposes to show how the various parameters of the network and encoder are linked together.

Figure 3-8: Adaptive Multi-layered Source Model

The problem is now reformulated to find the optimal control policy that will maximize the quality of the transmitted video without violating the required QoS in terms of delay and loss.

The problem can therefore be formulated as follows:

Define:

- Matrix $[VS]_{i \times j}$ which contains the VOPs sizes such that:

$i \in [1,2,3]$ (i represent the quality level).

$j \in [1-N]$ (j represent the j^{th} VOP and N is the total number of VOPs).

VS_{1j} corresponds to low quality VOP size for the j^{th} VOP.

VS_{2j} corresponds to medium quality VOP size for the j^{th} VOP.

VS_{3j} corresponds to high quality VOP size for j^{th} VOP.

- Vector $[L]_j$ which indicates the layer type of j^{th} VOP such that:

$L_j = 1$ if j^{th} VOP belongs to the basic layer.

$L_j = 0$ if j^{th} VOP belongs to the enhanced layer.

- Vector $[\mu']_{1 \times j}$ contains the average service rate for the j^{th} VOP
- Vector $[B]_{1 \times j}$ contains the values of buffer occupancy at time just before j^{th} VOP
- Matrix $[C]_{3 \times j}$ contains the value of the control signals such that:

if $C_{1j} = 1$ send the j^{th} VOP with low quality
 $= 0$ Do not send

if	C_{2j}	$= 1$	send the j^{th} VOP with medium quality
		$= 0$	Do not send
if	C_{3j}	$= 1$	send the j^{th} VOP with high quality
		$= 0$	Do not send

Problem

The problem can be formulated into the following form:

$$\max \sum_{j=1}^N \sum_{i=1}^3 C_{ij} \times VS_{ij} \quad (\text{EQ 3.8})$$

Subject to:

$$\sum_{i=1}^3 C_{ij} = 1 \quad \text{if} \quad L_j = 1 \quad \forall j \quad (\text{EQ 3.9})$$

$$\sum_{i=1}^3 C_{ij} \leq 1 \quad \text{if} \quad L_j = 0 \quad \forall j \quad (\text{EQ 3.10})$$

$$\sum_{i=1}^3 C_{ij} \times VS_{ij} \leq \mu'_j \times D_{\max} - B_j \quad \forall j \quad (\text{EQ 3.11})$$

Where:

$$B_j = \text{Max} \left[\left(B_{j-1} + \sum_{i=1}^3 C_{ij-1} \times VS_{ij-1} - \frac{\mu'_{j-1}}{SVR} \right), 0 \right] \quad (\text{EQ 3.12})$$

$$B_1 = 0 \quad (\text{EQ 3.13})$$

The above noted problem can be solved using mixed linear and integer programming techniques. In order to do so the problem need to be rewritten into the following form:

$$\max \sum_{j=1}^N \left(\sum_{i=1}^3 C_{ij} \times VS_{ij} \right) - B_j \quad (\text{EQ 3.14})$$

Subject to:

$$\sum_{i=1}^3 C_{ij} = 1 \quad \text{if} \quad L_j = 1 \quad \forall j \quad (\text{EQ 3.15})$$

$$\sum_{i=1}^3 C_{ij} \leq 1 \quad \text{if} \quad L_j = 0 \quad \forall j \quad (\text{EQ 3.16})$$

$$\sum_{i=1}^3 C_{ij} \times VS_{ij} \leq \mu'_j \times D_{\max} - B_j \quad \forall j. \quad (\text{EQ 3.17})$$

$$B_j \geq B_{j-1} + \sum_{i=1}^3 C_{ij-1} \times VS_{ij-1} - \frac{\mu'_{j-1}}{SVR} \quad \forall j \quad (\text{EQ 3.18})$$

$$B_j \geq 0 \quad \forall j \quad (\text{EQ 3.19})$$

Solving the above noted problem provides the optimal control policy $C [3*N]$. Applying this control policy to the system results in the following optimal equations for the performance metrics:

$$\rho = \frac{\sum_{j=1}^N \sum_{i=1}^3 C_{ij} \times VS_{ij} \times SVR}{N \times \sum_{i=1}^M \frac{\mu_i \times t_i}{T}} \quad (\text{EQ 3.20})$$

$$ATVR = \frac{1}{N} \times \sum_{j=1}^N \sum_{i=1}^3 C_{ij} \times SVR \quad (\text{EQ 3.21})$$

$$SNR = \frac{\sum_{j=1}^N \sum_{i=1}^3 C_{ij} \times SNR_{ij}}{\sum_{j=1}^N \sum_{i=1}^3 C_{ij}} \quad (\text{EQ 3.22})$$

$$CLR = 0 \quad (EQ\ 3.23)$$

The above analysis demonstrates that an optimal solution to this problem can be found as long as the traffic is known in advance. However, when the traffic is not known beforehand, one needs to find an approach different from the analytical one. Consequently, this approach is based on designing an intelligent source architectural model and a set of algorithms that determine the control signals based on on-line measurements.

To solve the problems defined in this chapter, Chapter 4 presents the studies that focus on the network domain and the performance of ATM/ABR networks. Then in Chapter 5 the traffic characteristics of MPEG-4 test sequences is studied in addition to the study of the performance of the transmission of smoothed uncontrolled MPEG-4 video. The proposed source architectural model and algorithms needed to achieve an efficient transmission of MPEG-4 over ABR networks are introduced in Chapter 6. Last but not least, Chapter 7 evaluates the performance of the proposed architecture and algorithms using the test sequences of Chapter 5.

Chapter 4

Performance Evaluation of the ATM/ABR Service

The ATM/ABR service implements a flow control, which enables the efficient use of network resources. It can be used to provide ATM bearer services, or to carry other cell services such as ATM VBR/UBR as well as packet services such as Frame Relay and IP. There are many ATM switch designs, ABR service parameters, and deployment options to consider. Judicious choice of these options and parameters could have a significant impact on the ABR performance in various networking scenarios.

4.1 Introduction

ATM networks are capable of offering a wide range of services with different QoS according to the requirements of users and applications. One of these services, the Available-Bit-Rate (ABR), is the only ATM service implementing a closed-loop flow control mechanism through which the network informs the source of the rate it is capable of handling. A closed-loop flow control allows the efficient use of the link capacity while providing a near zero-loss for the data cells. The ABR service was originally designed for the transport of loss-sensitive data such as TCP and Frame-Relay traffic. Figure 4-1 on page 83 shows a common usage of the ABR service as a backbone service for various enterprise data traffic types.

The first generation of ATM switches deployed the Explicit Forward Congestion Indication (EFCI) flow control mechanism. The EFCI is a simple binary feedback scheme by which the switches detect congestion and convey it either by setting the EFCI bit in the data cell header, or by setting a Congestion Indication (CI) and the No Increase (NI) bits of the Resource Management (RM) cells (see Appendix B).

Second generation ATM switches implement the Explicit Rate (ER) flow control mechanism for the ABR service [Bonomi95]. Studies have shown that the ER schemes perform better than EFCI in terms of higher link utilization, shorter queue length and better fairness [Lai98a]. Second generation ATM switches also introduce an optional feature which enables the segmentation of the ABR control loop. This feature is supported by the Virtual Source/Virtual destination (VS/VD) concept thus enabling any node to act as a virtual destination (VD) which can terminate the control loop, while concurrently acting as a virtual source (VS) which can originate a new control loop. Designing the VS/VD feature is not a trivial task. Jain has shown that if the VS/VD is not properly designed, it can lead to an unstable condition, where the switch queues do not drain [Kaly98]. On the other hand, research has shown that the VS/VD feature can prove to be very beneficial in the case of large round trip delays (RTD) when a faster response time is needed [Goyal98]

Figure 4-1: ABR Service Application

Despite available research [Woo98] [Lai98b], the design and engineering of an optimized ABR service remains a complex task. There are too many parameters and deployment options to choose from. Furthermore, the optimal configuration often depends on network environmental factors (such as the number of administrative domains, security and round trip-delay), switch design features, traffic patterns and the network performance objectives.

4.2 ABR Engineering Considerations

Successful ABR deployment must take several factors into consideration, including network environment factors such as offered services, number of administrative domains, ATM switch capabilities, and control loop options. It is very important to understand the engineering capabilities of ABR in order to judge its suitability for a specific application. This section provides an overview of the ABR engineering considerations, including network environment, switch capabilities, and control loop options.

4.2.1 Network Environment

An ABR service can span one or more *administrative domains*. Implementations of the ABR service in private networks or in single-carrier networks are considered as one administrative domain. In the case of Virtual Private Networks (VPN), where an enterprise leases an ABR VPN from a carrier, there are two administrative domains. Multi-domains are commonly found in global services crossing several carriers. In networks spanning more than one administrative domain, security issues and segment performance tuning

(independent of other network segments) become critical. For example, policing and shaping features are often required at the boundaries where non-conformed cells may be discarded. Also the optimal ABR service parameters [Ohsaki98] are often different for the individual segments to match their network structure and performance objectives.

The services offered by the network may limit the deployment options of the ABR service. For example, in configurations where the ABR service is available all the way from source to destination, the end-to-end control loop option can be implemented. However, in cases where multiple services between the source and destination are involved (for example, Frame Relay or IP over ATM, as shown in Figure 4-1 on page 83), the ABR service can only be used for the portion of the network which can support it. To be able to support such a network configuration, the ATM switch must have the capability to act as a virtual source and virtual destination to originate/terminate the control loop and allow for internetworking with other services.

Network configuration such as topology, round-trip delay (RTD), number of users and applications, and traffic characteristics should also be considered in the ABR design, as they can have great impact on the final engineering decisions.

Finally, the network performance objectives should be defined precisely, as in many cases trade-offs need to be made between various objectives such as efficiency vs. loss vs. delay [Balak97].

4.2.2 ATM Switch Capabilities

The ATM Forum specifies the minimum functions that an ABR switch should support. It also specifies several optional features. Consequently, ABR switches may come with varying capabilities such as: EFCI, ER, VS/VD, per-vc queueing and traffic shaping. Many studies, including the present one indicate that a good and flexible ABR switch design is critically important. Furthermore, during the next several years, one would expect to encounter many networks that consist of a combination of first (EFCI-based) and second (ER-based) generation ATM switches. In those cases, the second generation switches should be deployed at the hot spots where congestion is most likely to occur (for example low speed access links or overloaded transatlantic trunks). The ABR switch design requirements are elaborated in Section 4.3, together with an example of a modern ABR switch design.

4.2.3 Control-Loop Options

Based on the network environment and the switches' capabilities, the network engineer should select the optimal ABR control-loop design. Normally, there are three options to choose from: end-to-end, edge-to-edge and hop-by-hop.

As shown in Figure 4-2 on page 88 (a), *end-to-end* ABR service deployment is the simplest ABR configuration. It consists of only one ABR control loop. The Forward Resource Management (F-RM) cell, originating from the source, travels through all intermediate switches until it reaches the destination, where it is turned around into a Back-

ward Resource Management (B-RM) cell and sent back to the source. The returning B-RM cell contains the allowed rate information for the entire loop.

Figure 4-2 on page 88 (b) shows an *edge-to-edge* deployment option consisting of three control loops. Its main advantage is the ability of segmenting the ABR loop into access and backbone portions, which allows the backbone network provider to control/engineer the backbone loop independently of the access portion of the network.

The extreme deployment case, shown in Figure 4-2 on page 88 (c), is the *hop-by-hop*, where the control loop is segmented into up to $N-1$ loops (where N is the number of network elements (including source, destination and switches)). In Section 4.4, we provide a performance evaluation and comparison of the various options.

Figure 4-2: ABR Control-Loop Options

4.3 ABR Switch Design

The performance of the various ABR service depends heavily on the switch design and features. A good and flexible ABR switch design enables the achievement of high efficiency and low cell loss, provides the flexibility to segment the control loop as required, and can control the delay experienced in the connection queue. The key features of the ABR switch include a robust ER algorithm with actual per-vc and link measurements and good smoothing functions, use-it-or loose-it policy, (weighted) fairness among users, large dynamic buffering and flexible VS/VD options.

This section describes the requirements for the ABR switch design by describing the ERSA+ (Explicit Rate Switch Algorithm) as an example of a second generation ABR switch, which is also used for the simulation model that is presented in Section 4.4.

Figure 4-3: ERSA+ Switch Algorithm

The ERSA+ Algorithm, as shown in Figure 4-3 on page 90, performs the following major calculations:

Input rate measurement per connection

Through measuring and averaging the cell interarrival time (*ait*), the averaging is done using the Exponential Moving Weighted Average (EWMA) [Floyd93], which acts as a low-pass filter. This prevents short term normal traffic fluctuations from altering the ABR rates significantly. Upon each cell arrival the switch performs the following calculations:

$$\Delta t = \text{CurrentClock} - \text{OldClock} \quad (\text{EQ 4.1})$$

$$\text{OldClock} = \text{CurrentClock} \quad (\text{EQ 4.2})$$

$$\text{ait} = (1-\alpha) \times \text{ait} + \alpha \times \Delta t \quad (\text{EQ 4.3})$$

where α determines the time constant for the low pass filter

Availability calculations

Availability calculations seek to make the total elastic input bandwidth of all the connections equal to the total available elastic output bandwidth. In order to perform availability calculation, it is necessary to identify cell arrival and cell transmit opportunities, that constitute the elastic input and the available elastic output. A Generic Cell Rate Algorithm (GCRA) is used to determine which cell arrivals are in excess of MCR.

Fair Share calculations

Fair share calculations seek to adjust the elastic input rate of all connections relative to their weights. A piece-wise linear function is used to move connections quickly to their fair share while limiting any overshoot. It is also designed so that the connection must first use some of the additional offered bandwidth before subsequent increases are offered. This design prevents lazy or inactive connections from being offered a high bandwidth, which could cause undesirable transient behavior.

Congestion Calculations

Congestion calculations seek to protect resources from being depleted due to unusual transient behavior. The congestion factor is derived from the state of congested entities such as queues, links, buffer pools and switch fabric. This feature can also be used in order to control the delay in the connection queue, where a congestion factor can effectively reduce the explicit rate in the B-RM as the per-vc queue starts to build up.

Rate adjustments

Based on availability, congestion, and weight information, the ERSA+ algorithm adjusts the explicit rate and sends it back to the source through the B-RM cell.

Upon the reception of a B-RM cell, the source modifies the ACR (Allowed Cell Rate) as follows:

$$A = \text{Max}(\text{MCR}, \text{ER}) \quad (\text{EQ 4.4})$$

$$B = \text{Max}(\text{MCR}, \text{ACR} - \text{ACR} \cdot \text{RDF} \cdot \text{CI}) \quad (\text{EQ 4.5})$$

$$C = \text{Min}(\text{PCR}, \text{ACR} + \text{PCR} \cdot \text{RIF} \cdot \text{CI} \cdot \text{NI}) \quad (\text{EQ 4.6})$$

$$\text{ACR} = \text{Min}(A, B, C) \quad (\text{EQ 4.7})$$

where:

CI: is the Congestion Indication bit in the B-RM cell

NI: is the No Increase bit in the B-RM cell

RDF: is the Rate Decrease Factor parameter

RIF: is the Rate Increase Factor parameter.

Support of the VS/VD option requires the switch's support of per-vc queuing, and the ability to dynamically shape the connections traffic according to the calculated ACR. The shaper should include a "*recouping*" option, thus allowing a connection to recover some or all of the departure delay which may have been encountered during transient or congestion conditions.

The ERSA+ algorithm allows the switch to have different modes of operation depending on whether one chooses the option of one ABR control loop (ER mode) or multiple control loops (VS/VD mode). Through the rest of this chapter we use the ER mode to imply end-to-end and VS/VD to imply hop-by-hop loop configurations, as we switch context between switch and network control features. In the VS/VD mode, we assume that the switch implements all the source and destination behaviors as specified in the ATM Forum Traffic Management Specification Ver 4.0 [ATMF96].

As mentioned in the ABR introduction, the design of the ER switch algorithm is very critical for network performance. Some ER algorithms rely solely on the queue length or on the Current Cell Rate field (CCR) of the RM cell to estimate the traffic rate of the Virtual Connections (VCs). Such ER algorithms often lead to traffic oscillations, resulting in large cell losses and decreased network efficiency and fairness.

As noted, the ERSA+ algorithm achieves a good stability by calculating ER values for a connection based on the measured input rate. The ER is updated with the arrival of each backward RM cell. These calculations seek to adjust the input rates of the connections towards values that are fair, sustainable and yet utilize the link efficiently.

4.4 ABR Performance Evaluation

In this section the simulation results and performance analysis of various ABR configurations is presented. First, the performance metrics of interest to this study is described, followed by the network simulator model. Next, the simulation results and analysis for steady state, transient, and congestion conditions are presented.

4.4.1 Performance Metrics

The following metrics are used for the evaluation of ABR performance:

Link Utilization (LU_i)

LU_i evaluates the capacity of VS/VD and ER to fully utilize the link. Note, that for the purpose of this study, the ERSA+ has been tuned to offer a target utilization of 97%.

LU_i denotes utilization of link _{i} .

Maximum Queue Size (MQS_i)

MQS_i is the max queue size of the switch. An increase of MQS_i , as in the cases of transient or congested states, has a two-fold impact. In the case of a switch, which does not have enough buffering, some cells would be lost. If, on the other hand, the switch has enough buffering, an increase of MQS_i results in an increased delay. Throughout the simulation runs, the queue size is measured at steady, transient and congested states. MQS_i denotes the maximum queue size for switch _{i} .

Congestion Response Time (CRT)

CRT is the time taken until the network detects and reacts to a congestion condition and the reaction reaches the congested point. The larger the CRT is, the less effective is the network in controlling and recovering from congestion (and the larger the switch queue sizes become).

Transient Response Time (TRT)

TRT is the time needed for the network to reach the steady state following a change in the traffic or the capacity. The shorter the TRT, the better is the performance in terms of delay, loss, fairness and general network stability.

4.4.2 Simulation Model

A key objective of this study is comparing the performance of VS/VD (hop-by-hop) and ER (end-to-end) modes. To answer this question, an elaborate Bit-True Simulator for the ERSA+ algorithm was built which represents the exact switch implementation [Guirguis98]. The simulation model was built using the OPNET simulation tool of Mil3 corporation. Two models, Figure 4-4 on page 97 (a), Figure 4-4 on page 97 (b), were developed and were used to compare the two options for the steady, transient and congestion states. The simulation models assume three OC-3 access links that connect the three sources S1, S2 and S3 to Switch₁ (SW₁), an OC-3 trunk (Link₁) that interconnects SW₁ to Switch₂ (SW₂), and an OC-3 egress link (Link₂) that connects SW₂ to the three destinations D1, D2 and D3. The default round-trip delay is assumed to be 16ms. All three sources are greedy, always ready to send data according to the ABR rate control, have an OC-3 Peak Cell Rate (PCR) and a Minimum Cell Rate (MCR) set to zero. Table 4-1 on page 96 summarizes the key ABR parameters used in the simulation.

Table 4-1: ABR Parameters

Parameters	Value
Nrm	32
Rate Increased Factor (RIF)	1
Rate Decrease Factor (RDF)	1
Initial Cell Rate (ICR)	7000 cells/sec
Peak Cell Rate (PCR)	350,000 cell/s
Minimum Cell Rate (MCR)	0
Target utilization	97%

Figure 4-4: Simulation Model

4.4.3 Simulation Results

This section first presents the results for the steady state and transient conditions, assuming that the network is uncongested, defined as the condition when the network has enough buffering/capacity such that no cells loss would occur. Next, the simulation results and analysis under congestion condition, when some traffic gets discarded due to an increase in input traffic or to a decrease in the available bandwidth are presented.

Steady State

The steady state is defined as the state where the goals of efficiency and fairness have been achieved. One of the primary objectives of any scheme is its ability to converge to a steady state from any set of initial conditions provided that the demand and capacity remain constant.

Figure 4-6 on page 103 and Figure 4-7 on page 107 show that in both ER and VS/VD modes the connection rates converge to the target fair value of 50 Mbps/s, the link utilization achieves the high value of 97%, and the queue size stabilizes at the very small occupancy of 5 cells. It should be mentioned that traffic shaping could cause the VS/VD mode to achieve a smaller link utilization than the ER mode. This can be attributed to a delayed cell departure caused by contention or momentary overload, which causes the next cell to be scheduled based on the actual departure time if the ACR rate is not to be violated. This would result in under-utilization of the connection and the link bandwidth. This problem was solved in the ERSA+ implementation by *recouping* most of the depar-

ture delay. *Recouping* means that the departure delay, just experienced by the current cell, would be compensated for when scheduling the next cell. A programmable upper bound is placed on the amount of recouping in order not to violate the CDVT (Cell Delay Variation Tolerance) of the connection. Due to the use of recouping in our simulation both ER and VS/VD modes result in close link utilizations. Table 4-2 on page 99 summarizes these results for both modes.

Table 4-2: Summary of Results for the Steady State

Performance Metric	ER Mode	VS/VD Mode
MQS_1	5 cells	5 cells
MQS_2	3 cells	3 cells
LU_1	97%	97%
LU_2	97%	97%

Transient State

In general, the transient state can be defined as the state that the switch enters due to a change in demand or in resources until the switch is stabilized and reaches the steady state. A good scheme should converge quickly from transient to steady state. One should not underestimate the importance of transient analysis in large ATM/ABR networks. Much of the data applications traffic is bursty in nature, and many sources may join and leave the network every second. The expected increase in the volume and burstiness of compressed video traffic adds to the network stress, irrespective of whether the video is carried on the ABR or VBR service categories. In the latter case, the video traffic would

cause an increase in the ABR burstness indirectly, by reducing the aggregate ABR available capacity on the link.

To simulate the transient state, the three sources are started 100ms apart. The addition of each source changes the demand on the switch resources and causes the switch to enter the transient condition. At time zero when there is only one source, the fair share of this source should be 150Mbps, but when another source joins, the fair share should be reduced to 75Mbps for each source, and finally when the third source joins, the fair share should be 50Mbps each.

Analyzing the simulation results in the transient state (Figure 4-5 on page 102 and Figure 4-6 on page 103) one can observe that the VS/VD option results in a larger queue size and longer transient response times than the ER mode. This is due to the fact that in the VS/VD mode, the switch is required to shape the traffic according to the calculated ACR. Shaping means that for a given connection the time spacing between two successive cells should never be less than the reciprocal of its allowed cell rate. Note that when using traffic shaping, the queue buildup can no longer be an indication that the link is fully utilized. The link can be under utilized, but the queues build up due to the output restriction based on ACR. Note that both modes yield comparable performance in terms of link utilization because of the use of recouping in our simulation, as was described earlier. Table 4-3 on page 101 summarizes the simulation results for both ER and VS/VD modes in the transient state.

Table 4-3: Summary of Results for the Transient State

Performance Metric	ER Mode	VS/VD Mode
MQS_1	62 cells	250 cells
MQS_2	5 cells	10 cells
LU_1	85% - 100%	85% - 100%
LU_2	85% - 100%	85% - 100%
TRT	400 ms	430 ms

Figure 4-5: Simulation Results of ER Mode*Transmission of Multi-layered MPEG-4 Video over ATM/ABR Networks*

Figure 4-6: Simulation Results of VS/VD Mode

Transmission of Multi-layered MPEG-4 Video over ATM/ABR Networks

Congested State

A congestion state is entered when the switch starts to drop cells due to an increase in the demand over the available resources. In the case of ABR, the switch becomes more vulnerable to enter the congestion state when the Round Trip Delay (RTD) is large. This is due to the fact that it will take more time for the source to react to any decrease in the available bandwidth. To protect the ABR switch from entering the congestion state and to ensure zero-loss, the switch is required to allocate a buffer size according to the following equation:

$$\text{Required Buffer} = c \times \text{RTD} \times \text{Variation in Link Speed} \quad (\text{EQ 4.8})$$

where c is a multiplier greater than one depending on the speed of convergence of the deployed algorithm and on the ABR source parameters (RIF, RDF). Segmenting the ABR control-loop through the VS/VD mode provides a powerful technique to distribute the buffering requirements among several switches, as will be demonstrated below.

A congestion situation has been simulated by decreasing the capacity of link2 (between SW2 and destinations) from 150Mbps to 30Mbps. In real life, this situation could happen when higher priority CBR/VBR sources suddenly become active and grab most of the link bandwidth.

Figure 4-7 on page 107 (a) shows the results, assuming an RTD of 16ms. When the available bandwidth decreases the queue size in switch 2 starts to increase till it reaches a value of 4600 cells in the ER mode. This means that a buffer size of 4600 cells is required

to prevent cells from being dropped. Note that the queue size of switch 2 matches the theoretical value from EQ. 4.8.

$$\text{Required Buffer} = c \times 16 \times 10^{-3} \times (150-30) \times 10^6 / (53 \times 8) \quad (\text{EQ 4.9})$$

$$\text{Required Buffer} = c \times 4,528 \text{ cells.} \quad (\text{EQ 4.10})$$

This results in c very close to 1 which indicates the fast response of ERSA+ to changes in network conditions. In the case of VS/VD, we find that the buffer requirements are distributed between switch 1 and switch 2, resulting in 2300 cells each.

Figure 4-7 on page 107 (b) shows the results when increasing the RTD is increased to 40 ms. Once again, we find that the buffer requirement is 11,400 cells in switch 2 in the ER mode, while it equals 5700 cells in each of switch 1 and switch 2 in the VD/VD mode. Figure 4-7 on page 107 (c), and Table 4-4 on page 106 demonstrate the results in the case of a large RTD of 200ms. All these results confirm the buffer distribution benefits which can be achieved in the VS/VD mode. The results can be generalized to other network configurations and congestion conditions as follows. The ratio of the maximum buffer requirements of VS/VD over ER mode is linearly proportional to the values of the feedback delay in both cases. This delay is measured between the time the B-RM leaves the congested switch until it reaches the source/virtual source of the up-stream control-loop. This rule suggests that the VS/VD benefits are greatest in networks with large RTD, when congestion occurs downstream away from the real traffic sources.

Table 4-4: Summary of Results for the Congested State (RTD=200ms)

Performance Metric	ER Mode	VS/VD Mode
MQS_1 after congestion	5 cells	28,325 cells
MQS_2 after congestion	56,650 cells	28,325 cells
LU_1 before congestion	97%	97%
LU_1 after congestion	20%	20%
LU_2 before congestion	97%	97%
LU_2 after congestion	100%	100%
CRT	200 ms	100 ms
TRT	80 ms	90 ms

Figure 4-7: Queue Size in Congested State

4.5 ABR Network Delay

This section presents the analysis of the maximum end-to-end delay that can be experienced by an ATM cell in an ABR VC. Knowing the maximum per VC queue can provide us with the maximum delay in the vc queue. The same methodology used in determining the buffer size requirement in cases of congestion can be used to determine the maximum end-to-end delay.

A worst case scenario is a connection that is sending traffic with a rate equivalent to PCR and suddenly is forced to reduce its transmission rate to MCR. Note that with a good ER switch design this sudden reduction should be done gradually. However, we will assume this worst case scenario in order to get the worst case in terms of buffer requirements. Following the same logic and analysis as in the congestion case studied earlier we find that:

$$\text{Connection Queue} = \text{RTD drop in connection speed.} \quad (\text{EQ 4.11})$$

The above worst case scenario will result in a maximum connection queue according to EQ. 4.12 and maximum queueing delay according to EQ. 4.13.

$$\text{Maximum Connection Queue} = \text{RTD} * (\text{PCR} - \text{MCR}) \quad (\text{EQ 4.12})$$

$$\text{Maximum Queueing delay} = \text{RTD} * (\text{PCR} - \text{MCR}) / \text{MCR} \quad (\text{EQ 4.13})$$

The maximum network “ D_{network} ” can be calculated from EQ. 4.14 and it can be bounded by choosing PCR less than the value in EQ. 4.15.

$$\text{Max. } D_{\text{network}} = \text{Propagation delay} + \text{RTD} * (\text{PCR}-\text{MCR})/\text{MCR} \quad (\text{EQ 4.14})$$

$$\text{Max. PCR} = [(D - \text{RTD}/2) * \text{MCR}/\text{RTD}] + \text{MCR} \quad (\text{EQ 4.15})$$

Hence, by controlling the ratio between PCR and MCR, we can control the end_to_end delay. For example, if RTD were 16 ms and MCR were 256Kbps and we want the maximum D_{network} to be less than 56msec., then PCR should be less than or equal to 1Mbps.

According to the ATM Forum standard there are no constraints on the selection of PCR. However, if we were to impose such a constraint, we can guarantee a bounded delay for ABR, which is very critical in the case of video applications.

4.6 ABR Support for Video Applications

In order to support video applications, an ABR switch, in addition to the need to be appropriately engineered, should be capable of supporting an enhanced set of features. A summary of the ATM/ABR switch requirements to support video applications are:

- The ER switch should support Per-Vc queueing in order to be able to deploy the adequate delay control function and to be able to implement the VS/VD option if needed. Per-VC queueing provides a segregation between the ABR connections carrying data traffic and the ABR connections carrying video traffic.
- The ER switch should be capable of supporting MCR greater than zero. This allows for the guarantee of the minimum amount of resource allocation which is very critical

to video application. If the ER switch supports only an MCR of 0 then no guarantee will be offered, and it will not be considered suitable for video applications.

- The ER switch should be capable of supporting fairness criteria based on weighted allocation. The ATM Forum recommends several fairness criteria. However not all of them are suitable for video transmission. For video transmission, it is desirable to deploy a weighted allocation of bandwidth based on MCR where each connection is allocated its MCR and the elastic bandwidth is allocated to all the connections based on a weight equivalent to MCR. This criteria leads to several desirable behaviors. First, data will normally be assigned an MCR of 0, which guarantees that the bandwidth is allocated to a data connection only after all the video connection are allocated their needs differentiating between data and video traffic. Second, video connections with high bandwidth requirements will need both a higher bandwidth allocation and a higher minimum guarantee. The weighted fairness criteria supports these requirement by allocating more bandwidth to connections with higher MCR than to the other connections. Finally, it is expected that the ABR service will be priced based on the MCR, and hence from both the user and the carrier perspectives, it will be fair to allocate the elastic bandwidth based on what the customer will pay.
- Implementing a delay control function in the ER calculation.
- From an engineering perspective, the PCR to MCR ratio should be constrained based on the RTD and the acceptable network delay. The difference between PCR and MCR should decrease with the increase of RTD. The relation for PCR and MCR should fol-

low the model presented in Section 4.5.

- The VS/VD option proves to be beneficial in distributing the buffer requirements at the cost of adding extra delays due to shaping. As we want to minimize the delay when transmitting video applications, then the VS/VD option is not recommended. However, if the ER switch sets the ER value in the forward direction (F-RM), then the VS/VD will become very important in providing a fast feedback to the source. However, in the case that the ER switch sets the ER value in the backward direction, then the VS/VD is not recommended for the transmission of video applications. In short, if the ER value is set in F-RM, then the VS/VD option can prove to be beneficial in providing faster response time.

4.7 Conclusions

The ATM/ABR service can yield a high network utilization coupled with low network loss and a controlled delay. However, successful deployment of the ABR service requires a flexible ATM switch design which implements a robust measurement-based ER algorithm and the VS/VD control-loop segmentation option. In addition, the video traffic can be transmitted using the ABR service provided that the ER switch support the enhanced features mentioned in Section 4.6

In this chapter, it was assumed that the source should follow the source behavior as described in TM4.0 [ATMF96] However, this is not a trivial task for real time applications without any prior knowledge of the source traffic. As will be demonstrated in the

next chapter, for the MPEG-4 video traffic to follow the required behavior, a large smoothing buffer is needed which will violate the delay requirements for video applications. The problem is now shifted from the network domain to the end-system domain, and more specifically to the design of the source, which is the focus of the remainder of this thesis.

Chapter 5

Transmission of Uncontrolled Multi-layered MPEG-4 Video

The previous chapters presented the studies which focus on the network-related issues of this research which presented the performance studies of the ATM/ABR service and the required features to support video applications. This chapter shifts the focus of this thesis to the end-systems domain. The first part of this chapter covers the study of the nature of MPEG4 video traffic and the statistical characteristics of various MPEG4 test sequences. The second part presents the study of the performance of smoothed uncontrolled MPEG-4 video transmission over ATM/ABR and the various trade-offs involved. It is worth noting that this research focuses on channel rates between 128Kbps and 1Mbps as this is the range where MPEG-4 works very efficiently. Nonetheless, to my knowledge no previous research has been carried out for this channel rate range for VBR video due to its infeasibility with MPEG-1 and MPEG-2. In addition, the economic transmission rates most commonly available for end users are expected to fall within the channel range of this study.

5.1 Introduction

The MPEG-4 encoder has the capabilities to produce Constant Bit Rate (CBR) streams as well as Variable Bit Rate (VBR) streams. As was mentioned in Chapter 2, in

order to produce a CBR stream, the encoder uses an internal buffer and a rate control mechanism to adjust the encoding parameters. This leads to a variation in the encoded video quality. On the other hand, the VBR mode has the potential to provide a constant quality video, but imposes several challenges on the underlying transmission media due to its burstiness. In this research, the traffic characteristics of the VBR MPEG-4 video was studied.

The study of the traffic characteristics of the MPEG-4 video was motivated by the need for such information in order to design the capacity requirements of the underlying communications network. There is a wide variation in the volume of generated information when encoding the different video sources caused by the variation of the video content and the encoding parameters that are used. Teleconference video images are usually relatively static images of people where scene changes rarely occur. On the other hand, with broadcast media, there are no limitations on the content and scene changes may be frequent. This variation results in different traffic characteristics for each type of video as will be demonstrated in Section 5.3.

This research also investigates the extent to which the quality of the video signal is effectively degraded by the encoding process, and which can be quantitatively measured by the Signal-to-Noise Ratio (SNR). Note that SNR corresponds to the average power of the magnitude of pixel distortion caused by the encoding.

The SNR is calculated on a VOP basis according to the following equation:

$$\text{SNR for } n^{\text{th}} \text{ VOP} = \text{SNR}(n) = 10 \log_{10} \left(\frac{\sum_{i=1}^{N_p} (O_i(n))^2}{\sum_{i=1}^{N_p} (O_i(n) - d_i(n))^2} \right) \quad (\text{EQ 5.1})$$

Where:

$O_i(n)$: is the luminance value of the i^{th} pixel of the original n^{th} VOP.

$d_i(n)$: is the luminance value of the i^{th} pixel of the reconstructed n^{th} VOP.

N_p : is number of pixels in the VOP.

Three levels of encoding quality were considered, which were selected so that the difference of SNR between each level and the subsequent one is in the range of 1dB to 2 dB. The corresponding increase in the bit rate is on the average between 20% to 50% depending on the nature of the video sequence. The quantization indices (Q) 31, 20 and 15 were found to satisfy these preconditions. It should be noted, that normally the human eye does not perceive a difference in the quality if the SNR is within a 1dB range. Since this research focuses on the transmission at low bit rates, no indices beyond quantization level 15 were considered. Figure 5-1 on page 116 shows the relationship between the SNR, VOP size and the 31 levels of quantization provided by the MPEG-4 encoder for a sample VOP. It can be shown that the increase in VOP size is almost linear for Q between 15 and 31. However, the increase becomes exponential when Q decreases below 15. The VOP size increased from 10K bytes at a Q of 15 to over 120K bytes at a Q of 1. On the contrary, SNR increases with a much slower rate than the VOP size. Figure 5-1 on page 116 demonstrates that the SNR was equal to 28dB at a Q of 15 and increased to 42dB at a Q of 1.

Figure 5-1: VOP size and SNR vs. Quantization Index (Q)

5.2 MPEG-4 Source Modeling

Several studies investigated modeling of MPEG video traffic [Ni96] [Izquierdo98] [Rozić97] [Ohta94]. However, the typical approach to source modeling has been empirical in the sense that traces of the number of bits per VOP were analyzed and used to formulate source models that are valid for a class of sources. There has been little success in formulating source models that use the coding algorithm itself as a basis [Lakshman98]. Generally, source models have been useful for modeling video conferencing applications as they rely on a small set of physically meaningful parameters (the mean, variance, and one-lag correlation). However, adequate models for active sequences need many parameters given that there are three marginal distributions (scene lengths, scene change and intrascene), which can be useful only for stored video applications such as VoD. In addition, models based on a specific data set cannot be perfect when generalized. Since the focus of this research is on the transmission issues for real-time applications, where there is no prior information on the traffic, the most appropriate approach was to use several real MPEG-4 traces as a direct input to the ATM/ABR simulator developed for this study. Although, outside the scope of this thesis, yet for the interested researchers, the MPEG-4 trace information prepared for this research can be used to build multi-layered MPEG-4 traffic models using any of the well-known techniques such as Transform Expand Sample (TES) [Ismail95], Continuous-State Autogressive Markov Model, or Discrete-State Continuous-Time Markov process [Maglaris88].

5.3 MPEG-4 Video Test Sequences

The encoded VOP consists mainly of texture information and motion information. Movies with high level of motion activity are characterized by increased fluctuations in the VOP sizes due to the continuous scene changes which leads to bursty traffic. On the other hand, scenes with high level of spatial details are characterized by large amounts in texture coding and hence large IVOP sizes, but with lesser fluctuations of the VOP sizes for the same VOP type. Consequently, we should consider the level of both motion activity and spatial details when we classify video sequences. Four video sequences were chosen that represent the different types of video sources. Reference is made to the four test sequences as: movie, CNN, Soccer, and BBC. The movie sequence represents the video type with low spatial details and moderate motion activity, and this should result in maximum compression ratio and minimum transmission rate requirements. This sequence is represented by a 50 seconds trace from the movie "Ghost", which has one complete scene change and one partial scene change. The CNN sequence is characterized by a high level of spatial details, but no motion activity. This sequence can represent video conference applications and is represented by a 40 seconds trace from the CNN News, containing only one scene. The Soccer sequence is characterized by both a high level of spatial details and a high level of motion activity, and is represented by a 120 seconds trace of a soccer game. Finally, the BBC sequence is a hybrid type, which is represented by 90 seconds trace of the BBC news, and is characterized by no motion activity for a while before it switches to a video clip of news with high level of motion activity.

Table 5-1 on page 119 summarizes the classification details for the four test sequences and the corresponding compression ratios achieved for each type.

Table 5-1: Classification of the Test Sequences

Sequence	Spatial Detail	Motion Activity	Compression Ratio (Q=31)	Mean Rate (Kbps) B+E	Texture to Motion Ratio
Movie	L	M	302.8	91.3	1.9
CNN	H	L	157.94	175.05	8.52
Soccer	H	H	68.10	406	1.5
BBC	Hybrid	Hybrid	117.75	234.8	1.64

After obtaining the four test sequences in raw 4:2:0 yuv format, they were encoded into two layers as specified by the MPEG-4 standard. Setting the quantization index to 31, 20 and 15 provides three different levels of quality, and which are referred to in this thesis by low, medium and high quality levels. The quantization index 31 results in minimal VOP sizes and hence minimum transmission rate requirements that should be guaranteed by the network. The test sequences are coded with IBBPBBPBB sequence with temporal scalability type 1 as described in the standard [MPEG4-2202]. The VOP rate for the basic layer is 10 VOPs per second and the combination of both the basic and enhanced layer rate reaches 30 VOPs per second. All four sequences have a size of 360 by 240 pixels. The coding parameters that were applied to obtain the respective MPEG-4 test sequences are listed in Table 5-2 on page 120.

Table 5-2: MPEG-4 Coding Parameters for the Test Sequences

Parameter	Setting
Coding Sequence	IBBPBBPBBPBB
Basic Layer Rate	10 VOP per second
Basic & Enhanced Layer Rate	30 VOP per second
Scalability Type	Temporal Scalability type 1 (enh. layer is coded in BVOP)
Low Quality Level Quantization Index	31
Med. Quality Level Quantization Index	20
High Quality Level Quantization Index	15
Number of objects	one (rectangular based coding)
VOP size	360*240 pixels

Table 5-3 on page 121, Table 5-4 on page 122, Table 5-5 on page 122 and Table 5-6 on page 123 list the MPEG-4 video traffic statistics for the four sequences. Figure 5-2 on page 124, Figure 5-3 on page 125, Figure 5-4 on page 126, and Figure 5-5 on page 127 depict the VOP size and occurrence histograms for the four sequences. From these statistics and histograms the following can be observed: First, the resulting histograms for the soccer and the BBC sequences are more spaced out than those for the movie and the CNN sequences. This can be attributed to the fact that the fluctuations in the VOP size for the movie and the CNN sequences are mainly due to the different types of VOP (IVOP, PVOP and BVOP), and to the different nature of encoding related to each type. However, additionally for the soccer and the BBC sequences there are fluctuations triggered by the level of motion activity and scene changes. One can conclude that the prediction of the VOP size is much more difficult for the soccer and the BBC like video types than for the movie

and the CNN. Even though the CNN sequence has no scene changes, yet among the four sequences, it has the highest Peak-to-Mean Ratio (PMR). As will be demonstrated later in this chapter, the CNN sequence video type is the only type that can achieve reasonable efficiency just by smoothing the traffic contradicting with its high PMR. This can be explained by the fact that in spite of its high PMR, the changes in the VOP size have a periodic nature based on the period of the sequence type I, P and B. For this type of video, smoothing the traffic over a Group of Picture (GOP) duration can achieve high efficiency. In contrast the traffic bursts in the soccer and the BBC sequences are not periodic and are difficult to predict since they are based on scene changes which can occur at any time. In conclusion, although the statistics presented in the tables may give some insight on the nature of the types of video traffic, one cannot solely depend on them.

Table 5-3: Movie Sequence Statistics

	Quality Level 1		Quality Level 2		Quality Level 3	
	Basic	B&E	Basic	B&E	Basic	B&E
Mean VOP Size in cells	10.59	7.18	12.55	8.65	16.53	11.74
Max. VOP Size in cells	39.38	43.73	54.48	56.21	70.83	70.83
Min. VOP Size in cells	2.56	0.79	4	1.25	5.64	2.19
Peak-to-Mean Ratio (PMR)	3.72	6.09	4.34	6.49	4.28	6.03
Std. Deviation (VOP Size)	6.08	5.16	7.83	6.44	10.42	8.2
Mean Source Rate in cells/sec.	105.94	215.33	125.48	259.55	165.27	352.26
Mean Source Rate in Kbps	45	91.3	53.2	110	70	149.4

Table 5-4: CNN Sequence Statistics

	Quality Level 1		Quality Level 2		Quality Level 3	
	Basic	B&E	Basic	B&E	Basic	B&E
Mean VOP Size in cells	23.94	13.76	45.83	27.23	70.65	45.09
Max. VOP Size in cells	97.33	97.33	145.52	145.52	187.44	187.44
Min. VOP Size in cells	12.33	1.79	27.77	3.02	48.98	5.4
Peak-to-Mean Ratio (PMR)	4.06	7.07	3.17	5.34	2.65	4.16
Std. Deviation (VOP Size)	19.07	13.57	26.50	21.41	30.94	27.72
Mean Source Rate in cells/sec.	239.35	412.85	458.26	816.76	706.52	1,352.63
Mean Source Rate in Kbps	101.49	175.05	194.30	346.31	299.56	573.51

Table 5-5: Soccer Sequence Statistics

	Quality Level 1		Quality Level 2		Quality Level 3	
	Basic	B&E	Basic	B&E	Basic	B&E
Mean VOP Size in cells	37.46	31.92	58.82	50.85	82.16	72.06
Max. VOP Size in cells	120.48	120.48	181	188.54	237.79	262.06
Min. VOP Size in cells	6.92	2.1	13.12	3	23.3	6.42
Peak-to-Mean Ratio (PMR)	3.22	3.77	3.08	3.71	2.89	3.64
Std. Deviation (VOP Size)	17.97	19.72	26.43	29.88	35.26	40.03
Mean Source Rate in cells/sec.	374.53	957.55	588.2	1525.7	821.69	2161.79
Mean Source Rate in Kbps	158.8	406	249.4	646.9	348.4	916.6

Table 5-6: BBC Sequence Statistics

	Quality	Level 1	Quality	Level 2	Quality	Level 3
	Basic	B&E	Basic	B&E	Basic	B&E
Mean VOP Size in cells	22.71	18.46	31.18	25.97	43.17	36.25
Max. VOP Size in cells	65.08	79.59	94.35	119.29	149.08	158.77
Min. VOP Size in cells	8.58	2	9.29	2.02	15.44	4.48
Peak-to-Mean Ratio(PMR)	2.87	4.31	3.03	4.59	3.45	4.38
Std. Deviation (VOP Size)	10.66	12.4	14.81	17.64	21.41	23.79
Mean Source Rate (cells/sec.)	227.1	553.8	311.8	779.1	431.7	1087.5
Mean Source Rate (Kbps)	96.3	234.8	132.2	330.4	183	461

Figure 5-2: VOP Size and Occurrence Histogram for the Movie Sequence

Figure S-3: VOP Size and Occurrence Histogram for the CNN Sequence

Figure 5-4: VOP Size and Occurance Histogram for the Soccer Sequence

Figure 5-5: VOP Size and Occurrence Histogram for the BBC Sequence

Regarding the quality of the encoded video, although the four video sequences were encoded with the same fixed quantizer index Q , however the resulting SNR is different for each sequence and varies on a VOP basis. This is attributed to the fact that the distortion level may vary based on the content of the picture. We can see that the variation is small for the movie and the CNN sequences and increases for the soccer and BBC sequences. Once again, this is due to the level of motion activity involved in the soccer and the BBC sequences. We should note that if a variable quantizer index is used to obtain a fixed rate then the variation in SNR will be much larger.

The spatial details considerably affect the average SNR where CNN sequence with the highest level of spatial details has the lowest value of SNR. Table 5-7 on page 128 reports the average SNR values obtained for the four sequences. Figure 5-6 on page 129 demonstrates the resulting SNR with respect to time for the four video sequences.

Table 5-7: Mean SNR for Various Test Sequences

Sequence	Low Quality	Medium Quality	High Quality
Movie	31.4	32.9	34.4
CNN	24.32	26.56	28.31
Soccer	26.61	27.87	29.69
BBC	28.11	29.83	31.27

Figure 5-6: SNR for the MPEG-4 Test Sequences

5.4 Transmission of Smoothed Uncontrolled MPEG-4 Video

Given the bursty nature of the MPEG-4 traffic and the difficulties to accurately model traffic sources, smoothing the traffic before transmission has become of essence. The drawback of traffic smoothing is the extra delay it introduces which can violate the QoS requirements for video applications. In this section, the performance of uncontrolled smoothed MPEG-4 video is studied. The MPEG-4 video is uncontrolled in the sense that there is no control information sent to the encoder and the video is always encoded at a fixed quantization index. In other words, the smoothing buffer is linked to the User Network Interface (UNI), but not connected to the encoder. In this study, the encoder produces two layers of VBR-encoded bit streams. The two layers are then enqueued into the smoothing buffer and then served according to the ACR as shown in Figure 5-7 on page 132. In addition to smoothing the burstiness of the traffic, a source buffer is also needed to shape the traffic in order to comply with the traffic contract agreed upon by the user and the network operator. This is essential since the network polices the incoming traffic to ensure its conformance with the traffic contract and discards non-conforming cells. In our simulation model, we use a cell-spacing shaper to ensure that the cell spacing requirements are being satisfied.

The existence of the source buffer imposes a burstiness constraint on the source in order not to overflow the source buffer. We define the burstiness constraint $\text{Burst}(\tau)$ as the maximum amount of traffic that can be sent by the source during any interval τ . Hence, the source traffic should follow the following equations:

$$S(t_1, t_2) \leq \text{Burst}(t_2 - t_1) \quad \forall t_1 < t_2 \quad (\text{EQ 5.2})$$

Where $S(t_1, t_2)$ is the amount of traffic sent by the source during the interval $[t_1, t_2]$. The amount of burstiness allowed by a smoothing buffer depends on the source buffer size β and on the buffer service rate μ defined by the following equation:

$$\text{Burst}(\tau) = \beta + \mu \cdot \tau \quad (\text{EQ 5.3})$$

If the burstiness $\text{Burst}(\tau)$ exceeds the above calculated value, the buffer will overflow and cells will be dropped. Performance measures of interest are worst case delay $D_{\max}$, worst case queue size $Q_{\max}$, utilization ρ and cell loss ratio (CLR) due to buffer overflow. In order to evaluate these performance metrics, the ATM/ABR simulator of Chapter 4 was modified to accept the traffic from the MPEG-4 traces. Figure 5-7 on page 132 shows the simulation model used to study the behavior of smoothed uncontrolled MPEG-4 video. This chapter focuses on the first requirements of the research problem which is to guarantee the minimum level of quality through MCR. To this end, a congested network scenario is studied where only MCR is offered. First an infinite buffer capacity is assumed and the buffer occupancy for the four test sequences is compared under a normalized MCR of 1.33. Then the finite buffer case that allows a maximum smoothing delay of 150ms is examined. Any cell that experiences a delay more than 150ms will be dropped and considered lost. These sets of simulation experiments were also carried out with a normalized MCR of 1.25. Finally, the trade-offs between the MCR setting, smoothing delay and the cell loss ratio are studied.

Figure 5-7: Simulation Model

5.4.1 Infinite Smoothing Buffer

This section presents the results of the study conducted for the case of an infinite smoothing buffer. In this case, the CLR is equal to zero as the buffer never overflows and infinite burstiness is allowed. The performance metrics of interest are the buffer occupancy $Q(t)$, the maximum queue size Q_{max} and the corresponding smoothing delays $D(t)$ and D_{max} . The normalized MCR set to 1.33 resulted in an average utilization of 75% ($\rho = 0.75$). The normalized MCR is defined as the MCR divided by the mean source rate for the specific test sequence as indicated in Table 5-1 on page 119. A set of simulation experiments were carried out to observe these metrics for the different test sequences. Figure 5-8 on page 136 shows the results for $Q(t)$ and the corresponding $D(t)$ for the smoothing buffer of the four test sequences. The simulation results show that Q_{max} ranges from 102 cells for the movie sequence to 2947 cells for the soccer sequence which is translated into a D_{max} ranging from 180ms to 3.5seconds. It is obvious from the simulation results of Figure 5-8 on page 136 that, except for the CNN sequence, these values of smoothing delays are inadequate for video applications. This means that it is very difficult to achieve a 75% utilization for MPEG-4 video through traffic shaping only. Traffic smoothing alone can only be adequate for the CNN sequence type video which is characterized by low level of motion activity where there are no scene changes typical of video conference applications.

From Figure 5-8 on page 136, one can also deduce the level of motion activity for the various sequences through $Q(t)$ which starts to increase whenever motion activity is

present. For the movie sequence, there is a partial scene change at 19 seconds and a complete scene change at 33 seconds. These scene changes are reflected directly into burstiness in the traffic and hence a buffer fill-up. The same can be deduced for the other test sequences, whenever there is a scene change or high level of motion activity, the buffer queue starts to build up.

5.4.2 Finite Smoothing Buffer

Each type of application has specific QoS requirements that need to be satisfied. One of the most important QoS parameters for video applications are the end-to-end delay and the Cell Loss Ratio (CLR). Although smoothing the source traffic decreases bandwidth requirements, practically, the traffic smoothing cannot be performed over a long period of time, because this could result in an unacceptable delay as was demonstrated in Section 5.4.1.

In this study's simulations, it is assumed that the maximum smoothing delay is 150ms (i.e $D_{\max} = 150\text{ms}$). Based on these delay requirements and the buffer service rate (MCR), the buffer size is calculated and cells are dropped whenever the buffer overflows. It is worth noting that the normalized MCR, chosen to be equal to 1.25 of the average rate for the four sequences, led to an 80% utilization (i.e $\rho = 0.8$).

Figure 5-9 on page 137 shows the encountered delay $D(t)$ caused by traffic shaping and the cell losses $CL(t)$ caused by the buffer size limitation as imposed by the delay constraint. It is very interesting to analyze the nature of cell dropping in Figure 5-9 on

page 137. Although all four sequences suffer from a CLR in the order of 10^{-2} , they all differ in terms of cell dropping with respect to time. Each sequence has a completely different $CL(t)$. For the movie sequence, most of the cells were dropped at the scene change. For the CNN sequence there was continuous cell dropping which indicates that the CLR can be considerably enhanced with a little increase of $D_{\max}$ or MCR. However, for both the soccer and BBC sequences the cells are dropped in batches.

Figure 5-8: $Q(t)$ and $D(t)$ at $\rho = 0.75$ for Infinite Smoothing Buffer

Figure 5-9: D(t) and CL(t) at $\rho = 0.8$ and $D_{\max} = 150\text{ms}$

5.4.3 Setting of MCR

After studying the finite and infinite buffer cases, in this section an attempt is made to quantitatively establish the relation between MCR, CLR and average delay (D_{avg}) for a D_{max} of 150ms. Figure 5-10 on page 140 shows the relation between MCR and CLR. For all the video sequences, the CLR decreases with an increase of MCR until it reaches a CLR of zero. The same applies to D_{avg} , which decreases with an increase of MCR. However, the increase of MCR results in a lower bandwidth utilization and such a trade-off should be considered based on the application QoS requirements. It is worth mentioning here, that there are two aspects to the nature of encoded video that affect the transmission requirements differently. The first aspect is the level of spatial details in the encoded VOP. More spatial details result in higher average bit rates. The second aspect is the level of motion activity, which controls the amount of traffic burstiness. The results of Table 5-8 on page 139, show that although the CNN sequence required a higher average transmission rate than the movie sequence, yet a zero CLR with a 62% utilization was achieved with limited smoothing. In contrast for the movie sequence, this was achieved at 41% utilization.

Table 5-8: MCR vs. CLR and D_{avg} for $D_{max}=150ms$

Normal- ized MCR	ρ	Movie Seq.		CNN Seq.		Soccer Seq.		BBC Seq.	
		CLR	D_{avg}	CLR	D_{avg}	CLR	D_{avg}	CLR	D_{avg}
1.11	90%	1.04E-02	79.27	5.11E-02	48.96	13.8E-02	74.71	1.4E-01	81.57
1.25	80%	6.34E-02	62.85	2.03E-02	40.19	9.0E-02	60.39	9.08E-02	66.2
1.43	70%	3.25E-02	43.64	1.58E-02	34.69	4.92E-02	49.51	4.56E-02	53.2
1.61	62%	-	-	0	29.21	-	-	-	-
1.67	60%	2.09E-02	28.57	-	-	1.23E-03	34.95	8.63E-03	37.25
1.85	54%	-	-	-	-	-	-	0	21.78
2.0	50%	9.36E-03	21.91	-	-	1.71E-03	15.35	-	-
2.04	49%	-	-	-	-	0	14.25	-	-
2.44	41%	0	15.23	-	-	-	-	-	-

Figure 5-10: MCR vs. CLR and Average Delay (Max Delay=150ms)

5.5 Conclusion

One can conclude from the studies of this chapter that due to the burstiness of MPEG-4 video, traffic smoothing is essential. However, traffic smoothing introduces an extra delay which may be a critical issue with video application. In this chapter, the relationship between MCR, utilization and CLR for the various MPEG-4 test sequences was established quantitatively under a delay constraint condition. One conclude that, if the buffer size is restricted to satisfy the delay constraint, then the required MCR should be much higher than the average source rate. Nevertheless, this results in inefficient bandwidth reservation. The simulation results show that, to achieve a zero CLR, with the exception of the CNN sequence, the utilization will be less than 50% with a 150ms delay constraint. This leads to the conclusion that smoothing the traffic can be sufficient for video conference type applications yet for all other types more intelligent techniques should be deployed in order to achieve better utilization without violating the application QoS.

Chapter 6

Proposed Source Architectural Model

6.1 Introduction

Video traffic can be transmitted over two types of networks: *resource reservation networks* and *best-effort networks*. Resource reservation networks assume that the characteristics of the data-flow are known before-hand and that the network can reserve enough resources for the video session and consequently can guarantee the required QoS. Their major disadvantage is when users request only their minimum requirements, they will not get a better quality even when the network is under-utilized. On the other hand, if they request a better quality then this may block the admission of other users. In best-effort networks, no resource reservation takes place and accordingly no QOS guarantee is offered. The quality of the transmitted video depends on the traffic load in the network.

This thesis, proposes the transmission of MPEG-4 video over a third type of networks, namely *interactive networks*. Currently, the only implemented example of interactive networks are the ABR networks which allow for the bandwidth requirements to be split into hard requirements for guaranteeing minimum quality and elastic requirements for enhanced quality. However, this adds complexity to the design of the source architec-

ture. The source architecture is even further complicated due to the advanced feature of the MPEG-4 encoder which can produce multiple levels of quality and each level may consist of basic and enhanced layer VOPs. Hence, the need arises for a two-tier control system at the source, whereby the first tier controls the selection of the quality level and the second tier controls the transmission of the enhanced layer VOPs. Furthermore, in order to utilize the network bandwidth efficiently, the source should be capable of negotiating bandwidth requirements with the network and to react to changes in the network conditions. These requirements, in addition to the delay constraint imposed on the source, render the design of the source model a challenging problem. This chapter presents a proposed source architecture and the accompanying algorithms that are both network and application-aware and combine resource reservation and adaptation techniques. These features enable the proposed architecture to provide an efficient model for the transmission of MPEG-4 video over interactive networks with a minimum bandwidth guarantee.

6.2 Proposed Source Architectural Model

This section describes the proposed source architectural model that is capable of bridging the gap between application needs and network offerings. The basic idea is to make use of the MPEG-4 encoder to provide a network-aware application, where the application adapts to changes in the network conditions. Additionally, the proposed scheme makes use of ABR capabilities to provide an application-aware network, where the application and the network dynamically negotiate the required bandwidth.

Figure 6-1 on page 146 illustrates the interface requirements for the source, which interfaces with both the MPEG-4 encoder and the ABR network. The source accepts two streams of data from the MPEG-4 encoder, a basic layer stream and an enhanced layer stream. The source sends a control signal back to the encoder to set the quality level. At the network interface, the source transmits data according to the source behavior rules as defined in [ATMF96]. The source also periodically transmits forward-RM cells (F-RM) and receives backward-RM (B-RM) cells.

In resource reservation networks, traffic shaping is implemented at the source to smooth the source traffic and to conform to the reserved resources. With this study's interactive network, the implementation of a combination of rate and traffic shaping techniques is proposed. The traffic shaping is implemented via the adaptive traffic shaper algorithm (ATSA) with the objective to smooth the source traffic before entering the network, while the rate shaping is implemented via both QLSA and IVDA as will be described in details later.

With respect to traffic management, the proposed architecture implements both congestion control and congestion recovery mechanisms at the source level. The congestion control functionality is achieved through controlling the video quality level according to the available bandwidth. Nevertheless, due to the bursty nature of the traffic, a congestion may occur at the source buffer and the buffer can fill up and start dropping cells. Hence, it becomes very crucial to implement a congestion recovery mechanism in addition to the implemented congestion control mechanism. The congestion recovery mechanism,

that is implemented in the proposed architecture, is based on a novel intelligent VOP discard algorithm which will be applied to the enhanced layer VOPs. It should be noted, that one of the challenges of the source architecture is the discrepancy between what the source needs to control as required by the network (the rate) and what can be actually controlled by the MPEG-4 encoder. The MPEG-4 encoder enables the source to control the quality level (low, medium, high), but not the rate. Consequently, the source should implement several intelligent algorithms to be able to achieve this task.

The proposed architecture is depicted in Figure 6-2 on page 146 and consists of the following components:

- Intelligent VOP Discard Algorithm (IVDA)
- Quality Level Selection Algorithm (QLSA)
- Rate Request Algorithm (RRA)
- Adaptive Traffic Shaping Algorithm (ATSA)
- On-Line Measurements (OLM)
- Initialization
- Encoder State Tracking (EST)
- ABR Source Behavior

Throughout the rest of this chapter each component of the source architecture will be described in more details.

Figure 6-1: Source Interface

Figure 6-2: Source Architectural Model

6.3 VOP Discard Algorithms

Packet discard algorithms are commonly used in core networks to reduce congestion and increase goodput. They are normally implemented in the ATM switch to guarantee the transmission of whole AAL5 frames. However, in this study's proposed architecture, an intelligent VOP discard algorithm at the source is used. The deployment of an intelligent VOP discard algorithm is very crucial to the attainment of the required design objective. On one hand, it prevents the transmission of partial VOPs which may be dropped later at the destination after having consumed network resources. This ensures the transmission of whole enhanced layer VOPs without violation of the delay constraint. On the other hand, it provides the source with a congestion recovery mechanism which provides an immediate reaction to congestion created by changes in network conditions or arrivals of traffic bursts. In summary, VOP discard algorithms are buffer management techniques designed to ensure the protection of the basic layer VOPs during overload periods and to increase the goodput. In this section, the design objectives of the VOP discard algorithms are presented followed by the description of three different VOP discard algorithms and their performance evaluations.

6.3.1 Design Objectives

The design objectives of VOP discard algorithms can be summarized as follows:

Congestion Recovery Mechanism

The main objective of implementing a discard algorithm at the source buffer, is to provide the proposed source architecture with a congestion recovery mechanism. Congestion can occur at the source due to one or both of the following conditions: either the available bandwidth has been dropped leading to a buffer build-up and dropping of cells, or the arrival of a large VOP. In both cases, the source needs to be able to provide both congestion control and recovery mechanisms. Congestion recovery is provided through the implementation of a VOP discard algorithm.

Increases Goodput

The fragmentation of a VOP into cells may lead to the transmission of partial VOPs, which is not useful and is an unnecessary usage of bandwidth. The integrity of VOP transmission should be guaranteed at the source as a minimal requirement. This requirement will be achieved by implementing one of the proposed VOP discard algorithms and results in increasing the goodput.

Simplicity

The proposed algorithms should be as simple and as efficient as possible. Although there is a known trade-off between simplicity and overall performance, the general requirement is that the algorithms should not depend on complicated parameters which need to be differently tuned for each video session. The algorithms should perform well with the same parameter setting for all various types of videos. The performance of the proposed algorithms will be evaluated with the same parameter tuning for the MPEG-4 video sequences with different traffic characteristics.

Adaptive

Due to the continuous changes in network conditions and in the available bandwidth, the proposed VOP discard algorithm should be able to respond to these changes by changing the admission criterion. The initial setting of the threshold should be dynamically changed in response to changes in the available bandwidth.

6.3.2 Adaptive Early VOP Discard (AEVD)

The AEVD is based on the EPD algorithm introduced in 1995 [Romanow95]. With the EPD, the ATM switch drops an entire packet prior to the overflow of the buffer using a pre-set threshold as an indication to a possible overflow. When the threshold is reached, the switch drops the first arriving cells and all subsequent cells within the same VOP.

The AEVD is very similar in its operation to the EPD, however, it is used for a different function. It is implemented at the source to control the discarding of enhanced layer VOPs. Whenever the encoder produces an enhanced layer VOP which is ready to be enqueued, the queue controller decides whether to accept it, or discard it based on a specific threshold in a manner similar to the EPD. The major difference is the need to be able to dynamically change the threshold values with changes in the queue service rate (ACR). Figure 6-3 on page 150 describes the buffer dynamics for the AEVD where both the maximum buffer size and the threshold are functions of the buffer service rate (ACR). Figure 6-4 on page 150 lists the pseudo code for AEVD.

Figure 6-3: Buffer Model in AEVD

Algorithm AEVD:

Adaptive Early VOP Discard Algorithm is based on EPD algorithm introduced by Romanow [Romanow95]. Contrary to EPD, which has a static discard threshold, the discard threshold of AEVD is adaptively changed based on network information.

Step 1: Initialization:

- Set maximum allowable delay D_{max}
- Set initial ACR to ICR (Initial Cell Rate)
- Set Maximum Buffer Size $B_{max} = D_{max} \times ICR$
- Set Threshold_{AEVD} to $l \times ICR \times \frac{1}{VOPrate}$ (l is a constant)

Step 2: Admission decision:

For each enhanced layer VOP
 If $B(n) < Threshold_{AEVD}$
 accept the VOP
 else
 discard
 endif

Step 3: Adjusting discard threshold:

With the arrival of B-RM cell
 If $ACR_{new} \neq ACR_{old}$
 $B_{max} = D_{max} \times ACR_{new}$
 $Threshold_{AEVD} = l \times ACR \times \frac{1}{VOPrate}$
 endif

Figure 6-4: AEVD Algorithm

6.3.3 Modified Adaptive Early VOP Discard (M-AEVD)

A simple modification is added to the AEVD which enhances the performance of the VOP discard algorithm as will be seen from the performance evaluation in Section 6.3.5. The idea is based on the fact that from the network's perspective, it is more appealing in cases of congestion to discard the larger size VOPs and to favour the transmission of smaller size VOPs. Since the VOP discard algorithm applies to enhanced layer VOPs, all of type BVOP, then it is adequate to favor the transmission of smaller VOPs. This desired behavior of the M-AEVD has been achieved by modifying the basic AEVD to include the size of the current VOP in its decision. In the model implementation, the AEVD threshold was set to a buffer size equivalent to 1 VOP_time. In the M-AEVD, this threshold is increased to $(l+1)$ VOP_time and compared to sum of the buffer occupancy and the current VOP size before deciding to accept or reject the VOP. The pseudo-code for M-AEVD is listed in Figure 6-5 on page 152.

Algorithm M-AEVD:

Modified Adaptive Early Packet Drop Algorithm considers the current VOP size in the admission process in addition to the buffer occupancy and the discard thresholds. M-AEVD favours the transmission of smaller VOPs.

Step 1: Initialization:

- Set maximum allowable delay D_{max}
- Set initial ACR to ICR (Initial Cell Rate)
- Set Maximum Buffer Size $B_{max} = D_{max} \times ICR$
- Set Threshold_{MAEVD} to $(l+1) \times ICR \times \frac{1}{VOPrate}$ (l is a constant)

Step 2: Admission decision:

For each enhanced layer VOP
 If $B(n) + VOP_Size(n) < Threshold_{MAEVD}$
 accept the VOP
 else discard
 endif

Step 3: Adjusting discard threshold:

With the arrival of B-RM cell
 If $ACR_{new} < > ACR_{old}$
 $B_{max} = D_{max} \times ACR_{new}$
 $Threshold_{MAEVD} = (l+1) \times ACR \times \frac{1}{VOPrate}$
 endif

Figure 6-5: M-AEVD Algorithm

6.3.4 Intelligent VOP Discard Algorithm (IVDA)

The IVDA is an adaptive network-aware and application-aware VOP discard algorithm, which makes use of the knowledge of the current network state, the encoding sequence, nature of the various VOP types, and the buffer occupancy in its operation. Due to the built-in intelligence of the IVDA, it can be tuned to yield a superior CLR equivalent to the CLR obtained when transmitting only basic layer VOPs and discarding all the enhanced layer VOPs. The IVDA provides a flexible mechanism that can be tuned to provide various degrees of protection. The algorithm has three tunable risk parameters, which are separately applied to the three VOP types (IVOP, PVOP, BVOP). By increasing the risk parameters, better performance is obtained in terms of matching efficiency (ME) and the percentage the discarded VOPs at the risk of a higher CLR for the corresponding VOP type. Figure 6-6 on page 154 depicts the information that is used by the IVDA to make its admission decision.

Figure 6-6: Intelligent VOP Dropping Algorithm

The operation of the IVDA can be summarized as follows:

- Basic layer VOPs are always accepted.
- The delay experienced by the enhanced layer VOP is calculated based on the current ACR, VOP size and the current buffer occupancy. The VOP will be discarded early if it is expected to experience a delay greater than $D_{\max}$ before being completely transmitted.
- The enhanced layer VOP presents an opportunity to drain the source buffer, i.e. never admit an enhanced layer VOP if it will result in a queue build-up before the arrival of the next VOP.
- The basic layer VOP should be always protected. An enhanced layer VOP should not be accepted if it can result in the cells of future basic VOPs being dropped. Knowing the current state and the coding sequence, the enhanced layer VOP will only be accepted if it can be completely transmitted before the next basic layer VOP.
- If we observe the number of cells in the source buffer controlled by IVDA, we will observe a cyclic behavior in which the number of cells in the buffer rises during the transmission of basic layer VOP and then the IVDA forces the buffer to drain during the transmission of enhanced layer VOP. The IVDA is expected to yield the buffer behavior shown in Figure 6-7 on page 156.

Figure 6-7: Buffer Behavior under IVDA

- The IVDA does not directly apply the concept of thresholds. However, the admission decision of an enhanced layer VOP can be viewed as being based on a buffer threshold that becomes in effect at the arrival of the next basic layer VOP. IVDA refers to these three thresholds as *I_risk_level*, *P_risk_level*, and *B_risk_level* to differentiate them from the thresholds that are used in AEVD and M-AEVD. This naming reflects the fact that the larger the risk parameter the larger the probability of a higher CLR. If any of the risk level thresholds is set to zero, this enforces the admission of the enhanced layer VOPs only if they will be completely transmitted by the arrival of the next basic layer VOP belonging to the VOP type affected by the respective risk level threshold. This will result in more frequent discards of enhanced layer VOPs, but with the advantage of granting full protection to basic layer VOPs. This yields a CLR equal to the case of sending only basic layer VOPs as will be demonstrated by the simulation results in the next section.

Algorithm IVDA:

Intelligent VOP Discard Algorithm makes use of the network information, buffer state, current VOP size and the encoder state in order to decide on the admission of enhanced layer VOPs. IVDA introduces three configurable protection levels for basic layer VOPs which can be applied to the three different VOP types (I,P,B). The strictest protection configuration in the transmission of enhanced layer VOPs is only permitted if the VOP will be completely transmitted before the arrival of the next basic layer VOP. The other possible configuration allows a buffer occupancy equivalent to fractions or (multiples) of VOP_time.

Step 1: Initialization:

- Set maximum allowable delay D_{max}
- Set initial ACR to ICR (Initial Cell Rate)
- Set Maximum Buffer Size
- Set I_risk_level to zero $B_{max} = D_{max} \times ACR$
/* setting risk_level to zero provides the strictest protection level as it only admits enhanced layer VOPs if the buffer occupancy will be equal to zero at the arrival of next basic layer VOP*/
- Set P_risk_level to zero
- Set L_risk_level to zero
- Initialize the encoder state diagram

Step 2: Calculating cell budget until the arrival time of the next basic layer VOP:

For each enhanced layer VOP

$cell_budget = ACR \times time_to_next_BasicVOP - B(n)$

/* cell budget is calculated based on a zero buffer occupancy at the arrival time of next basic layer VOP*/

Step 3: Modify cell budget based on encoder state and risk level

if (next basic layer VOP is IVOP)

$cell_budget = cell_budget + I_risk_level$

if (next basic layer VOP is PVOP)

$cell_budget = cell_budget + P_risk_level$

if (next basic layer VOP is BVOP)

$cell_budget = cell_budget + B_risk_level$

Step 4: Admission decision

if ($(VOP_size < cell_budget) \&\& (B(n) + VOP_size < B_{max})$)

admit enhanced layer VOP

else

discard enhanced layer VOP

Figure 6-8: IVDA Algorithm

6.3.5 Performance Study of the VOP Discard Algorithms

In this section, the IVDA, AEVD and M-AEVD are evaluated in terms of performance. The performance of the three algorithms is also compared with the performance of the two extreme mechanisms which discard all the enhanced layer VOPs (discard all), and the mechanism that transmits all VOPs without any control (uncontrolled). The performance of the five techniques is evaluated with respect to CLR, matching efficiency (ME), discarded enhanced layer VOPs, and average smoothing delay at the source. It should be noted, that there is always a trade-off between these four performance metrics. For example, if we transmit all VOPs without any VOP discards then this results in a minimal value of VOP discards (zero), and a maximum value of matching efficiency. However, both CLR and the average delay will be at the maximum values. It should be emphasized that these performance metrics need to be carefully analyzed. For instance, in the case of uncontrolled traffic, the number of discarded VOPs at the source is zero. However this is accompanied with a high CLR, which leads to several corrupted VOPs that will be eventually discarded at the destination. The same may be applied to the matching efficiency where in cases of uncontrolled traffic the matching efficiency will be the highest of all the other algorithms but once again the goodput may not be as good as the matching efficiency due to the transmission of partial VOPs which will not be useful and will be discarded later.

In order to evaluate the performance of the various VOP discard algorithms under congestion conditions, this study's simulation experiments are set such that the network will always be congested, and the offered ACR is always equal to MCR. The various tech-

niques are evaluated under different MCR settings. As already mentioned, one of the main design criteria is simplicity so the three algorithms are evaluated with the same setting for all the four video traces. Additionally, the IVDA is set to provide the maximum protection by setting all risk level thresholds to zero. This setting will provide the best CLR that can be offered at the cost of more discarded VOPs and smaller matching efficiency.

After describing the simulation environment and settings the simulation results for each of the MPEG-4 video traces are now analyzed. Figure 6-9 on page 164 shows the simulation results for the movie trace. First, the cell loss ratio, shown in Figure 6-9 on page 164(a), is examined. The IVDA performs exactly as in the transmission of only basic layer VOPs, yielding the best performance in terms of CLR. Following the IVDA is the M-AEVD which outperforms the AEVD. It is obvious that in cases where no VOP discard algorithms are applied, the CLR will be significantly higher. Figure 6-9 on page 164(b) illustrates the matching efficiency, the performance of all the VOP discard mechanisms is similar except for the IVDA which yields a lower matching efficiency as it discards more enhanced layer VOPs. However, at the operating range of interest to this study, where zero CLR is achieved, the IVDA outperforms the other mechanisms. Figure 6-9 on page 164(c) shows the performance with respect to the discarded VOPs and as expected the IVDA protects the basic layer VOPs and achieves a good CLR by discarding enhanced layer VOPs. However, the percentage of the discarded VOPs is still within a very reasonable range as it varies between 16% to less than 2%. Finally, Figure 6-9 on page 164(d) depicts the results for the average smoothing delay, which shows that the IVDA outperforms all of the other

techniques.

It is clear that for the specific channel bandwidth, the IVDA outperforms the other algorithms in terms of CLR and average smoothing delay at the cost of a higher percentage of discarded VOPs and a lower matching efficiency. These simulation results can be analyzed in a different ways which may give us more insight on the performance of these algorithms. Instead of evaluating these techniques at specific channel rates, we need to evaluate for each technique the minimum bandwidth that is needed to be reserved in order to achieve a zero loss ratio. From both the network and the user's perspectives, it is very appealing if we can reserve less bandwidth while still being able to guarantee the required QoS. The tables, Table 6-1 on page 168 through Table 6-4 on page 168, provide this evaluation view for the four traces. The normalized MCR represent the minimum bandwidth that needs to be reserved to achieve a zero loss. Note that the lower the normalized MCR the smaller the reserved bandwidth, i.e. the more calls can be admitted. The results for the movie trace reported in Table 6-1 on page 168 demonstrate that in the case of the IVDA we need to reserve 1.17 times the average source rate (including basic and enhanced layer VOPs) to achieve a zero CLR as opposed to 2.28 when compared to the uncontrolled case or to 1.56 as in the cases of AEVD and M_AEVD. Generally, this great gain is achieved at the cost of discarding VOPs. In this case, the percentage of discarded VOPs is 11.45%, which is considered to be very reasonable. Finally, we can see that with IVDA we are able to achieve zero CLR while having a matching efficiency of 72.25% as opposed to 43.9% for the uncontrolled case and approx. 62% for both AEVD and M-

AEVD.

Figure 6-10 on page 165, Figure 6-11 on page 166, and Figure 6-12 on page 167 present the simulation results for the soccer, BBC, and CNN traces respectively. Although all the algorithms have the same parameter settings for all the MPEG-4 traces which have completely different characteristics, yet the results show a common trend in achieving a zero cell loss ratio. This trend can be summarized as follows:

Required Bandwidth Reservation

The IVDA requires the least bandwidth reservation among all the other mechanisms required to achieve zero loss for all the four traces. Comparing the IVDA with the uncontrolled case, the IVDA's bandwidth requirements are less by 58% for the soccer trace, by 55% for the BBC trace, by 49% for the movie trace, and by 3.1% for the CNN trace. Comparing the IVDA with M-AEVD, IVDA provides a reduction of the minimum required bandwidth by 25% for the movie trace, by 21% for the BBC trace, by 11% for the soccer trace and by 2.4% for the CNN trace.

Average Delay

The IVDA outperforms AEVD and M-AEVD in terms of average delay for all the traces with the exception of the movie trace. However, the difference is small in all the cases.

Matching Efficiency (ME)

The IVDA outperforms AEVD and M-AEVD in terms of ME for all the traces except for the soccer trace. Once again the difference is minimal.

Percentage of Discarded VOPs

The IVDA achieves its enhanced performance at zero loss at the cost of discarding enhanced layer VOP. Comparing the IVDA with the uncontrolled case, the percentage of discarded enhanced layer VOPs is 26.66% for the BBC trace, 25.64% for the soccer trace, 11.45% for the movie trace and 7.16% for the CNN trace. Comparing the IVDA with M-AEVD, the excess percentage of discarded enhanced layer VOPs by IVDA is 17.15% for the BBC trace, 12.54% for the soccer trace, 10.32% for the movie trace and 3.25% for the CNN traces.

In summary, the IVDA requires less bandwidth reservation at the cost of a higher rate of VOP discards. In addition, the IVDA parameters are much easier to configure given that the IVDA can always be configured to operate in the most conservative mode if it is strictly important not to loose any basic layer VOPs. With AEVD and M-AEVD, such a threshold setting is not a trivial task. Although this evaluation of the VOP discard algorithms under congested network conditions, where only MCR is provided, gives us only a mere insight into their behavior, our ultimate goal remains to evaluate their performance as part of the proposed architecture in a dynamic environment, where network conditions change affecting the allowed cell rate (ACR). This goal is fulfilled by the performance study presented in Chapter 7.

Figure 6-9: Comparison of VOP Discard Algorithms (movie trace)

Figure 6-10: Comparison of VOP Discard Algorithms (soccer trace)

Figure 6-11: Comparison of Various Dropping Techniques (BBC trace)

(a) Cell Loss Ratio

(b) Matching Efficiency

(c) Dropped VOPs

(d) Average Smoothing Delay

Figure 6-12: Comparison of VOP Discard Algorithms (CNN trace)

Table 6-1: Performance Results at CLR=0 (movie trace)

Metric	Uncontrolled	Discard All	AEVD	M-AEVD	IVDA
Normalized MCR	2.28	1.17	1.56	1.56	1.17
ME	43.9%	NA	62.5%	62.2%	72.5%
Discarded VOPs	0	66%	1.26%	1.13%	11.45%
Average Delay (ms)	21.01	NA	25.34	24.64	28.4

Table 6-2: Performance Results at CLR=0 (soccer trace)

Metric	Uncontrolled	Discard All	AEVD	M-AEVD	IVDA
Normalized MCR	1.93	0.82	1.03	0.92	0.82
ME	51.8%	NA	80.6%	83.7%	79.2%
Discarded VOPs	0	66%	12.24%	13.1%	25.64%
Average Delay (ms)	17.2	NA	54.46	49.07	36.6

Table 6-3: Performance Results at CLR=0 (BBC trace)

Metric	Uncontrolled	Drop All	AEVD	M-AEVD	IVDA
Normalized MCR	1.77		1.08	1	0.79
ME	56.5%	NA	78.8%	81%	81.2%
Discarded VOPs	0	66%	10.62%	9.51%	26.66%
Average Delay (ms)	25.81	NA	52.89	44.31	39.4

Table 6-4: Performance Results at CLR=0 (CNN trace)

Metric	Uncontrolled	Discard All	AEVD	M-AEVD	IVDA
Normalized MCR	1.28	1.24	1.27	1.27	1.24
ME	79%	NA	76.3%	77%	77.1%
Discarded VOPs	0	66%	6%	3.91%	7.16%
Average Delay (ms)	42.55	NA	35.36	36.59	35.45

6.4 Quality Level Selection Algorithm (QLSA)

The main functionality of the QLSA is to try to maximize ME by controlling the Quality Level (QL). The QLSA can be viewed as a rate controller (RC). However, there are some different requirements that distinguish the QLSA from a normal RC which can be summarized as follows:

- According to the traditional control theory, it is normally assumed that we can control the output rate of the coder so that a zero error between the target rate and the output rate can be achieved. In this case, we cannot claim that we have full control over the rate as we are only able to control the quality level selection which has an implication on the output rate but in a very coarse and indeterministic manner.
- Another challenge is introduced by the bursty nature of the traffic whereby it can be normally expected that the size of an IVOP from the low quality level is larger than those of BVOP of the high quality level. So we cannot assume that increasing the quality level will result in an immediate increase in the source rate. The increase will be achieved on the average over several VOPs.
- The definition of stability is not trivial in this problem. As with the continuous fluctuations of the input traffic it is not expected that the system will reach a steady state. An adequate definition for stability in our problem should be related to the QL stability and not to the source rate stability.

The QLSA is the core component of the proposed architecture, and which ensures

the attainment of the overall operation objectives. In order to achieve such, the QLSA depends on information from the network, the encoder, the application, and from the other components in the source. Figure 6-13 on page 173 illustrates the QLSA interface. The input to the QLSA can be categorized as follows:

From the Network (through B-RM cells and source behavior calculations):

- Allowed Cell Rate (ACR)
- Congestion Indication (CI)

From on-line measurements:

- EWMA_ME
- Number of cells dropped in short cycle
- Number of VOPs discarded in short cycle
- Number of cells dropped in long cycle
- Number of VOPs discarded in long cycle

From Initialization:

- Time constant
- Maximum target ME
- PCR
- Acceptable VOP discarded range

From the Encoder State Tracking:

- Current encoder state

Input to RRA from QLSA:

- Quality level increase
- Last increase

The pseudo-code for the operation of the QLSA is listed in Figure 6-14 on page 174. It should be noted, that the QLSA does not depend on a minimum value of ME to increase the quality level (QL). The purpose of this design is to avoid the aggregation of the problem in the case of any inaccuracy of RRA. In addition, the acceptable value of ME depends on the type of video and burstiness level of the traffic. In order to be independent of the video type and to avoid a different configuration for different streams, QLSA mainly depends on the percentage of discarded enhanced layer VOPs and the knowledge of the network status as a measure to increase the quality. The response time of the QLSA to various events differs greatly based on the importance of the event. For example, it is desirable to decrease the quality level faster than to increase it. Hence, the QLSA defines a *QL increase opportunity* and a *QL decrease opportunity*, whereby the QL level can be increased or decreased only at these opportunities. The increase opportunity is defined to be immediately after an IVOP with two consecutive increases that should be apart by a least T seconds (where T is a QLSA parameter) to avoid unnecessary fluctuations between QLs. The rationale for choosing a QL increase after an IVOP is attributed to the fact that an IVOP is normally larger in size than a PVOP or a BVOP while being the most important VOP that needs protection. Hence, if the QL increase is performed any time before an

IVOP then there may be a higher risk of dropping IVOP cells if the wrong decision was taken. One must not forget that the objective of the QLSA is not to provide an immediate reaction to bandwidth availability but to gradually with time reach the QL that is capable of consuming the available bandwidth.

With respect to QL decrease, it is more important to respond more frequently to avoid dropping cells. Therefore, the decrease opportunity is offered just before a PVOP or an IVOP. Once again, we decrease the quality before a PVOP or an IVOP in order to protect them from being dropped. Decreasing the quality level is normally associated with signs of congestion where the risk of dropping cells starts to be high. A QL decrease opportunity was not defined with every VOP to avoid too many QL fluctuations and given that there is enough enhanced layer VOPs which can be discarded first before decreasing the QL. This feature provides the QLSA with the ability to absorb any sudden glitches in the bandwidth without fluctuating the QL and to provide a graceful degradation if the congestion condition persists.

Figure 6-13: QLSA Interface

Algorithm QLSA:

Quality Level Selection Algorithm (QLSA) controls the increase/decrease of MPEG-4 encoder quality level. It depends on information from the network, encoder state and on-line measurements to take the required decision. The QLSA works in coordination with RRA.

Step 1: Initialization:

```

Set T
/*T determines the minimum time between two consecutive increases*/
Set Max_ME
/* Maximum ME above which there will a greater risk of losing cells*/
Set allowed_VOP_discard
/*Acceptable VOP discard level*/
Set Min_VOP_discard
/*The limit below which the QL starts considering increasing QL*/

```

Step 2: Check for increase opportunity:

```

if (VOP_type == IVOP)      /*increase opportunity*/
{
    if ((network not congested && PCR > RR) && (last_increase > T)
        && (long_cycle_cell_dropped=0)
        && (long_cycle_discarded_VOP < Min_VOP_discard))
    {
        increase QL
        Request Rate with double aggressive parameter
    }
}

```

Step 3: Check for a decrease opportunity:

```

if (VOP_type==PVOP || IVOP) /*decrease opportunity*/
{
    if ((network is congested || ACR=PCR)
        && ((EWM_ME > Max_ME) ||
            (short_cycle_dropped_VOP > allowed_VOP_discard))
        || (short_cycle_cells_dropped>0))
    {
        decrease QL
    }
}

```

Figure 6-14: QLSA Algorithm

6.5 Rate Request Algorithm (RRA)

The ABR service provides the capability for the application to dynamically re-negotiate the required bandwidth through the RM cells. However, in the case of real-time video applications it is not a trivial task for the source to determine the required rate. The Rate Request Algorithm (RRA), specifies how the source determines the value of the requested rate for MPEG-4 video streams.

MPEG-4 video traffic, like other MPEG video, is characterized by its burstiness and continuous fluctuations of its rate, which poses a problem in estimating the requested bandwidth. As mentioned in Chapter 3, AT&T research lab, among other research groups, has proposed mechanisms that rely on using a video traffic model for predicting future traffic. This predicted rate is then used as the requested rate and inserted into the RM-cell. This mechanism has its drawbacks with respect to finding an adequate traffic model in the case of real-time applications, that are characterized by the lack of previous knowledge of the video sequence. If the requested rate is less than the actual source rate, the video sequence will suffer from cell loss and hence quality degradation. On the other hand, if it is greater the link will be under-utilized as the ER algorithm may reserve more bandwidth than the connection needs. In other words, the RRA tries to answer the question of how to request a rate which may or may not be used by the application.

The proposed Rate Request algorithm relies on setting the requested rate based on previous bandwidth consumption. The idea is to request a rate with a certain percentage

higher than the previous source rate. The requested rate will not be further increased unless the source starts using the previous increment. Two major problems have to be solved in order to implement this technique successfully. The first is how to measure the source rate, and the second is to find a suitable function to calculate the requested rate. As it is known that the MPEG-4 traffic is bursty in nature and fluctuates greatly so we need some sort of averaging functions. However, the most recent samples should be given higher weight in the used average function. These two requirements suggested the use of an Exponential Weighted Moving Average (EWMA) function, which acts as an aging function and a lowpass filter as will be described in Section 6.7.3.

With respect to the second problem, the use of a simple first order function is suggested for ease of implementation. The options are either multiplicative or additive functions. With multiplicative functions, the difference between the requested rate and the EWMA_SR increases with the increase in EWMA_SR. However, this difference is always constant with additive functions regardless of the EWMA_SR value. Due to the requirement that the initial setting of RRA parameters be the same for all the traces and under various quality levels, and due to the fact that the bandwidth requirements for the test sequences differ greatly, it is very difficult to find a common constant K for all of them with an additive function whereas this can be achieved with a multiplicative function. Hence a multiplicative function was chosen for the RRA Figure 6-15 on page 178 illustrates the multiplicative and additive function options for the RRA.

The RRA is required to be able to operate efficiently for two scenarios. The first scenario is when we have a fixed QL and the RRA adjusts the requested rate between the ACR and the source rate to enhance the matching efficiency. The second scenario is when the RRA requests extra bandwidth in order to enable an increase in QL whenever network conditions permit. Despite the apparent simplicity of the proposed algorithm, the simulation results demonstrate its efficiency for both scenarios eliminating the motivation to design more complex algorithms. Moreover, the RRA performs well for the different test sequences using a single configuration and parameter settings. This alleviates the user from the burden of predicting the future traffic.

The RRA is controlled by the aggressiveness parameter defined as K (for the multiplicative function). This parameter is set to 0.35 which provides a minimum ME of 75%. Larger values of K provide a more conservative operation in terms of cell loss but at the expense of a smaller ME. On the other hand, smaller values of K provide higher ME at the expense of a higher risk in terms of cell loss. The application can set K according to its loss tolerance and ME requirements.

Figure 6-15: Requested Rate Function Options

Figure 6-16: Rate Request Algorithm (RRA) Interface

Figure 6-16 on page 178 shows the interface for the RRA. The RRA works in conjunction with the QLSA in two cases. First, when the QL is stable for a considerable duration of time and the network is not congested, the RRA acts more aggressively in terms of the Requested Rate (RR). This behavior gives an indirect opportunity for the QLSA to increase QL if the other conditions permit. Second, whenever the QLSA increases QL, the RRA takes an extra measure of precaution and increases its RR more aggressively. Note that the response times to increase QL and to grant the requested rate are different. Whenever the QLSA decides to increase QL it takes effect at the arrival of the next VOP. However, if the network decides to grant the RR to the source, this request is placed with the next F_RM cell and takes effect after a round-trip delay. It should be mentioned that the RRA is activated only for MPEG-4 video applications. In the case of data applications the requested rate is always set to PCR with MCR set to zero to provide video applications with a higher priority.

Concluding, the RRA features can be summarized as follows:

- The RR is based on the most recent consumption of bandwidth.
- The most recent consumption of bandwidth is measured using EWMA function.
- The RRA operates in co-ordination with the QLSA.
- The aggressiveness parameter is directly derived from the target ME.

The above four features provide a simple to implement algorithm. Yet, the simulation results demonstrate that a high utilization can be achieved without violating the required QoS. The pseudo-code for the RRA is listed in Figure 6-17 on page 180.

Algorithm RRA:

The Requested Rate Algorithm (RRA) is responsible for setting the rate that is requested by the source from the network via F-RM cells. RRA works in co-ordination with QLSA to ensure the availability of resources before increasing the quality level (QL).

Step 1: Initialization:

- Set $\text{target_ME} = 75\%$
- Set $k = 1 - (1/\text{target_ME})$
- Set $\alpha_{SR} = 0.125$
- Set $\alpha_{ME} = 0.05$
- $\text{EWMA_SR} = \text{ICR}$
- $\text{EWMA_ME} = 0.75$
- $\text{Requested_Rate} = (1+K) \times \text{ICR}$

Step 2: With the occurrence of a short cycle event

*/*Adjust source rate and matching efficiency using EWMA */*

$$\begin{aligned}\text{EWMA_SR} &= (1 - \alpha_{SR}) \times \text{EWMA_SR} + \alpha_{SR} \times \text{SR} \\ \text{EWMA_ME} &= (1 - \alpha_{ME}) \times \text{EWMA_ME} + \alpha_{ME} \times \text{ME}\end{aligned}$$

/ Request double K if QL has been recently increases */*

```
if (last_increase < time_constant)
    Requested_Rate = (1+2K) × EWMA_SR
else
    Requested_Rate = (1 + K) × EWMA_SR
if (Requested_Rate > PCR)
    Requested_Rate = PCR
```

Step 3: With the transmission of an F-RM cell

Set ER field in F-RMcells to Requested_Rate

Step 4: With Quality Level Increase

$$\text{Requested_Rate} = (1+2K) \times \text{EWMA_SR}$$

Figure 6-17: Rate Request Algorithm

6.6 Adaptive Traffic Shaping Algorithm (ATSA)

The function of traffic shaping is two-fold. First, it ensures that the source conforms to the traffic contract and hence no cells will be dropped if the traffic is policed by a down-stream policer. Second, it smoothes the traffic and reduces the burstiness of video traffic. One of the most common techniques used for traffic shaping is the leaky bucket. Regarding smoothing, traffic smoothing, it normally occurs on a GOP level to be able to smooth the fluctuations caused by the various compression techniques used in the IVOP, PVOP and BVOP.

In this model, an adaptive cell spacing shaper is proposed that adapts easily to changes in ACR according to EQ 6.1. The shaper serves the source buffer with a variable rate equal to ACR which meets the conformance requirements for the ABR service. Regarding the smoothing effect, we cannot smooth on a GOP level as this leads to a smoothing delay of 1.2 sec. which is not acceptable for real-time traffic. Instead the maximum allowable smoothing delay ($D_{\max}$) is negotiated between the application and the source. The application sets $D_{\max}$ based on end-to-end delay requirements and the other delay components as described in Chapter 3. Any cell that exceeds $D_{\max}$ in the smoothing buffer will be dropped and considered lost. In this architecture, the buffer size is dynamically set with the change of service rate according to EQ 6.2 in order to be able to guarantee a constant $D_{\max}$ with a variable service rate. The maximum buffer size requirement occur when the service rate is equal to PCR according to EQ 6.3.

$$\text{Cell_Space}(t) = 1/\text{ACR}(t) \quad (\text{EQ 6.1})$$

$$\text{Buffer_Size}(t) = D_{\max} * \text{ACR}(t) \quad (\text{EQ 6.2})$$

$$\text{Max_Buffer_Requirements} = D_{\max} * \text{PCR} \quad (\text{EQ 6.3})$$

6.7 On-line Measurements

The RRA and QLSA depend to a great extent on the on-line measurements to achieve their design objective. Video traffic in general is characterized by both short and long range dependencies. Hence, the operation of the RRA and the QLSA should depend on both short-term and long-term measurements. Three types of statistical measurements are gathered in the proposed source model and defined in the following subsections.

6.7.1 Short-Cycle Measurements

These measurements are used to provide a fast response to the events that need a fast reaction. One of the events that require a fast response is the dropping of cells from a basic layer VOP. The duration of the short cycle is programmable in our architecture and can be altered based on the required response time and the application loss tolerance. One of the most important decisions based on the short-cycle measurements is the quality level (QL) reduction. The short-cycle is chosen, in this study's model, to be the time between two consecutive I or P VOPs which is equivalent to 300 msec in the chosen encoding sequence.

The major measurements that are collected in the short-cycle are:

- Number of transmitted cells
- Number of cell opportunities
- Number of cells dropped
- Number of VOPs discarded
- Source Rate (SR)
- Matching Efficiency (ME)

6.7.2 Long-Cycle Measurements

In addition to the events that need a fast response, there are some events that require a slower reaction in order to prevent too frequent fluctuations in the decision of the QLSA and to ensure its stability. An example of these events is the case of a low-matching efficiency. One of the reactions that is taken based on the long-cycle is to increase the QL. Before reacting to such events, we need to ensure that other important performance metrics are protected for a sufficiently long period of time. Such metrics that are observed for the long-cycle events are:

- Number of cells dropped
- Number of VOPs discarded

The duration of the long-cycle is programmable and controlled by a time constant parameter (T) which is a multiple of GOP time. In our setting, the GOP time is set to 1.2 sec and T was chosen to be 3 which results in a long cycle duration of 3.6 sec.

6.7.3 Exponential Weighted Moving Average (EWMA) Measurements

Due to the nature of both the ABR service and the MPEG-4 video traffic, one can experience events to which a reaction based on short-cycle or long-cycle measurements does not yield the desired behavior. In particular, measurements that are directly related to the source rate and the matching efficiency. Reacting directly to these measurements will lead to a high rate of fluctuations. Hence, a lowpass filter that can filter any transient measurements for these two metrics is needed. Although one need to consider these metrics for a considerable duration of time, the most recent values are more important than the older ones. This means that the values of these metrics age with time. One function that can satisfy these two needs in terms of acting as a lowpass filter and an aging function is the Exponential Weighted Moving Average (EWMA). The EWMA is used to measure the following:

- EWMA Matching Efficiency (EWMA_ME)
- EWMA Source Rate (EWMA_SR)

Hence, no decision will be taken based on the short-term increase or decrease of the above two metrics that results from the bursty traffic. EWMA measures is calculated according to EQ 6.4 and EQ 6.5.

$$\text{EWMA_SR} = (1-\alpha\text{SR})\text{EWMA_SR} + \alpha\text{SR}\text{SR} \quad (\text{EQ 6.4})$$

$$\text{EWMA_ME} = (1-\alpha\text{ME})\text{EWMA_ME} + \alpha\text{ME}\text{ME} \quad (\text{EQ 6.5})$$

The weight α determines the time constant of the lowpass filter. Two different val-

ues for α have been chosen for the two calculations. The chosen values are 0.125 for α_{SR} and 0.05 for α_{ME} . This choice leads to a decay of the measured rate in about 10 sec while the matching efficiency decays in approx. 30 sec. Figure 6-18 on page 186 shows the aging weight for the above values of α . Note that the usage of EWMA has been discussed earlier in the context of the QLSA and RRA. It should be mentioned, that the usage of EWMA was originally proposed by Floyd as a lowpass filter for the calculation of the average queue length in the Random Early Detection (RED) Gateway [Floyd93].

Figure 6-18: Aging Property of EWMA

6.8 Source State Diagram

The various components of the source model are invoked reacting to the occurrence of various events during the life-time of a connection. Following are the major events:

6.8.1 Cell Transmit Opportunity (Cell Tx opp.)

This event is activated when the connection is given an opportunity to transmit one cell. When this event is activated, the source performs the required ABR source behavior in terms of transmitting a data cell or inserting an F-RM cell. In the case of the F-RM cell, the most recently calculated value for the Requested Rate is used. The statistics are updated based on the type of the performed task.

6.8.2 Backward-RM Cell Arrival (B-RM Rx)

This event is activated upon the arrival of the B-RM cell from the network. When this event is activated, the ABR source behavior module calculates the ACR according to the source behavior rules. Then it conveys this information to the adaptive shaper to adjust both the cell spacing requirements and the buffer size. The congestion indication and ACR become available for use with the other components whenever needed.

6.8.3 VOP Arrival (VOP Rx)

This event is activated upon the arrival of a VOP to the source buffer. This event occurs once every 33 ms in the case of 30 VOPs/sec video. When this event is activated the source determines the type of the VOP stream. If the VOP is from the basic layer stream, then it will be segmented into cells and enqueued. If it is from the enhanced layer

stream, then its size will be passed to IVDA to decide whether to accept or discard it. It should be noted that in this research it was assumed that one complete VOP arrives to the source every VOP_time.

6.8.4 Short-Cycle Timer (Short Cycle)

This event occurs when a VOP is received and the short-cycle timer has expired. The statistics for the short-cycle are reset and QLSA is invoked. The short cycle presents an opportunity to decrease the quality level.

6.8.5 Long-Cycle Timer (Long Cycle)

This event occurs when a VOP is received and the long-cycle timer has expired. The statistics for the long cycle are reset and QLSA is invoked. The long-cycle timer is also an opportunity to increase the quality level. Figure 6-19 on page 189 shows the source state diagram and Table 6-5 on page 188 summarizes the relations between the source components and the state diagram.

Table 6-5: Components Invoked with Various Events

Event	Functions Invoked
Initialization	Initialization
Cell Tx Opp	ABR Source Behavior, On-line Measurements, RRA
B-RM Rx	ABR Source Behavior, Adaptive Shaper, RRA
VOP Rx	IVDA, State Tracking, RRA, On-line Measurements
Short Cycle	On-line Measurements, QLSA
Long Cycle	On-line Measurements, QLSA

Figure 6-19: Source State Diagram of the source

Chapter 7

Performance Evaluation

7.1 Introduction

This chapter presents the performance evaluation and analysis of the proposed architecture and algorithms introduced in Chapter 6. Section 7.2 describes the performance metrics and the simulation model used. In Section 7.3, the simulation results and analysis are presented followed by the conclusions in Section 7.4.

7.2 Performance Metrics and Simulation Model

The most significant design objective of the proposed architectural model and algorithms is to maximize the perceptual quality of the received video at the destination while targeting a high resource utilization. Therefore, to evaluate their performance the perceived video quality was measured. A subjective measure of the perceptual quality, referred to as the Mean Opinion Score (MOS), has been defined and extensively used to design and compare video coding algorithms [Wakeman95]. However, a MOS-based feedback mechanism is not a possible option due to the unavailability of public MPEG-4 players at the time of this research. Only the encoder and decoder were publically available. Consequently, one has to rely on objective measures. One of the most commonly used objective measures is the Signal to Noise ratio (SNR). Other measures are the loss rate, which can be represented by CLR and the VOP rate (VOP_Rate) that should be also considered. In addition to the video quality the CDF for the source delay is also used as a per-

formance metrics.

In order to evaluate the performance of transporting MPEG-4 video over ATM/ABR service using the proposed architectural model, a trace-driven simulator was built to simulate the source architecture and the ATM/ABR network. Figure 7-1 on page 192 shows the simulation model, where five sources are connected to the first ER switch via a 1Mbps link, which is connected to a second ER switch via link₁, which is connected to the destination via a 4Mbps link.

The round-trip propagation delay between source and destination is set to 17ms. The maximum allowable delay in the source buffer is set to 200ms. Based on a rough knowledge of the various video sequences, the MCR is set to 384Kbps for the soccer trace where a high level of motion activity is expected, and set to 256Kbps for both the CNN and BBC trace. The CNN trace is characterized by a high level of spatial details and low motion activity. The BBC news trace is characterized by moderate motion activities and moderate spatial details. Finally, the movie trace is assigned 128Kbps MCR due to its low spatial details and motion activity. The PCR for all of the four video sessions is set to 1Mbps, which is the link speed that connects them to the first ER switch. The data source has an MCR set to zero and a PCR to 4Mbps which enables this source to grab any unused bandwidth from the video sources. With respect to the other ABR parameters, both the RIF and RDF are set to 1, Nrm is set to 32 and ICR is set to MCR. The setting of the various source algorithms parameters are as described in Chapter 6, where the same setting applies to all the video sessions.

Figure 7-1: Simulation Model

7.3 Simulation Results and Analysis

In this section, the results and the analysis of several simulation experiments are presented to evaluate the performance of the proposed architecture. All the simulation experiments are carried for the four MPEG-4 video traces. The simulation experiments study the performance under normal conditions where the link bandwidth is greater than Σ MCR and less than Σ PCR for the video sources. The behavior of the Rate Request Algorithm is studied as well. The impact of VOP discard algorithms and QLSA on the quality of the transmitted video is then evaluated. The proposed architecture is also evaluated under heavy load traffic which resulted in a congested network. The proposed source architecture and its interaction with a background traffic that can frequently grab a large percentage of the bandwidth and leaving only MCR for the video sources is also evaluated. Finally, the impact of the round trip-delay and MCR settings is mentioned.

7.3.1 Performance of the Proposed Architecture & Algorithms

In this section, the performance of the proposed architecture and algorithms under normal operating conditions is evaluated and analyzed. The performance results of this study will be used as the base reference case when evaluating the different VOP discard algorithms.

In this case, it is assumed that bandwidth of link₁ is 2Mbps, which is less than Σ PCR for the video sources (4Mbps), and larger than Σ MCR of the video sources (1Mbps). It is expected that, the video session will be offered a bandwidth between its MCR and

Transmission of Multi-layered MPEG-4 Video over ATM/ABR Networks

PCR. The performance of the proposed architecture is studied with respect to video quality, CDF of the source delay, and overall link utilization.

Figure 7-2 on page 196 shows the attained video quality for the four MPEG-4 traces, which shows a zero loss ratio for all the four traces. The average VOP rate ranges from 25.97VOPs/sec for the CNN trace to 27.48VOPs/sec for the movie trace. This accounts to an average drop of 10% from the original VOP rate which is 30 VOPs/sec including both the basic and enhanced layer VOPs. The average quality level (QL) ranges from 1.85 for the soccer trace to 2.83 for the movie trace. Regarding SNR, an almost stable SNR for the BBC, CNN, and movie traces was achieved. However, for the soccer trace, the SNR fluctuations are more apparent which is attributed to the high level of burstiness in the soccer trace requiring more frequent fluctuations in the quality level, and hence in the SNR.

Figure 7-3 on page 197 shows the Cumulative Distributive function for the source delay. The source buffer was tuned to provide a maximum delay of 200ms. However, the delay that the cells experience inside the source buffer is normally much less than the maximum delay. We can see from Figure 7-3 on page 197 (a), that in the case of the soccer trace, 92% of the cells experienced a source delay less than 50ms and 99% of the cells experienced a source delay less than 100ms. Figure 7-3 on page 197(b), Figure 7-3 on page 197(c), and Figure 7-3 on page 197(d) show a similar trend for the rest of the traces where the probabilities that the source delay be less than 50ms are 0.91, 0.80 and 0.90 and

the probabilities that the source delay be less than 100ms are 0.99, 0.95, and 0.98 for the BBC, CNN and movie traces respectively. It should be noted that the delay is normally controlled by the burst size of the traffic, and hence one may expect the soccer trace to have a delay greater than the CNN trace given that the motion activity in the CNN trace is minimal. However, the traffic of the CNN trace is still considered a bursty traffic with a large burst size caused by the different VOP encoding of IVOP, PVOP and BVOP.

From Figure 7-4 on page 198(a), which illustrates the behavior of the data source and the final utilization achieved on link₁, one can see that the data source was able to grab any bandwidth, not used by the four video sources, resulting in an average transmission rate of 0.5Mbps for data source. Figure 7-4 on page 198(b) shows that the proposed architecture has attained a utilization of over 90% of the link, for both video and data traffic, where over 70% of the traffic is video traffic which is considered a very good ratio and not easy to achieve [Luo97].

Figure 7-2: Video Quality of the Proposed Architecture

Transmission of Multi-layered MPEG-4 Video over ATM/ABR Networks

Figure 7-3: CDF of Source Delay

Figure 7-4: Data Source & Link₁ Utilization

One can conclude from Figure 7-2 on page 196, Figure 7-3 on page 197, and Figure 7-4 on page 198 that the proposed architecture and algorithms are capable of the following:

- Utilizing over 70% of the available bandwidth for the video traffic with zero CLR. This is translated into higher QL. Note, that it was shown in Chapter 5 that without any control less than 50% utilization was achieved
- Achieving a VOP rate of 27 VOPs/sec on the average. This counts for discarding 10% of the enhanced layer VOPs. The effect of such a drop is much less annoying to the viewer than the effect caused by lost cells.
- Achieving 90% overall link utilization for both video and data traffic. This high utilization is feasible due to the adaptive feature of the source model and the closed-loop feedback of ER networks.
- Supporting various MPEG-4 traces with different traffic characteristics with the same set of parameters setting.
- Achieving an average smoothing delay between 20ms and 30ms with the probability that the smoothing delay will be larger than 100ms being less than 0.05 for all the traces. This delay is achieved when setting the maximum delay to 200ms.

7.3.2 Behavior of Rate Request Algorithm

The network offers the source bandwidth based on the requested rate by the source as calculated by Rate Request Algorithm (RRA) and based on the network availability. In order to demonstrate the behavior of the RRA, the bandwidth of link₁ and link₂ was increased to 4Mbps, which is equal to the sum of PCR of the four video traces. By doing such, enough bandwidth was provided inside the network to support all the rates requested by the sources (less than or equal to PCR).

Figure 7-5 on page 202 depicts the behavior of the RRA through plotting both ACR and the source rate for the four video traces. The allowed cell rate (ACR) is calculated based on the information provided by the network and the ABR source behavior as explained in Chapter 4. Hence, the ACR becomes equal to RR with a time lag equal to the round-trip delay. Note that the source rate is the actual rate at which the source sends, which should always be less than the ACR. It is obvious, that the RRA is able to estimate the future requirements of the video sources with an acceptable level of accuracy, and is able to achieve a matching efficiency in the range of 75% between the forecasted video traffic and the real one. After analysis of the results, the need for any more complex or sophisticated forecasting mechanism has been eliminated considering the significant difference between the traffic characteristics of the four traces and the attained matching efficiency.

The results for the CNN trace, shown in Figure 7-5 on page 202 (c), show a noticeable difference between the ACR and the results for the source rate. This can be attributed to its being the least bursty trace. However, the RRA always expects some sort of future burstiness which never happens in the case of the CNN trace. Considering that the aggressiveness parameter (K) is set to the same value for all the test sequences, this result is considered acceptable. If it is definite, beforehand, that one video session will have no scene changes, as in the case of video conference sessions, then k can be set to a lower value, and consequently the difference between RR and SR will be smaller at the cost of a higher risk to violate the QoS in cases of sudden traffic burstiness caused by high motion activity.

Figure 7-5: Behavior of RRA with an Uncongested Network

7.3.3 Impact of VOP Discard Algorithms

This section presents the study of the impact of VOP discard algorithms on the quality of the transmitted video. For the purpose of this study, the simulation experiments of Section 7.3.1 are modified to run with the different VOP discard algorithms. First, the case where no discard algorithms are implemented is studied. This case is equivalent to the transmission of single layer video or multi-layer video without deploying any control on the enhanced layer. Then the case where M-AEVD is implemented instead of IVDA is studied. The results of both cases are then compared to the IVDA results of Section 7.3.1.

Figure 7-6 on page 205 illustrates the video quality when disabling IVDA. As expected, the main benefit is sensed in the VOP rate which is stable at 30 VOPs/sec as there is no VOP discard. But this gain cannot justify the degradation in the performance of all the other quality parameters. In terms of dropped cells, all four traces experienced cells being dropped ranging from 63 cells for the BBC trace to 494 cells for the CNN trace. The number of dropped cells can have a dramatic impact on the quality of the transmitted video as was earlier described. The quality level (QL) and SNR, suffer more fluctuations than in the case of IVDA. The main reasons for these fluctuations can be attributed to the fact that in the absence of a discard mechanism, the QLSA is used as both a congestion recovery and a congestion avoidance mechanism resulting in more frequent fluctuations in the quality level, as shown in Figure 7-5 (a) through Figure 7-5 (d). However, when the IVDA is activated, the discarding of the enhanced layer VOPs acts as congestion recovery mechanism and the QLSA is used to control the quality level to avoid future congestion.

This architecture results in a fairly stable quality level. These fluctuations, when deactivating the IVDA, become more severe when the network starts to become congested under heavy load traffic. Due to the obvious benefits of the transmission of multi-layer video and the control of the enhanced layer, no further studies are carried out for this specific case.

The second study evaluates the performance of M-AEVD as the deployed VOP discard algorithm. Figure 7-7 on page 206 shows that the performance of M-AEVD is very similar to that of IVDA and even better than that of the IVDA when considering the VOP discarding, which on the average is 6% compared to 10% in the case of IVDA. However, we should realize that IVDA is mainly designed with the mandate to provide the best protection for CLR, equivalent to the case when only the basic layer VOPs are transmitted. This is achieved at the cost of more VOP discards. The advantage of such is clearer in the case of a congested network as will be demonstrated in Section 7.3.5. However, in the case of an uncongested network, as for these experiments, we cannot realize the benefits of IVDA over M-AEVD as both of them achieve a zero cell loss. In summary, it can be concluded that if the network is uncongested, a simple threshold-based discard algorithm such as M-AEVD can do the required functionality. However, under heavy load traffic, IVDA is expected to significantly outperform M-AEVD as will be demonstrated in Section 7.3.5.

Figure 7-6: Video Quality of Proposed Architecture without IVDA

Figure 7-7: Video Quality of Proposed Architecture with M-AEVD

7.3.4 Impact of QLSA

This section evaluates the performance of the proposed QLSA versus the most commonly used buffer-based control mechanism. In a buffer-based control mechanism, the quality level (QL) is adjusted based on the buffer occupancy to prevent overflow or underflow of the buffer. As shown in Figure 7-8 on page 209, a feedback from the source buffer adjusts the MPEG-4 encoder quality level, where QL is increased if the buffer occupancy is less than threshold_1 and decreased if the buffer occupancy is larger than threshold_2 .

Figure 7-9 on page 210 illustrates the video quality results for the buffer-based quality control mechanism. It should be noted that all of the other components of the proposed architecture are not modified in order to be able to study the performance of QLSA in isolation. The threshold values chosen for the simulation experiments are the equivalent of two VOP_time for threshold_1 and three VOP_time for threshold_2 . Comparing Figure 7-9 on page 210(a), Figure 7-9 on page 210(b), and Figure 7-9 on page 210(d) to the respective graphs in Figure 7-6 on page 205, it becomes obvious that QLSA provides a more stable operation in terms of QL. This behavior can be expected as QLSA depends on several parameters and on on-line measurements to decide about the QL whereas the buffer-based mechanism depends only on a single metric; namely the buffer occupancy. Another major drawback of the buffer-based control is the need to set the threshold and the sensitivity of the performance to this parameter. This problem is clearly demonstrated in Figure 7-9 on page 210(d), where the threshold setting of the other video traces was not suitable for the

CNN trace. In the case of the CNN trace, there is no motion activity and hence the traffic is less bursty if smoothed over several VOPs resulting in a buffer occupancy that is always greater than threshold_1 . Hence, QL is never increased even though the network is able to provide the CNN trace with more bandwidth. This problem would have been resolved if different threshold values had been chosen. It is worth noting, that although QLSA is more complicated than the simple buffer-based control, however from the user's perspective its operation is easier as QLSA does not rely on complicated parameters that are sensitive to the video trace type as is the case with the threshold value of the buffer-based control mechanism.

For a more comprehensive study, the behavior of the buffer-based control mechanism is also studied under heavy load conditions in Section 7.3.5.

Figure 7-8: Buffer-Based Quality Level Control Mechanism

Figure 7-9: Video Quality of Proposed Architecture with Buffer-Based QL Control
Transmission of Multi-layered MPEG-4 Video over ATM/ABR Networks

7.3.5 Impact of Traffic Load

This section presents a study of the impact of heavy load traffic on the proposed architecture for several configurations. The proposed architecture is first configured with the QLSA and IVDA. Other studied configurations replace IVDA with M-AEVD and QLSA with a buffer-based control mechanism. These sets of experiments are very important as they simulate the congestion condition where the network offers only MCR most of the time. In order to simulate this condition, the link_1 capacity is reduced to 1Mbps which is equal to ΣMCR for the four video traces.

We start by analyzing the performance of the proposed architecture with QLSA and IVDA by observing the different metrics for video quality. The results shown in Figure 7-10 on page 213 demonstrate that the IVDA behaves as desired with respect to the VOP rate. By discarding a VOP IVDA provides an initial reaction with the ultimate goal to stabilize at the adequate quality level resulting in a discarded VOP rate in the range of 10%. This objective was achieved for the BBC, CNN, and movie traces where the VOP discard rate was in the required range. However, for the soccer trace approximately 20% of the VOPs were discarded. This is attributed to the bursty nature of the soccer trace, where the IVDA needs to protect the video quality from any cell loss at the cost of extra VOPs being dropped. Nonetheless, this ratio can be improved by increasing MCR. Regarding the cell loss ratio, both the soccer and the BBC traces achieve zero cell loss in spite of their bursty nature. Yet, the CNN trace experienced a drop of 21 cells belonging to

one VOP. Similarly, the movie trace experienced a drop of 10 cells belonging also to a single VOP. It should be noted, that although IVDA tries to protect the basic layer cells from being dropped, this can still occur when the network is congested, especially when MCR is chosen in a very rough and coarse manner as in our case due to the lack of knowledge of the exact traffic characteristics.

**Figure 7-10: Video Quality of Proposed Architecture with IVDA
(Congested Network)**

Figure 7-11 on page 215 illustrates the results pertaining to the video quality of the proposed architecture when studied with M-AEVD. As expected, under congested conditions, M-AEVD is not capable of providing the same guarantee in terms of CLR as IVDA. With M-AEVD, the soccer trace suffers a drop of 42 cells and the BBC trace suffers a drop of 11 cells, as opposed to zero cell loss with IVDA for both traces,. The CNN traces incurs a drop of 152 cells spanning multiple VOPs compared to 21 in the IVDA case and finally the movie trace incurs a drop of 19 cells compared to 10 cells again with the IVDA. A minor enhancement in the VOP rate and QL of around 1% is noticeable, nevertheless it cannot justify the large number of dropped cells and the effects of such on the overall video quality.

Figure 7-12 on page 216 evaluates the video quality of the proposed architecture with a buffer-based quality control mechanism. It is obvious from Figure 7-12 on page 216(a) and Figure 7-12 on page 216(b) that the buffer-base control mechanism provides frequent fluctuations in the quality level. These fluctuations are obvious for video traces with high level of motion activity such as the soccer and BBC traces. However, for traces with low level of motion activity such as the CNN and the movie trace, the buffer-based control mechanism is adequate.

Figure 7-11: Video Quality of Proposed Architecture with M-AEVD (Congested Network)

Figure 7-12: Video Quality of Proposed Architecture with Buffer-Based Control Mechanism (Congested Network)

7.3.6 Impact of Background Traffic

In order to simulate a realistic network environment, we need to study the effect of background traffic on the proposed architecture. A square wave background traffic is applied to switch1, which periodically grabs 1Mbps of the bandwidth. The period was chosen to be every 10 seconds. As a result of the background traffic, the bandwidth available for the ABR sources (both video and data) fluctuates between 2Mbps to 1Mbps.

The objective of this study is to investigate how the video quality is affected by the background traffic. The introduction of background traffic leads to the reduction of the bandwidth available to the ABR sources and hence reducing their ACR. Figure 7-13 on page 220(a), Figure 7-13 on page 220(b), Figure 7-14 on page 221(a), and Figure 7-14 on page 221(b) illustrate the Allowed Cell Rate (ACR) for the four MPEG-4 sources. Due to the RDF and RIF setting, the ACR is dramatically reduced when the bandwidth is grabbed by VBR traffic. This setting allows for the examination of the proposed architecture under severe unstable conditions where ACR experiences a reduction from near PCR to near MCR almost instantly. The proposed architecture should be capable of adjusting the QL and the VOP rate such that it matches the available bandwidth while ensuring that restricting the source delay to $D_{\max}$ does not result in any cell loss.

Figure 7-13 on page 220(c), Figure 7-13 on page 220(d), Figure 7-14 on page 221(c) and Figure 7-14 on page 221(d) depict the resulting video quality for the four traces. It can be seen that the result of four traces seems to bare a common behavior, which can be obvious when the ACR is dropped by the first reaction of the source being to dis-

card enhanced layer VOPs, and then adjusting the QL. The results for the observed video quality metrics are reasonable with two exceptions. The first exception is that the discarded VOP rate of the soccer trace is around 20%, which is much higher than the other traces which are around 10%. However, due to the high level of motion activity in the soccer trace any other mechanism that reduces the discarded VOP rate could result into higher CLR, which has an effect worse than the effect of the discarded VOPs. The second exception is that the CNN trace is not capable of achieving zero cell loss and having four cells dropped at the source. These results could be considered as unexpected for the CNN trace, which has the least motion activity among all the other traces. Nevertheless, if we review the CNN traffic characteristics we notice two things. First the CNN has no scene changes and the second is that spatial detail are high. This combination leads to the following: large IVOP (due to the spatial details) and small PVOP and BVOP (due to having only one scene). This results in the CNN trace, among all the traces, having the highest Peak-to-Mean Ratio (PMR). The PMR for the first QL for the CNN is 7.07, whereas for the BBC trace the PMR is 4.32, for the soccer trace 3.77, and for the movie trace 6.09. This bring up a very interesting point given the general approach of directly relating the level of burstiness to the level of motion activity. The author believes that this trend was build based on MPEG-1 and MPEG-2 characterized by a common practice of buffering a GOP. In the case of smoothing the traffic on a GOP level, the CNN and movie traces will be almost perfectly shaped with very little burstiness. However, as we restrict the smoothing buffer to a value much less than a GOP this results that the CNN traffic with its single

scene can be more bursty than the soccer trace with continuous change of scenes. The only difference between them is that the burstiness of the CNN is periodic and can be predicted whereas for the soccer trace the burstiness is unpredictable.

As one of the design objectives is to support various types of video traffic with a single configuration, this results in QL being compromised in some situations. On the contrary, if the algorithms are specifically tuned for each trace then a better video quality could be achieved.

Figure 7-13: Video Quality of Proposed Architecture with Background Traffic (Soccer and BBC traces)

Figure 7-14: Video Quality of Proposed Architecture with Background Traffic (CNN and movie traces)

7.3.7 Impact of Round Trip Delay

The focus of this chapter is to evaluate the performance of the source architecture, and therefore changing the round-trip delay will not affect the performance metrics being observed. However, it has a greater impact on the overall performance specifically on the delay that the cells experience inside the ER switch. Note that the effect of the round-trip delay on the switch delay has already been studied in Chapter 4.

7.3.8 Impact of MCR Setting

The setting of the value of MCR has a great impact on the video quality as MCR defines the guaranteed portion of the bandwidth. However, an exact setting of MCR requires a detailed knowledge of the traffic characteristics, which normally are not known before-hand. MCR is very critical under congestion conditions where the network only offers MCR. The relation between MCR and CLR and the average smoothing delay for a single layer MPEG-4 video can be found in Chapter 5. In addition, the relation between the setting of MCR and CLR, VOP drop, matching efficiency and the average delay for various VOP discard algorithms was studied in Chapter 6. In general, increasing MCR provides the user with more protection and guarantee in cases of congestion. However, it is expected that the pricing of the ABR service will be based on the MCR value. Two users with different MCR values can have the same bandwidth if the network is not loaded but they will be billed differently based on their chosen MCR. It should be noted that in addition to the effects of MCR on the guaranteed bandwidth, MCR can be also used as a

weight in the distribution of the elastic bandwidth resulting in more bandwidth being allocated to the sources with higher MCR. This is one of the critical factors that may differentiate ABR data sources from ABR video sources.

7.4 Conclusions

In this chapter, the performance of the proposed source architecture for the transmission of MPEG-4 video over ATM/ABR networks is evaluated. The architecture is evaluated with various VOP discard algorithms and quality level control algorithms. The performance is evaluated under different background traffic loads and different network conditions. The use of the ABR service seems promising from both the user and network's perspectives. Users are relieved from the burden of defining parameters for a traffic they are not aware of. Users are also guaranteed a minimum acceptable level of quality, which can be increased as more resources are made available by the network. Moreover, users are provided with a service that is more cost-effective relatively to other ATM services. It can be concluded that controlling both the quality level and transmission of enhanced layer VOPs provides a lot of benefits. Regarding the VOP discards and the control algorithms, we can conclude, that for videos with low levels of motion activity, a buffer-based quality control mechanism is adequate. However, QLSA performs better for video traces with high level of motion activity. Similarly, deploying IVDA is critical for congested networks. In summary, for networks, that are unlikely to be congested, a threshold-based algorithm is adequate.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

Chapter 1 presented the motivations of this research and the expected benefits of MPEG-4 video transmission over ATM/ABR networks as well as the current obstacles that prevent this approach.

From the network perspective in Chapter 4, it has been shown that if the ER switch implements certain enhanced features, if the ABR service is well engineered, and if the source follows the predefined ABR source behavior, then the ATM/ABR service can be suitable for MPEG-4 video transmission.

From the perspective of the source, Chapter 6 of this thesis proposed the source architecture and the set of algorithms that are capable of matching the source rate with the rate offered by the networks. Moreover, the proposed source architecture provides a guarantee of the maximum delay at the source. This proposed source architecture and algorithms combines Adaptive Traffic Shaping Algorithm (ATSA) with Intelligent VOP Discard Algorithm (IVDA), and a Quality Level Selection Algorithm (QLSA) in addition to Rate Request Algorithm (RRA). The proposed architecture and algorithms were evaluated in Chapter 7 for various network environments. This performance evaluation and analysis affirms the capability of the proposed source architecture to guarantee the mini-

mum required QoS, to adapt to changes in network conditions, and to efficiently utilize the available resources. This can provide a more economical way to transmit MPEG-4 video with a minimum quality level guarantee reaching a good compromise to the video quality vs. cost trade-off that has been and still is a major challenge to the large-scale deployment of multimedia video applications. The resulting enhanced performance, evaluated in Chapter 7, is not attributed to a single component, but rather to the combination and the interaction of these components that rely on on-line measurements that measure the achieved performance, compare it to the target performance and take the appropriate action accordingly. As well of significant positive performance impact is the fact that the proposed source architecture makes use of the information made available by both the MPEG-4 encoder and the network.

The source architecture can be considered as both application-aware and network-aware, whereby the source is granted access to the network condition, the MPEG-4 encoder state, and to the state of its own resources (source buffer).

8.2 Summary of Thesis Contributions

In brief, the major contributions of the research presented in this thesis lie in the design and performance evaluation of a novel source architecture, and a set of supporting algorithms which are both network- and application-aware, and which interact and interface with both the ATM/ABR networks and the MPEG-4 coder. In summary, the major contributions coupled with the results of the corresponding performance evaluations are:

- ***Contributions of the designed novel source architecture***

The novel source architecture, that is proposed and studied under a wide range of network conditions in this thesis, introduces and incorporates in its designed model several new ideas [Guirguis00a] [Guirguis00b]. The most significant ideas can be summarized as follows:

- The source relies on information from the network, the application, and its own internal components in order to perform the required rate matching between the rate offered by the network and the MPEG-4 encoder rate combining adaptive traffic shaping, quality control and congestion recovery techniques.
- Keeping track of the cyclic behavior of the VOP encoding sequence and using this information to provide protection to IVOPs.
- Maintaining on-line measurements of the required performance metrics and the capability to adapt the decision accordingly.

Another novelty of the source architecture is the integration and the interaction between its various components.

The performance evaluation and analysis, presented in Chapter 7, demonstrates that with the proposed architectural design of the source four simultaneous MPEG-4 video sessions and one data session were transmitted over a 2Mbps link, and a link utilization of over 90% was achieved. Note that over 70% of the traffic was video traffic. Between 80% to 92% of the video cells experienced a smoothing delay less than 50ms with an absolute

maximum delay of 200ms. Moreover, the source was capable of achieving a zero loss and stabilized the quality of the video session based on the available resources which resulted in an average SNR of 30db.

- ***Contributions of the proposed set of novel source algorithms***

Several algorithms were designed during this research to support the newly designed source architectural model in attaining its main objective in the efficient transmission of MPEG-4 video over ATM/ABR networks [Guirguis00a] [Guirguis00b] The performance of the algorithms was studied along with the source to confirm the attainment of the thesis objectives, and can be briefly reviewed as follows:

- A Rate Request algorithm (RRA), which depends in its design on a simple EWMA of the source rate to predict the future demand. In spite of the apparent simplicity of the RRA, the performance analysis, presented in Chapter 7, demonstrates that it is able to achieve a matching efficiency in the range of 75% between the forecasted traffic and the actual traffic.
- An Intelligent VOP dropping algorithm (IVDA), which is designed to consider the combination of network conditions, coder state, buffer occupancy, and delay constraints in its decision. The analysis results reported in Chapter 6 and Chapter 7 illustrate the superior performance of the IVDA in protecting the enhanced layer VOP compared to the other techniques, especially in the case of a congested network. Note that when the network is guaranteed to be uncongested a

simpler discard algorithm such as M-AEVD can be implemented.

- A Quality Level Selection algorithm (QLSA), which is designed to rely on on-line measurements, network conditions and coder state in its decision. The performance analysis, presented in Chapter 7, shows the superior performance of QLSA over buffer-based quality level section algorithms in terms of providing a more stable quality level and the ability to absorb any sudden glitches in network resources. In addition to the capability to gracefully degrade the quality if the congestion persists, the advantages of QLSA are even more obvious for a video session with a high level of motion activity.

- ***Contributions of MPEG-4 video traffic and transmission studies***

This research has contributed several important studies and results to the network research community. However, most significant, this research yielded the first published studies on the transmission of smoothed uncontrolled MPEG-4 video using real MPEG-4 traces at low-bit rates for the finite and infinite smoothing buffer cases [Guirguis99a]. The trade-offs between MCR, delay, and utilization for various MPEG-4 video traffic were also presented in Chapter 5. The simulation results illustrate that for uncontrolled MPEG-4 video a utilization of less than 50% is attained in order to achieve a zero loss with a delay constraint of 150 ms. On the other hand, to achieve a utilization of 75% the smoothing delay can be as high as 3.5 sec.

As pointed out in Chapter 1, the research presented in this thesis adds significant

contributions to previous research in terms of the used input traffic, the expansion of the studied set of performance metrics, and the underlying ABR network, and its compliance with the existing standards.

8.3 Suggestions for Future Research

Due to the novelty of MPEG-4, this research can be considered one of the early studies conducted in the area of the MPEG-4 video transmission. The matter of this thesis can be further researched when the MPEG-4 player becomes publically available and the full features of the MPEG-4 encoder are implemented.

Suggested research topics for future research include:

- Extending the algorithm to handle VOPs of arbitrary shapes and to control multiple VOPs with different priorities.
- The results of this research can be further evaluated by means of subjective testing in addition to the objective evaluation presented in this research.
- From a network perspective, the proposed source architectural model can be extended to handle various networks such as DiffServ and IntServ. The performance of the transmission of MPEG-4 video over these networks on the MPEG-4 quality can be studied and evaluated.

References

- Aboul-Magd94 Osama Aboul Magd, "Contribution/94-0767", *ATM Forum*, September, 1994.
- Adanez98 Xavier Garcia Adanez, "Designing New Network Adaptation and ATM Adaptation Layers for Interactive Multimedia Applications", *Ph.D. Thesis, Swiss Federal Institute of Technology*, March 1998.
- Ahn97 Byungchan Ahn, Ki-Ho Cho, Hyoieong Song and Jaehyung Park, "Design of Rate-based Congestion Control Scheme for MPEG Video Transmission in ATM Networks", *Proceedings of GLOB-COMM 97*, Vol. 3, pp. 1690-1695, November 1997.
- Alpert97 T. Alpert, V. Baroncini, D. Choi, L.Contin, R. Koenen, F. Pereira, H. Peterson, "Subjective Evaluation of MPEG-4 Video Codec Proposals: Methodological Approach and Test Procedures", *Image Communication Journal*, Vol. 9, No. 4, pp. 305-325, May 1997.
- ATMF96 "Traffic Management Specification Version 4.0", *ATM Forum*, April 1996.
- Balak97 Meera Balakrishnan, Antonio Puliafito, Kishor Trivedi and Yanniz Viniotis "Buffer Losses vs. Deadline Violations for ABR Traffic in an ATM switch: A Computational Approach", *Telecommunication Systems*, Vol. 7, No. 1-3, pp. 105-123, 1997.
- Bonomi95 Flavio Bonomi and Kerry W. Fendick, "The Rate-Based Flow Control Framework for the Available Bit Rate ATM Service", *IEEE Network*, Vol. 9, No. 2, pp. 25-39, March/April 1995.
- Casoni97 Maurizi Casoni and Jonathan S. Turner, "On the Performance of Early Packet Discard", *IEEE Journal on Selected Areas in Communications*, Vol.15, No. 5, pp. 892-902, June 1997.
- Chiariglione97 Leonardo Chiariglione, "MPEG and Multimedia Communications," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No.1, pp. 5-18, February 1997.
- Chong97 Song Chong and San-gi Li, "Probabilistic Burstiness-Curve-Based Connection Control for Real-Time Multimedia Services in ATM Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No.6, pp. 1072-1086, August 1997.

- Colonnese97a S. Colonnese, U. Mascia, G. Russo, "Core Experiment N2: Preliminary FUB Results on Combination of Automatic Segmentation Techniques" *ISO/IEC JTC1/SC29/WG11/MPEG97/2365*, July 1997.
- Conti97 M. Conti and E. Gregori, "Traffic Characterization and Bandwidth Allocation for MPEG-2 VBR Video Traffic," *IEEE GlobComm '97*, Vol. 1, pp. 443-447, 1997.
- Cruz97 Luis de la Cruz, Juan Jose Alins and Jorge Mata, "Prediction Techniques for VBR Traffic Shaping", *IEEE GlobComm '97*, Vol. 3, pp. 1434-1439, 1997.
- Dag99 Tamer Dag and Ioannis Stavrakakis, "QoS Based ABR Service under PF Based Flow Control Scheme", *Proceedings of ICC 99*, Vol. 1, pp. 124-128, June 1999.
- Dalgic97 Isamil Dalgic, Fouad A. Tobagi, "Performance Evaluation of ATM Networks Carrying Constant and Variable Bit-Rate Video Traffic", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 6, pp. 1115- 1131, August 1997.
- Duffield98 N.G. Duffield, K.K. Ramakrishnan, Amy R. Reibman, "SAVE: An Algorithm for Smoothed Adaptive Video over Explicit Rate Networks", *IEEE/ACM Transactions on Networking*, Vol. 6, No. 6, pp. 717-728, December 1998.
- Ebrahimi00 Touradj Ebrahimi and Casper Horne, "MPEG-4 Natural Video Coding- An Overview", *Image Communication Journal*, Vol. 15, No. 4-5, pp. 619-629, January 2000.
- Ebrahimi97 Touradj Ebrahimi, "MPEG-4 Video Verification Model: A video encoding/decoding algorithm based on content representation", *Image Communication Journal*, Vol. 9, No. 4, pp. 367-384, May 1997.
- Elefth95 Alexnadros Eleftheriadis, "Dynamic Rate Shaping of Compressed Digital Video", *Ph.D. Thesis, Columbia University*, 1995.

- Eun97 Doyoung Eun, Heejune Ahn, Byeong-heee Roh, and Jae-Kyoon Kim, "Effects of Long-Range Dependence of VBR Video Traffic On Queueing Performance", *IEEE GlobComm 97*, Vol. 3, pp. 1440-1444, 1997.
- Floyd93 Sally Floyd and Van Jacobson, "Random Early Detection Gateways for Congestion Avoidance", *IEEE/ACM Transactions on Networking*, Vol. 1, No. 4, pp. 397-413, August 1993.
- Fornes98 J.M. Fornes, J.A. Ternero and F.R. Rubio, "MPEG Video Aggregation over ABR-ATM", *Proceedings of the 1st IEEE International Conference on ATM*, pp. 59-65, June 1998.
- Francesch00 G. Franceschini, "The Delivery Layer in MPEG-4", *Image Communication Journal*, Vol.15, No. 4-5, pp.347-363, January 2000.
- Goyal98 Rohit Goyal, Xiangrong Cai, Raj Jain, Sonia Fahmy and Bobby Vandalore, "Per-VC Rate Allocation Techniques for ATM-ABR Virtual Source, Virtual Destination Networks", *Proceedings of IEEE Global Telecommunications Conference*, Vol. 3, pp. 1778-1785, 1998.
- Guirguis00a Ramy M. Guirguis and Samy Mahmoud, "A Rate Request Mechanism for the Transmission of Real-Time MPEG-4 Video over ATM/ABR Networks", *Proceedings of 8th International Conference on Telecommunication Systems Modeling and Analysis*, pp. 450-461, March 2000.
- Guirguis00b Ramy M, Guirguis and Samy Mahmoud, "Transmission of Real-Time Multi-Layered MPEG-4 Video over ATM/ABR Service", *Proceedings of IEEE ICC2000*, Vol. 1, pp. 259-263, June 2000.
- Guirguis98 Ramy M. Guirguis, Sameh Rabie, Osama Aboul Magd, Samy Mahmoud, "The use of Bit True Simulation in ATM Switch Design", *Proceedings of IEEE Symposium on Planning & Design of Broadband Networks*, Mont Tremblant, pp. 23-33, October 1998.

- Guirguis99a Ramy M. Guirguis and Samy Mahmoud, "Traffic Characterization and Transmission of Uncontrolled Multi-layered MPEG-4 Video over ATM/ABR Service", *Proceedings of The Third Canadian Conference on Broadband Research*, Ottawa, pp. 202-213, November 1999.
- Guirguis99b Ramy M. Guirguis, Sameh Rabie, Osama Aboul-Magd, Samy Mahmoud, "Engineering and Performance Evaluation of the ABR Service in Second Generation ATM Networks", *Proceeding of Globecom '99*, Vol. 2, pp. 1275-1282, December 1999.
- Hamidi97 Maher Hamdi, James W. Roberts and Pierre Rolin, "Rate Control for VBR Coders in Broad-band Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No.6, pp. 1040-1051, August 1997.
- Hegde99 Manju V. Hedge, Otto Schmid, H.Saidi, Paul S. Min, "Real Time Adaptive Bandwidth Allocation for ATM Switches", *Proceedings of ICC 99*, Vol. 3, pp. 1831-1835, June 1999.
- Hsu97 Chi-Yuan Hsu, Antonio Ortega, Amy R. Reibman, "Joint Selection of Source and Channel Rate for VBR Video Transmission Under ATM Policing Constraints", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 6, pp. 1016- 1028, August 1997.
- Huard98 Jean-Francois Huard, Aurel A. Lazar, Koon-Seng Lim, George S. Tselikies "Realizing the MPEG-4 Multimedia Delivery Framework", *IEEE Network*, Vol. 12, No. 6, pp. 35-45, November 1998.
- Ismail95 M. Riyaz Ismail, I.E. Lambadaris, M. Devetsikiotis and A.R. Kaye, "Modelling Prioritized MPEG Video Using TES and a Frame Spreading Strategy for Transmission in ATM Networks", *IEEE INFOCOM' 95*, Vol. 2, pp. 762-770, April 1995.
- Izquierdo98 Michael Izquierdo, "Modelling, Transmission and Multiplexing of MPEG VBR Sources Over Packet Switched Networks", *Ph.D. Thesis, North Carolina State University*, October 1998.
- Jain96 Raj Jain, Shivkumar Kalyanarman, Rohit Goyal, Sonia Fahmy, "Real Time ABR: Proposal for a New Work Item", *ATM Forum/ 96-1760*, December 1996.

- Jelenkovic97 Predrag R. Jelenkovic, Aurel A. Lazar and Nemo Semret, "The Effect of Multiple Time Scales and Subexponentiality in MPEG Video Streams on Queueing Behavior", *IEEE J. Select. Areas Commun*, Vol, 15, No.6, pp. 1052-1071, August 1997.
- Kaly00 Shivkumar Kalyanarman, Raj Jain, Sonia Fahmy, Rohit Goyal "The ERICA Switch Algorithm for ABR Traffic Management in ATM Networks", *IEEE/ACM Transaction on Networking*, Vol. 8, No. 1, pp. 87-98, February 2000.
- Kaly97b Shivkumar Kalyanaraman, "Traffic Management for the Available Bit Rate (ABR) Service in Asynchronous Transfer Mode (ATM) Networks", *Ph.D. Thesis, The Ohio State University*, 1997.
- Kaly98 Shiv Kalyanaraman, Raj jain, Jianping Jiang, Rohit Goyal, Sonia Fahmy and Pradeep Samudra,"Design Considerations for Virtual Source/Virtual Destination (VS/VD) Feature in the ABR Service of ATM Networks", *Journal of Computer Networks & ISDN Systems*, Vol. 30, No. 19, pp. 1811-1824, October 1998.
- Kanakia95 Hemant Kanakia, Partho P. Mishra, Amy R. Reibman,"An Adaptive Congestion Control Scheme for Real Time Packet Video Transport", *IEEE/ACM Transactions on Networking*, Vol. 3, No. 6, pp. 671-682, December 1995.
- Kanakia96 Hemant Kanakia, Partho P. Mishra, "Packet Video Transport in ATM Networks with single-bit feedback", *Multimedia Systems*, Vol. 4, Issue 6, pp 370-380, December 1996.
- Kesidid96 George Kesidis, "*ATM Network Performance*," Kluwer Academic Publishers, 1996.
- Kim99 Yonhwan Kim and San-qi Li, "Performance Analysis of Data Packet Discarding in ATM Networks", *IEEE/ACM Transactions on Networking*, Vol 7, No.2, pp. 216- 227, April 1999.
- Koenen97 Rob Koenen, "MPEG-4: Context and Objectives", *Image Communications Journal*, Vol. 9, No. 4, pp. 295-304, May 1997.

- Krunz99 Marwan Krunz, "Bandwidth Allocation Strategies for Transporting Variable-Bit-Rate Video Traffic", *IEEE Communications Magazine*, Vol. 37, No. 1, pp. 40-46, January 1999.
- Labrador99 Miguel A. Labrador and Sujta Banerjee, "Packet Dropping Policies for ATM and IP Networks", *IEEE Communications Surveys*, Vol. 2, No. 3, pp 2-14, Third Quarter 1999.
- Lai98a Y-C Lai and Y-D Lin, "Interpretability of EFCI and ER Switches for ABR Service in ATM Networks", *IEEE Network*, Vol. 12, No. 1, pp. 34-42, Jan./Feb. 1998.
- Lai98b Yuan-Cheng Lai and Ying Dar Lin "Performance analysis of rate-based flow control under variable number of sources", *Computer Networks and ISDN Systems*, Vol. 29, No. 17-18, pp. 2139-2152, February 1998.
- Lakshman97 T.V. Lakshman, Partho P. Mishra, K.K Ramakrishnan, "Transporting Compressed Video over ATM Networks with ABR Feedback Control", *ATM Forum Contribution 97-0652*, July 1997.
- Lakshman98 T.V. Lakshman, A. Ortega, and A. R. Reibman, "VBR Video: Trade-offs and Potentials", *Proceedings of the IEEE*, Vol. 86, No.5, pp. 952-973, May 1998.
- Luo97 Wenjun Luo and Magda El Zarki, "Quality Control for VBR Video over ATM Networks" *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 6, pp. 1029-1039, August 1997.
- Maglaris88 Basil Maglaris, "Performance Models of Statistical Multiplexing in Packet Video Communications", *IEEE Transactions on Communications*, Vol. 36, No. 7, pp. 834-843, July 1988.
- Mech97 R. Mech, M.Wollborn, "A Noise Robust Method for Segmentation of Moving Objects in Video Sequences", *Proceedings of ICASSP-97*, Vol. 4, pp. 2657-2660, 1997.
- Mehaoua99 Ahmed Mehaoua, Raouf Boutaba, Song Pu, Yasser Rasheed, Alberto Leon-Garcia, "Towards An Efficient ATM Best Effort Video Delivery Service", *Proceedings of ICC 99*, Vol. 2, pp. 1352-1356, June 1999.

- MPEG4-0999 *"MPEG-4 Testing and Evaluation Procedures Document"*, ISO/IEC JTC1/SC29/WG11/N0999, July 1995.
- MPEG4-1713 *"Frequently Asked Questions on MPEG-4 Video"*, ISO/IEC JTC1/SC29/WG11/N1713, April 1997.
- MPEG4-1936 *"MPEG Systems (1-2-4) FAQ Version 4.0"*, ISO/IEC JTC1/SC29/WG11/N1936, October 1997.
- MPEG4-2201 *"MPEG-4 Systems"*, ISO/IEC JTC1/SC29/WG11/N2201, May 1998.
- MPEG4-2201 *"Part 1: MPEG-4 Systems Committee Draft"*, ISO/IEC JTC1/SC29/WG11/N2201, May 1998.
- MPEG4-2202 *"Part 2: MPEG-4 Coding of Audio Visual Objects: Visual Committee Draft"*, ISO/IEC JTC1/SC29/WG11/N2202, May 1998.
- MPEG4-2456 *"MPEG-4 Requirements, Version 10"*, ISO/IEC JTC1/SC29/WG11/N2456, December 1998.
- MPEG4-2457 *"MPEG-4 Applications"*, ISO/IEC JTC1/SC29/WG11/N2457, October 1998.
- MPEG4-2458 *"Overview of MPEG-4 Profiles and Levels"*, ISO/IEC JTC1/SC29/WG11/N2458, October 1998.
- MPEG4-2459 *"MPEG-4 Overview"*, ISO/IEC JTC1/SC29/WG11/N2459, October 1998.
- MPEG4-2604 *"Report of the Formal Verification Tests on MPEG-4 Video Error Resilience"*, ISO/IEC JTC1/SC29/WG11/N2604, December 1998.
- MPEG4-2605 *"Report of the Formal Verification Tests on MPEG-4 Temporal Scalability in Simple scalable Profile"*, ISO/IEC JTC1/SC29/WG11/N2605, December 1998.
- MPEG4-2606 *"DMIF ISO/IEC 1449-6 Extensions Amd1 for Version 2 WD 5.0"* ISO/IEC JTC1/SC29/WG11/N2606, December 1998.

- MPEG4-2720 *"Very Low Bit Rate Audio Visual Coding - Part 6: Delivery Multi-media Integration Framework (DMIF)"*, ISO/IEC JTC1/SC29/WG11/N2720, April 1999.
- MPEG4-2773 *"MPEG-4 Industry Forum Announcement"*, ISO/IEC JTC/SC29/WG11 N2773, July 1999.
- MPEG4-2825 *"Report of Formal Verification Test on MPEG-4 Advanced Real Time Simple Profile (Error Robustness, Temporal Resolution Stability)"*, ISO/IEC JTC1/SC29/WG11/N2825, July 1999.
- MPEG4-3202 *"DMIF Roadmap"*, ISO/IEC JTC1/SC29/WG11/N3202, December 1999.
- Ni96 J. Ni, T. Yang, D.H.K. Tsang, "Source Modeling, Queueing Analysis, and Bandwidth Allocation for VBR MPEG-2 Video Traffic in ATM Networks," *IEE Proceedings Communications*, Vol. 143, No. 4, pp.197-205, August 1996.
- Ohsaki98 Hiroyuki Ohsaki, Masayuki Murata and Hideo Miyahara, "Parameter Tuning of Rate-Based Congestion Control Algorithm For ABR Service Class in ATM Networks", *International Journal of Communication Systems*, Vol. 11, No. 2, pp. 103-128, 1998.
- Ohta94 Naohisa Ohta, *"Packet video Modeling and Signal Processing"*, Artech House, 1994.
- Ohta99 Chikara Ohta and Masakatu Morii, "Feedback Picture Sequence Control for Real-Time MPEG Video Transfer through ATM Networks", *Proceedings of ICC 99*, Vol. 3, pp. 1819-1825, June 1999.
- Onvural94 Raif O. Onvural, *"Asynchronous Transfer Mode Networks Performance Issues"*, ARTECH HOUSE, 1994.
- Ostermann97 Jorn Ostermann, "Methodologies Used for Evaluation of Video Tools and Algorithms in MPEG-4", *Image Communication Journal*, Vol. 9, No. 4, pp. 343-365, May 1997.
- Pancha94 Pramod Pancha and Magda El Zarki, "MPEG Coding for Variable Bit Rate Video Transmission" *IEEE Communications Magazine*, Vol. 32, No. 5, pp. 54-66, May 1994.

- Pereira99 Fernando Pereira, "MPEG-4: Why, What, How and When?", *Image Communication Journal*. Vol. 15, No. 4-5, pp. 271-279, January 2000.
- Perrira97 Fernando Perrira and Thierry Alpert, "MPEG-4 Video Subjective Test Procedures and Results," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No.1, pp. 32-51, February 1997.
- Romanow95 Allyn Romanow and Sally Flyod, "Dynamics of TCP Traffic over ATM Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 4 pp 633-641, May 1995.
- Rose97 O. Rose, "Simple and Efficient Models for Variable Bit Rate MPEG Video Traffic," *Performance Evaluation*, Vol. 30, No.1-2, pp.69-85, July 1997.
- Rozic97 Nikola Rozic, "Source Modeling of MPEG Video" *Proceedings of IEEE GlobComm '97*, Vol. 3, pp. 1429-1433, 1997.
- Saha97 Debanjan Saha, Sarit Mukherjee, and Stish K. Tripathi, "Multirate Scheduling of VBR Video Traffic in ATM Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 6, No. 4, pp. 397-410, August 1997.
- Salehi98 James D. Salehi, Zhi-LiZhang, Jim Kurose, Don Towsley, "Supporting Stored Video: Reducing Rate Variability and End-to-End Resource Requirements Through Optimal Smoothing", *IEEE/ACM Transaction on Networking*, Vol. 6, No. 4, pp.397-410, August 1998
- Sikora97 Thomas Sikora, "MPEG-4 Video Standard Verification Model", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 7, No.1, pp. 19-31, February 1997.
- Suh98 Doug Young Suh, Seung Gi Chang, Seung Gyun.Han, "H.263v2 Video Transportation over ATM ABR Channel", *ATM Forum Contribution 98-0677*, October 1998.

- Vandalore00 Bobby Vandalore, "Traffic Managment to Enhance Quality of Service (QoS) of Multimedia over Avaialbe Bit Rate (ABR) Service in Asynchronous Transfer Mode (ATM) Networks", *Ph.D. Thesis, The Ohio State University*, 2000.
- Vandalore98a Bobby Vandalore, Raj Jain, Rohit Goyal, Sonia Fahmy, "Design and Analysis of Queue Control Functions for Explicit Rate Switch Schemes", *Proceedings of IC3N'98*, pp. 780-786, October 1998.
- Vandalore99a Bobby Vandalore, Sonia Fahmy, Raj Jain, Rohit Goyal, Mukul Goyal, "QoS and Multipoint Support for Multimedia Applications over the ATM ABR Service", *IEEE Communications Magazine*, pp.53-57, January 1999.
- Vandalore99b Bobby Vandalore, Raj Jain, Rohit Goyal, Sonia Fahmy, "Dynamic Queue Control Functions for ATM ABR Switch Schemes: Design and Analysis", *International Journal of Computer and Telecommunication Networking*, Vol. 31, No. 18, pp. 1935-1941, August 1999.
- Vickers97a Brett J. Vickers, Meejeong Lee and Tatsuya Suda, "Feedback Control Mechanisms for Real-Time Multipoint video services", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 3, pp. 512-530, April 1997.
- Vickers97b Brett J. Vickers, Tatsuya Suda, "An ATM service architecture for the transport of adoptively encoded live video", *Computer Communications* Vol. 20, No. 12, pp. 1078-1088, December 1997.
- Vickers98a Brett J. Vickers, Celio Aluquerque and Tatsuya Suda, "Adaptive Multicast of Multi-Layered Video: Rate-Based and Credit-Based Approaches", *Proceedings of IEEE INFOCOM'98*, Vol. 3, pp. 1073-1083, 1998.
- Vickers98b Brett J. Vickers, "Feedback Control Mechanisms for the Transport of Adaptively Encoded Real-Time Video", *Ph.D. Thesis, University of California, Irvine*, 1998.
- Wakeman95 Ian John Wakeman, "Congestion Control for Packetised Video in the Internet", *Ph.D. Thesis, University of London*, 1995.

- Whyte97 Bill Whyte, "Multimedia Telecommunications", Chapman & Hall, 1997.
- Woo98 Tai-Kuo Woo "A predictive approach for ABR traffic management in ATM Networks" *Computer Communications*, Vol. 21, No. 1, pp. 81-87, 1998.
- Younes98 Khaled Ben Younes and Khalid Begain, "Scalable Video on Demand on ABR in ATM Networks", *Proceedings of the 1st IEEE International Conference on ATM*, pp. 51-58, June 1998.
- Zheng99 Bing Zheng and Mohammed Atiquzzaman, "Traffic Management of Multimedia over ATM Networks" *IEEE Communication Magazine*, Vol. 37, No. 1, pp33-38, January 1999.

Appendix A : MPEG-4 Encoder

A.1. Software Implementation of the MPEG-4 standard

The MPEG-1 and MPEG-2 standards have been subject of extensive studies, where simulation programs written in C language were used to implement the simulation models of MPEG-1 and the test models of MPEG-2. Part 5 of both standards gives the software implementation of both the encoder and the decoder.

However, with MPEG-4, substantial innovation has been made with the definition of *Reference Implementation of MPEG-4 Audio, Video and System Verification Model (VM)*, written in C and C++ [Chiariglione97].

The primary intent of the *Video Verification Model (VM)* [Ebrahimi97] [Sikora97], is to provide a fully defined core video coding algorithm platform for the development standard. It evolves over time by means of core experiments. New algorithms and tools are added to the VM and the older ones are replaced in the VM by successful core experiments. Until October 1996, the MPEG-4 video group had focused its efforts on a single VM [Sikora97], which has gradually evolved from version 1.0 to version 4.0 [Ebrahimi97].

There are currently two official implementations of the MPEG-4 Video Codec. The C implementation is provided by the European projects ACTS-MONUSYS whereas the C++ by Microsoft Corp. Both implementations abide by the VM description documents. This research will use the Microsoft MPEG-4 Codec version 0.0 released in May 1998 to generate the required MPEG-4 traces.

To encode a sequence, a parameter file is needed in order to control the encoder operation.

The source file must be in 4:2:0 raw-YUV format. For object-based coding, a segmentation mask file should be prepared. This file should contain a binary or an alpha-plane data-mask, where a transparent pixel has a value of 0 and an opaque pixel is represented by 255.

Figure A.1 shows the required input and the expected output of the coder. All the coder parameters, that must be specified in the parameter file, are listed in Table A.1. The encoder will accordingly generate a bit-stream file, a reconstructed output file and an optional trace file. The decoder, shown in Figure A.2, receives a bit-stream file as input and decodes it, generating a file with a “.yuv” postfix. If the object-based decoding is enabled, the decoder will also generate a segment file. The decoder is capable of decoding of both base and enhancement layers.

Figure A.1: Microsoft MPEG-4 Coder

Figure A.2: Microsoft MPEG-4 Decoder

A.2. MPEG-4 Controlling Parameters

The MPEG-4 Codec offers a wide range of parameters that can control the encoding process and the resulting compressed output.

Table A-1: MPEG-4 Codec Parameters

Parameter Name	Default Value	Remarks
Frame Width, Height	320, 240	
First and Last Frame	varies	
Not 8-bit flag	0	if 1 source must have 2 bytes/pixel
First, Last Object Index	1,1	
Quantization Precision	5	ignored if "Not 8-bit flag"=0
Number of Bits per Pixel	8	ignored if "Not 8-bit flag"=0
VO Scalability	0	0= no, 1= temporal, 2= spatial scalability (SS)
SS: Spatial option	N.A.	enhancement layer prediction type: Type 0: PBBBBB.... Type 1: PPPPPP
SS: Frame size of Enhancement Layer	N.A.	used only when spatial scalability is activated
SS: Horizontal Up-Sampling Factor n/m	N.A.	Used only when spatial scalability is activated
SS: Vertical Up-Sampling Factor n/m	N.A.	
Temporal Scalability Type	0	
Temporal Scalability Enhancement type	0	must be 0 when encoding rectangle object
Rate Control for each VOP	0	0:No, 1:Yes (on only when no B-VOP)

Table A-1: MPEG-4 Codec Parameters

Parameter Name	Default Value	Remarks
Total of Bits for each VOP	500,000	
Alpha Usage for each object	0	0: rectangle (original frame), 1: binary, 2: 8-bit, 3:shape only binary
Binary Alpha Rounding	0	
<NSS>mb level Binary Alpha Size Conversion Disable	1	
Binary Alpha Refresh Rate	1	≥ 0
Error Resilient Disable Flags	1	1st bit (RVLC): 0=no, 1=yes 2nd bit (Data Partition): 0:no, 1=yes 3rd bit (Video Packet): 0=no, 1=yes Example: 1= no RVLC, No data partition, no video packet 0= no RVLC, No data partition, Yes video packet 4= Yes RVLS, No data partition, Yes video packet
Video Packet Length	0	for error resilient coding
Interlaced Coding Flag	0	0= progressive, 1= interlace
Quantizer	0	0 = H.263, 1 = MPEG
Load Non-Default Intra Q-Matrix	0	0 = False, 1 = True
Load Non-Default Inter Q-Matrix	0	0 = False, 1 = True

Table A-1: MPEG-4 Codec Parameters

Parameter Name	Default Value	Remarks
Threshold to code Intra DC as AC coefficients	0	0=never do that 7= always do that 1= do when qp>=13 2= do when qp>=15 3= do when qp>=17 4= do when qp>=19 5= do when qp>=21 6=do when qp>=23
Stepsize for IVOP DCT	10	1 - 31
Stepsize for PVOP DCT	10	1 - 31
Stepsize for BVOP DCT	10	1 - 31
Load non default gray alpha intra Q-Matrix	0	0 = False, 1 = True and matrix follows
Load non default gray inter Q-Matrix	0	0 = False, 1 = True and matrix follows
Stepsize for IVOP gray alpha DCT	16	1 - 16
Stepsize for PVOP gray alpha DCT	16	1 - 16
Stepsize for BVOP gray alpha DCT	16	1 - 63
Decouple Gray Quantization Update from Texture Dquant (remains constant within VOP)	0	
<NSS> Number of PVOP's between 2 IVOP's	-1	negative values means only 1 I at the start
<NSS>Number of BVOP's between 2 PVOP's	2	

Table A-1: MPEG-4 Codec Parameters

Parameter Name	Default Value	Remarks
GOV period	0	0=never
Flag for loop filter disable	1	
temporal sampling rate	3	≥ 1
maximum search range	16	≥ 1
Use the original reference VOP for ME	1	0 = False, 1 = True
Advanced Prediction Disable	0	
Input Source Frame Rate	30	
Top Field First Flag	0	relevant to interlaced source only
Alternative Scan	0	relevant to interlaced source only
Direct Mode Search Radius	2	relevant to B-VOP only
Motion Vector File Usage	0	0=not used, 1=read motion vectors from file, 2=write motion vectors to file
Motion Vector File Name		
Source File Name		
Chrom Type	2	0 = 4:4:4, 1 = 4:2:2, 2 = 4:2:0
Microblock Statistics Dump Option	0	0 = no, 1 = yes
Generate Trace File	0	0 = no, 1 = Yes
Sprite Usage	0	0 = no sprite, 1 = static sprite (current support mode) 2 = on-line Sprite 3 = GMC

Table A-1: MPEG-4 Codec Parameters

Parameter Name	Default Value	Remarks
Warping Accuracy	0	0 = 1/2 1 = 1/4 2 = 1/8 3 = 1/16 No harm in case of no sprite
No Points for Sprite	-1	-1 if no sprite, no harm in case of no sprite
Sprite Point File		
Sprite Mode	0 0	0 = no sprite 1 = low latency

Appendix B : ABR Service Parameters

B.1. RM Cell Structure

RM Cell Format

Message Type Field (MT- octet 7)

DIR = Indicates which direction of data flow is associated with RM cell.

0 for forward RM cells

1 for backward RM cells

BN = Indicated whether RM-cell is a Backward Explicit Congestion Notification (BRCN) cell or not

0 for source generated RM cells

1 for non source generated RM cells

CI = Congestion Indication allows a network element to indicate that there is congestion in the network. When a source receives a backward RM-cell with CI=1 it decrease its ACR. When turning around a forward RM-cell, a destination will set CI=1 to indicate that the previous received data cell had the EFCI state set

1 indicates congestion
0 otherwise

NI = No Increase, it is used to prevent a source from increasing its ACR.
1 to indicate no additive increase is allowed
0 otherwise

RA = not used for ABR

Table B-1: Fields and their Position in the RM-cell

Field	Description	Initial Value if source-generated	Initial value if switch or destination generated
Header	ATM Header		
ID	Protocol Identifier	1 for ABR service	1
MT	Message type field		
ER	Explicit Cell Rate	a rate not greater than PCR parameter	any rate
CCR	Current Cell Rate	ACR parameter	0
MCR	Minimum Cell Rate	MCR parameter	0
QL	Queue length	0 or set in accordance with I.371 draft	same as source
SN	Sequence Number	0 or set in accordance with I.371 draft	same as source
Reserved			
CRC-10			

B.2. ABR Service Parameters

Table B-2: ABR Service Parameters

Parameter		Description	Units & Range	Type ^a	Negotiation	Default value
PCR	Peak Cell Rate	The source cannot exceed this rate	cells/sec	S	down	NA
MCR	Minimum Cell Rate	The rate at which the source is always allowed to send	cells/sec	S	down to MCR min if signaled	0
ICR	Initial Cell Rate	The rate at which a source should send initially and after an idle period	cells/sec	S&C	down	PCR $ICR = \min \left(ICR, \frac{TBE}{FRRT} \right)$
RIF	Rate Increase Factor	Controls the amount by which the cell transmission rate may increase upon receipt of an RM cell	1/32768 -1	S	down	1
Nrm		Maximum number of cells a source may send for each forward RM cell	2-256	O	no	32
Mrm		Controls allocation of bandwidth between forward RM cells, backward RM cells and data cells	Constant fixed at 2	----	---	----
RDF	Rate Decrease Factor	Controls the decrease in the cell transmission rate	1/32768 -1	S	down	1/32,768

Table B-2: ABR Service Parameters

Parameter		Description	Units & Range	Type ^a	Negotiation	Default value
ACR	Allowed Cell Rate	Current cell rate at which the cell is allowed to send	cells/sec	----	----	----
CRM	Missing RM cell count	CRM limits the number of forward RM cells which may be sent in the absence of received backward RM-cells	implementation specific	C		$\left\lceil \frac{TBE}{N_{rm}} \right\rceil$
ADTF	ACR Decrease Time Factor	the time permitted between sending RM cells before the rate is decreased		O	down	0.5
Trm		provides an upper bound on the time between forward RM-cells for an active source	ms	O	no	100
FRTT	Fixed Round-Trip Time	the sum of fixed and propagation delays from the source to a destination and back	us 0-16.7 sec.	S	accumulated	----
TBE	Transient Buffer	negotiated number of cells that the network would like to limit the source sending during startup period before the first RM cell returns	cells 0 - 16777215	S	down	16,777,215

Table B-2: ABR Service Parameters

Parameter		Description	Units & Range	Type ^a	Negotiation	Default value
CDF	Cutoff Decrease Factor	controls the decrease in ACR associated with CRM	0, 1/64-1	O	up	1/16
TCR	Tagged Cell Rate	limits the rate at which a source may send out-of-rate forward RM cell	constant 10 cells/sec	F		10

a. S: Signaled parameters, O: Optionally Signaled parameters, C: Computer after call setup, F: Fixed