1

128

OF/DE

1

PM-1 3 1/2" x 4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

PRECISIONSM RESOLUTION TARGETS

Dynamic Rate Control for MPEG Compressed Video in ATM Networks

by

S. Karademir. BSc.
Istanbul Technical University

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

Master of Engineering

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

Carleton University
Ottawa, Ontario
September 9, 1996
© copyright
1996. S. Karademir

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395 rue Wellington
Ottawa (Ontario)
K1A 0N4

pour le : votre référence

pour le : votre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

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

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

ISBN 0-612-19495-7

Canada

Name Sibel KARADENIR

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided

Electronics and Electrical

SUBJECT TERM

0544

U-M-I

SUBJECT CODE

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0113
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523

Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language
 General 0679
 Ancient 0289
 Linguistics 0290
 Modern 0291
Literature
 General 0401
 Classical 0294
 Comparative 0295
 Medieval 0297
 Modern 0298
 African 0316
 American 0591
 Asian 0305
 Canadian (English) 0352
 Canadian (French) 0355
 English 0593
 Germanic 0311
 Latin American 0312
 Middle Eastern 0315
 Romance 0313
 Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion
 General 0318
 Biblical Studies 0321
 Clergy 0319
 History of 0320
 Philosophy of 0322
Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
 Archaeology 0324
 Cultural 0326
 Physical 0327
Business Administration
 General 0310
 Accounting 0272
 Banking 0770
 Management 0454
 Marketing 0338
Canadian Studies 0385
Economics
 General 0501
 Agricultural 0503
 Commerce-Business 0505
 Finance 0508
 History 0509
 Labor 0510
 Theory 0511
Folklore 0358
Geography 0366
Gerontology 0351
History
 General 0578

Ancient 0579
Medieval 0581
Modern 0582
Black 0328
African 0331
Asia, Australia and Oceania 0332
Canadian 0334
European 0335
Latin American 0336
Middle Eastern 0333
United States 0337
History of Science 0585
Law 0398
Political Science
 General 0615
 International Law and Relations 0616
 Public Administration 0617
Recreation 0814
Social Work 0452
Sociology
 General 0626
 Criminology and Penology 0627
 Demography 0938
 Ethnic and Racial Studies 0631
 Individual and Family Studies 0628
 Industrial and Labor Relations 0629
 Public and Social Welfare 0630
 Social Structure and Development 0700
 Theory and Methods 0344
Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
 General 0473
 Agronomy 0285
 Animal Culture and Nutrition 0475
 Animal Pathology 0476
 Food Science and Technology 0359
 Forestry and Wildlife 0478
 Plant Culture 0479
 Plant Pathology 0480
 Plant Physiology 0817
 Range Management 0777
 Wood Technology 0746
Biology
 General 0306
 Anatomy 0287
 Botanics 0308
 Botany 0309
 Cell 0379
 Ecology 0329
 Entomology 0353
 Genetics 0369
 Limnology 0793
 Microbiology 0410
 Molecular 0307
 Neuroscience 0317
 Oceanography 0416
 Physiology 0433
 Radiation 0821
 Veterinary Science 0778
 Zoology 0472
Biophysics
 General 0786
 Medical 0760
Earth Sciences
 Biogeochemistry 0425
 Geochemistry 0996

Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleocology 0426
Paleontology 0418
Paleozoology 0985
Polynology 0427
Physical Geography 0348
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
 General 0566
 Audiology 0300
 Chemotherapy 0992
 Dentistry 0567
 Education 0350
 Hospital Management 0769
 Human Development 0758
 Immunology 0982
 Medicine and Surgery 0564
 Mental Health 0347
 Nursing 0569
 Nutrition 0570
 Obstetrics and Gynecology 0380
 Occupational Health and Therapy 0354
 Ophthalmology 0381
 Pathology 0571
 Pharmacology 0419
 Pharmacy 0572
 Physical Therapy 0382
 Public Health 0573
 Radiology 0574
 Recreation 0575

Speech Pathology 0460
Toxicology 0383
Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences
 Chemistry
 General 0485
 Agricultural 0749
 Analytical 0486
 Biochemistry 0487
 Inorganic 0488
 Nuclear 0738
 Organic 0490
 Pharmaceutical 0491
 Physical 0494
 Polymer 0495
 Radiation 0754
 Mathematics 0405
 Physics
 General 0605
 Acoustics 0986
 Astronomy and Astrophysics 0606
 Atmospheric Science 0608
 Atomic 0748
 Electronics and Electricity 0607
 Elementary Particles and High Energy 0798
 Fluid and Plasma 0759
 Molecular 0409
 Nuclear 0610
 Optics 0752
 Radiation 0756
 Solid State 0611
 Statistics 0463
Applied Sciences
 Applied Mechanics 0346
 Computer Science 0984

Engineering
 General 0537
 Aerospace 0538
 Agricultural 0539
 Automotive 0540
 Biomedical 0541
 Chemical 0542
 Civil 0543
 Electronics and Electrical 0544
 Heat and Thermodynamics 0348
 Hydraulic 0545
 Industrial 0546
 Marine 0547
 Materials Science 0794
 Mechanical 0548
 Metallurgy 0743
 Mining 0551
 Nuclear 0552
 Packaging 0549
 Petroleum 0765
 Sanitary and Municipal 0554
 System Science 0790
 Geotechnology 0428
 Operations Research 0796
 Plastics Technology 0795
 Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

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

Dynamic Rate Control for MPEG Compressed Video in ATM Networks

submitted by S. Karademir, BSc.

in partial fulfillment of the requirements for
the degree of Master of Engineering

Thesis Supervisors

Chair, Department of Systems and

Computer Engineering

Carleton University

September 9, 1996

Abstract

Traffic congestion avoidance and control is a major design concern in ATM (Asynchronous Transfer Mode) networks. The large propagation delay bandwidth product and the diverse service requirements make it difficult to control the congestion in ATM networks. Particularly, the effect of the information loss on compressed video due to congestion is vital because the real-time requirements of video do not tolerate re-transmission of the lost cells.

This thesis proposes an effective rate control scheme for MPEG compressed variable bit rate video sources in ATM networks. The scheme controls the rate at which traffic is submitted to the network depending on the congestion status information provided from the network. A congestion notification mechanism which was proposed for rate control of available bit rate traffic to the ATM Forum is utilized.

The simulation results show that the proposed scheme achieves maintaining a low output link buffer occupancy, resulting in smaller cell delay variation for video connections. Furthermore, it is demonstrated that the number of video sources which the network can accommodate can be increased significantly without violating the quality of service requirements of the connections.

Acknowledgements

I would like to thank Dr. M. Devetsikiotis and my thesis supervisors, Prof. I.E. Lambadaris and Prof. A.R. Kaye for their guidance and encouragement throughout the course of my thesis. I appreciate their helpful suggestions and the patience they shown with me during my study period here at Broadband Networks Laboratory.

I gratefully acknowledge the funding support provided to my research work by the Telecommunications Research Institute of Ontario.

I would like to thank to my colleagues Ms. Rukhsana Ansari and Mr. Changcheng Huang for their valuable suggestions.

Last but not the least, I thank to my husband, Hilmi, for his continuous support, endless patience and many valuable suggestions. Special thanks are due to my family, especially, to my father for his constant encouragement.

To my father

Table of Contents

Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	vi
List of Tables	ix
List of Figures	xii
List of Acronyms	xiii
List of Symbols	xiv
1 Introduction	1
1.1 Background	1
1.2 Previous Research	3
1.3 Thesis Objective	6
1.4 Thesis Organization	7
1.5 Publication	8
2 ATM and MPEG Compressed VBR Video	9
2.1 ATM Overview	9
2.1.1 B-ISDN Protocol Reference Model	10

2.1.2	Operation and Maintenance of the B-ISDN UNI at ATM Level	16
2.2	MPEG Compression	18
2.3	Variable Bit Rate Coded Video	20
3	Composite Rate Control Scheme (CRCS)	23
3.1	Introduction	23
3.2	Overview of Composite Rate Control Scheme	24
3.3	Quantizer Control	25
3.4	Feedback Mechanism	26
3.5	Switch Behavior	28
3.6	Adaptive Rate Control	28
3.6.1	Balanced Mechanism	28
3.6.2	Modified Algorithm	34
4	Simulation Study	36
4.1	Network Configuration	36
4.2	Source Modeling	38
4.2.1	Data Source Model	38
4.2.2	Video Source Model	38
4.3	Quality of Service Requirements for MPEG Video	42
4.4	Calculation of End-to-End Delay for Video	43
4.5	Output Analysis of the Simulation Study	44
4.5.1	Transient Removal	44
4.5.2	Simulation Run Length and Number of Runs	46

5	Experimental Results	48
5.1	Introduction	48
5.2	Comparison of Balanced Mechanism and Modified Rate Control Algorithms	49
5.3	The Effect of Buffer Size on the Performance Metrics	53
5.3.1	Cell Loss Rate	56
5.3.2	Cell Delay	57
5.4	Efficiency of the Algorithm in Heavily Loaded Networks	59
5.5	Implications of the Algorithm on Queueing Delay Distribution	62
5.6	The Effect of Propagation Delay on Performance Metrics	65
5.7	The Effect of Background Source Traffic on the Performance Metrics	66
6	Conclusions and Recommendations for Future Research	71
6.1	Conclusions	71
6.2	Recommendations for Future Research	73
	References	74

List of Tables

2.1	Properties of ATM/B-ISDN service classes.	14
2.2	ATM Layer OAM Functions at the UNI	17
4.1	Statistics of the sequence "football.mpg"	39
5.1	Data source traffic parameters	67

List of Figures

2.1	B-ISDN Protocol Reference Model	11
2.2	ATM cell payload	15
2.3	Constant and variable bit rate transmission.	21
3.1	Block diagram of the composite rate control scheme.	25
3.2	Structure of ATM layer OAM cell.	27
3.3	Operating cycles and important parameters of Balanced Mechanism. .	29
3.4	Flow chart of the rate control algorithm used in BM.	31
3.5	(a) Optimistic and (b) Conservative rate adjustment functions vs. q (feedback) for various values of $f(q)$ denoted by y	32
4.1	Simulation network topology.	37
4.2	Finite state machine model of the video source.	40
4.3	Slice content of the frames between frame 218 and frame 250 of "foot- ball.mpg" sequence.	41
4.4	Averaged batch means for end-to-end cell delay.	45
4.5	Cumulative average of end-to-end cell delay.	46
5.1	The traces of calculated transmission rate and QIL using the BM in medium network load.	50

5.2	The traces of calculated transmission rate and QIL after the first set of modifications in medium network load	51
5.3	Flow chart of the rate adjustment algorithm after the second set of modifications.	52
5.4	The trace of Quantization Index Level of the reference connection after the second set of modifications.	53
5.5	The trace of total queueing delay and its marginal histogram	54
5.6	The histograms of delay and QIL in heavy load.	55
5.7	The effect of switch buffer size on average cell loss rate of reference video cells.	57
5.8	The effect of switch buffer size on average buffer occupancy of the bottleneck switch.	58
5.9	The effect of switch buffer size on average total queueing delay of reference video cells.	59
5.10	Number of video sources that the network can accommodate versus the utilization of the output port buffer in the bottleneck switch	60
5.11	Mean QIL vs. number of video sources.	61
5.12	CDF of Quantization Index Level of (a) I-P frames, (b) B frames	63
5.13	The histogram and the CDF of total queueing delay.	64
5.14	The effect of propagation delay on (a) mean cell loss rate and (b) mean quantization index level of the reference connection.	65
5.15	The effect of total propagation delay on (a) mean total queueing delay of the reference connection cells and (b) mean buffer occupancy of the bottleneck switch.	67

congestion, and the N -th outgoing cell to $EFCI=0$ implying that there is no congestion. It continues marking its cells in this fashion until the end of the connection. If a cell marked to $EFCI=0$ is able to reach the destination end, the destination end generates a RM cell and sends it back to the source. Unless the source receives such a RM cell it decreases its cell rate. The increments and the decrements in the cell rate of the VC are proportional to the current traffic rate of that VC.

Although the PRCA prevents the potential control failure in FECN, it still has a disadvantage which is known as the "beat down" problem: When a VC passes through more congested links, its cells are marked to $EFCI=1$ more often. As a result, this VC will have a lower allowed cell rate than other VCs going through fewer congested links. This may cause severe unfairness among the sources. The improvement to this scheme is brought by the APRC (Adaptive Proportional Rate Control) [14]. APRC uses a special type of RM cells which travel from the source to destination and back to the source again. This RM cell carries modified allowed cell rate calculated by the intermediate switches. Upon reception of the RM cell, the source adapts its rate according to the information written on the RM cell.

The EFCN, PRCA and APRCA are among the best-known algorithms proposed and discussed in the ATM Forum [14]. All of these proposed schemes use closed loop control which has been accepted as an effective way to control congestion in ATM networks.

Incorporating network status with the video coding mechanism has been investigated for frame relay [20] and, in general, for fast packet switching networks [21]. In [20], it has been shown that transmission of H.261 compressed video over frame relay is feasible if a BECN type feedback is used and the H.261 codec is modified

List of Acronyms

AAL	ATM Adaptation Layer
ABR	Available Bit Rate
APRC	Adaptive Proportional Rate Control
ATM	Asynchronous Transfer Mode
BECN	Backward Explicit Congestion Notification
BISDN	Broadband Integrated Services Digital Network
BM	Balanced Mechanism
CAC	Call Admission Control
CBR	Constant Bit Rate
CCITT	International Telegraph and Telephone Consultative Committee
CDF	Cumulative Distribution Function
CODEC	Coder/Decoder
CPU	Central Processing Unit
CRCS	Composite Rate Control Scheme
CS	Convergence Sublayer
DCT	Discrete Cosine Transformations
EFCI	Explicit Forward Congestion Indication
FECN	Forward Explicit Congestion Notification
FIFO	First In First Out
ITU-T	International Telecommunications Union-Telecommunications
JPEG	Joint Picture Expert Group
MPEG	Motion Picture Experts Group
NTSC	National Television System Committee
OAM	Operation, Administration and Management
PDU	Protocol Data Unit
PRCA	Proportional Rate Control Algorithm
QIL	Quantization Index Level
RM	Resource Management
SAR	Segmentation and Reassembly
TCP	Transmission Control Protocol
TES	Transform-Expand-Sample
UNI	User Network Interface
UPC	Usage Parameter Controller
VBR	Variable Bit Rate
VC	Virtual Channel
VCI	Virtual Channel Identifier
VP	Virtual Path
VPI	Virtual Path Identifier

List of Symbols

C	cell transmission rate
$[C_m, C_M]$	the limits for the degree of confidence
D	buffer division factor
D_D	decoding delay
D_E	encoding delay
D_P	packetization delay
D_M	maximum allowable end-to-end delay
D_N	maximum allowable network delay
f_c	degree of confidence used in F_c calculation
f_o	degree of confidence used in F_o calculation
F_c	conservative function
F_o	optimistic function
l	average burst length
L	buffer occupancy level of Codec buffer
q	utilization obtained from the RM cell
Q	quantization index level of CBR Codec
Q_{offset}	offset defining the minimum of Q
q_t	target utilization
r_i	token rate during the cycle i
$\hat{r}_i$	estimated token rate
r_m	minimum token rate/minimum video source transmission rate
r_M	maximum token rate/maximum video source transmission rate
s_i	source rate during the cycle i
$\hat{s}_i$	estimated source rate
s_{i-1}	source rate measured during the cycle $i - 1$
t_i	transmission rate during the cycle i
t_{i-1}	transmission rate measured during the cycle $i - 1$
$\hat{t}_i$	estimated transmission rate
T_s	averaging period for utilization in the switch
λ	average burst arrival rate
ρ	average data source load

Chapter 1

Introduction

1.1 Background

The Asynchronous Transfer Mode (ATM) technology has been identified as the transport mechanism that will unify the present voice, image and data networks [1]. Advantages of ATM include flexibility, bit rate independence and the ability to switch various types of traffic in a single fabric. ATM also offers potential gains in bandwidth efficiency by buffering and multiplexing bursty traffic [2, 3, 4, 5].

The services ATM will support have a wide range of QoS (Quality of Service) requirements. Data applications are very sensitive to loss, even though, they can tolerate delay. At the same time, voice and video applications have very stringent end-to-end delay requirements [6]. In addition to diverse QoS requirements, some services are expected to generate smooth traffic (e.g. voice), while others, such as video and data, are expected to generate bursty traffic making these services ideal for statistical multiplexing.

If the connections are assigned a bandwidth lower than their peak bit rate, the network can serve more sources. The disadvantage of this approach is that it can cause congestion. Congestion occurs when the network becomes unable to guarantee

the negotiated QoS requirements to the already established connections and to the new connection requests [7]. During the congestion period, the delay increases and cells are discarded causing more retransmissions by the applications and resulting in an even higher load in the network. Real-time applications such as digitized voice and video which have very stringent performance requirements in terms of delay, delay jitter and loss rate cannot afford retransmissions: therefore, congestion control is a challenging problem for networks serving these types of applications [8, 9].

A very high bit rate and relatively large propagation delay make the control issue even more challenging especially for guaranteed quality services. Due to the large propagation delay-bandwidth product, there will be a very large number of cells in transit at any time in the network. This makes the network volatile in terms of traffic control.

In the past, many studies claimed that the dynamic control schemes using feedback from the network would fail because of the large propagation delay [10]. Preventive control techniques including Call Admission Control (CAC), bandwidth allocation and traffic policing schemes have been proposed as a solution [11, 12, 13]. It has been suggested that real-time services such as constant bit rate (CBR) and variable bit rate (VBR) voice and video can be accommodated according to the peak cell rate, cell delay variation tolerance (for CBR) and maximum burst length (for VBR) [14]. If the network resources required by a real-time connection cannot be allocated, the request is rejected at call setup time. This would guarantee the QoS requirements but would cause under-utilization of network resources, especially for VBR services that are bursty in nature.

1.2 Previous Research

Closed loop control in ATM networks targeting data traffic has become the main focus of research recently. In the ATM Forum, very important steps have been taken towards standardization of congestion control for ABR (available bit rate) [15].

Originating from the idea of providing an effective means for users to utilize network resources in a dynamic manner, FECN (Forward Explicit Congestion Notification) schemes use the payload type identifier in ATM cells to carry the congestion information [16, 17]. The most commonly used mechanism to detect the congestion in these schemes is monitoring the buffer status. The feedback signal is generated when the switch buffer occupancy exceeds a target value. Upon reception of the feedback the source adjusts its rate using additive increases and multiplicative decreases at fixed intervals. This policy was first suggested by Jain, Ramakrishnan and Chiu in 1987, and was implemented in the TCP (Transmission Control Protocol) [18].

There is a drawback for this type of notification which is based on *negative* feedback. The control may fail if the cells carrying feedback encounter heavy congestion on the feedback route since the source increases its rate unless it receives a congestion indication from the network. In order to overcome such a drawback, the PRCA (Proportional Rate Control Algorithm) is proposed by using a *positive* feedback, in which the source would increase its rate only if it receives an explicit *positive* indication from its destination in the form of a RM (Resource Management) cell (see [14, 19] and references within). The EFCI (Explicit Forward Congestion Indication) state is used indirectly in the generation of the RM cell as follows. Starting from the beginning, the source marks its $N-1$ outgoing cells to $EFCI=1$ implying that there is

congestion, and the N -th outgoing cell to $EFCI=0$ implying that there is no congestion. It continues marking its cells in this fashion until the end of the connection. If a cell marked to $EFCI=0$ is able to reach the destination end, the destination end generates a RM cell and sends it back to the source. Unless the source receives such a RM cell it decreases its cell rate. The increments and the decrements in the cell rate of the VC are proportional to the current traffic rate of that VC.

Although the PRCA prevents the potential control failure in FECN, it still has a disadvantage which is known as the "beat down" problem: When a VC passes through more congested links, its cells are marked to $EFCI=1$ more often. As a result, this VC will have a lower allowed cell rate than other VCs going through fewer congested links. This may cause severe unfairness among the sources. The improvement to this scheme is brought by the APRC (Adaptive Proportional Rate Control) [14]. APRC uses a special type of RM cells which travel from the source to destination and back to the source again. This RM cell carries modified allowed cell rate calculated by the intermediate switches. Upon reception of the RM cell, the source adapts its rate according to the information written on the RM cell.

The EFCN, PRCA and APRCA are among the best-known algorithms proposed and discussed in the ATM Forum [14]. All of these proposed schemes use closed loop control which has been accepted as an effective way to control congestion in ATM networks.

Incorporating network status with the video coding mechanism has been investigated for frame relay [20] and, in general, for fast packet switching networks [21]. In [20], it has been shown that transmission of H.261 compressed video over frame relay is feasible if a BECN type feedback is used and the H.261 codec is modified

to control the rate. The adjustment of the rate in response to BECN is achieved by changing either the video frame rate or the quantization factor. It is concluded that quantizer control is more effective in terms of peak buffer occupancy and average throughput.

The authors of [21] propose an adaptive congestion control scheme for packet video transmission to study the transient effects of network congestion on the perceptual quality of video images. In their control scheme, the predicted system state is used for adapting the encoding rate. Each switch in the network keeps track of the buffer occupancy value and the service rate for every single connection. The target sending rate for each frame is calculated on the basis of the difference between the estimated buffer occupancy level and the target value for the connection at the bottleneck switch. The estimated buffer occupancy is calculated using the last known service rate and buffer occupancy level extracted from the feedback packet. In this calculation the change in the buffer occupancy between the current time and the arrival time of the last feedback is taken into consideration making the algorithm still efficient for the networks with large propagation delay. After calculating the target sending rate, the coder tries to match this rate by adjusting the quantization factor. A damping mechanism is imposed to the quantization factor in order to reduce the fluctuations in the video quality.

In [21] it is shown that, by modulating the transmission rate of a video encoder according to the congestion signals from the network, the video quality deteriorates within acceptable limits and the transmission capacity of the network is used efficiently.

Although sound algorithms have been proposed for congestion control for

packet switching networks in general, the integration of network status and the video coding algorithm into a control scheme in order to achieve better usage of network resources has not been fully explored for ATM networks. This area, exploiting strong interactions between the coder and the network, will be the main focus of this thesis.

1.3 Thesis Objective

The objective of this thesis is to:

- i Propose a composite rate control scheme for MPEG coded VBR video sources in ATM networks in order to achieve better usage of network resources. The main point of the control scheme is the control of the rate at which traffic is submitted to the network depending on the congestion status information provided from the network.
- ii Present a practical means for an efficient congestion notification mechanism.
- iii Develop an ATM network simulator using the OPNET simulation package to study the performance of the proposed scheme.
- iv Research and implement the video and data traffic models for ATM networks in OPNET as processes based on finite state machines.
- v Explore, through simulation, the effectiveness of the proposed scheme in terms of resource sharing and video quality.
- vi Study the effect of network parameters such as switch buffer size, propagation delay and the burstiness of the background traffic on the performance of the video.

1.4 Thesis Organization

The remaining chapters of this thesis are organized as follows:

Chapter 2: Reviews the fundamentals of ATM and B-ISDN protocol reference model with a particular emphasis on the functions of the User and the Management Plane, MPEG compression scheme and VBR coded video.

Chapter 3: Presents the integrated approach to implement the rate control for MPEG video traffic. The main idea in this approach is to incorporate the condition of network elements in the source rate control by means of a feedback mechanism. Three aspects of Composite Rate Control Scheme (CRCS) are described in detail: Feedback mechanism, rate controller, and multiple rate video coder.

Chapter 4: Describes the ATM network simulation topology. The OPNET video and the data models used in the simulation are explained. The Quality of Service requirements for the MPEG video are discussed.

Chapter 5: Investigates the performance of the scheme in terms of efficiency and robustness. First, it demonstrates the performance of the Balanced Mechanism's rate control algorithm and the modified rate control algorithm. Second, the effect of the switch buffer size on the performance metrics such as average cell loss rate and the average queueing delay is investigated. Then, the efficiency of the algorithm in terms of resource sharing and bandwidth allocation is demonstrated. The implication of CRCS on queueing delay distribution is

also explored. Finally, the effect of propagation delay and the burstiness of the background sources on the performance metrics of the reference connection is investigated.

Chapter 6: Presents conclusions and recommendations for future research.

1.5 Publication

At the time of submission of this thesis a paper had been accepted for publication [22].

Chapter 2

ATM and MPEG Compressed VBR Video

2.1 ATM Overview

Originally conceived as a switching technique, ATM is now becoming the universal transport technology which is deployed by service providers around the world. Increasing demand for the higher bandwidth services in a more cost effective way by the enterprise network customers, and the drivers for the residential broadband services such as internet access, home-based businesses and entertainment service features are the main driving forces behind the ATM broadband networks.

In Asynchronous Transfer Mode the transmission bit stream is divided into a number of fixed-length *cells*. The fixed cell size simplifies the implementation of the ATM switches and the multiplexers and enables implementations at very high speeds [23]. The term "asynchronous" in ATM refers to the fact that cells from any particular source need not be periodically spaced within the overall cell stream. Unlike in synchronous transfer mode (STM) which identifies a data unit associated with a given connection by its position in the transmission frame, in ATM a cell

belonging to a connection may take place at any position in the output stream.

Connections are established by setting up translation tables at switching and multiplexing points which map an incoming connection to an outgoing connection. The connections established in this manner are called *Virtual Circuits* since no bandwidth is consumed unless traffic is actually being transmitted. ATM connections may operate at either a Constant Bit Rate (CBR) or a Variable Bit Rate (VBR).

ATM protocols are designed to handle time critical data such as telephony and video in addition to more conventional inter-computer data communications. The following section describes the B-ISDN protocol model with a particular emphasis on the first three layer of user plane, and the management plane.

2.1.1 B-ISDN Protocol Reference Model

The ATM based B-ISDN Protocol Reference Model (PRM) which defines a set of principles such as protocol layering and layer service definition is given in Figure 2.1 [24]. The B-ISDN PRM consists of three planes, namely, User plane, Control plane and Management Plane.

The User Plane provides for the transfer of user information. The Control Plane is responsible for connection control functions. These are signaling functions which are necessary to set up, supervise and release a connection. The Management Plane performs management functions related to resources and the parameters of the protocol entities. It is further broken down into layer management and the plane management. The plane management is an abstract concept which is planned to provide coordination between all the planes. The layer management handles the

Figure 2.1: B-ISDN Protocol Reference Model

Operations, Administration and Maintenance (OAM) information flows specific to the related layer.

The overview of the functions of the physical layer, the ATM layer and the ATM adaptation layer are given in the following subsections. Then, the OAM functions of the ATM layer at User Network Interface (UNI) are explained in order to provide an introductory information for the congestion notification mechanism used in the proposed Composite Rate Control Scheme (CRCS).

Physical Layer

The physical layer supports transmission of ATM cells over a physical medium that connects two ATM devices. It consists of two sublayers: the Physical Medium Dependent (PMD) Sublayer and the Transmission Convergence (TC) Sublayer. The PMD

Figure 2.2: ATM cell payload

At the SAR sublayer, a four-octet overhead consists of a two-octet header and a two-octet trailer. Following is the explanation of the fields that are shown in Figure 2.2:

- **SN (Serial Number)** is an eight-bit field that is used for detecting cell losses during transmission.
- **B/E (Begin and End Tag)** is a two-bit code that is used to inform the receiving node about the boundaries of a particular slice.
- **RES (Reserved)** field
- **PAD (Padding bits)** field is used to fill the cell payload in order to complement it to 48 octets.
- **LI (Length Indicator)** informs the receiver about the useful content of the cell when the cell is partially filled.

The Convergence Sublayer assures the necessary error control and sequencing as well as the sizing of information. On transmission, the CS accepts variable length USER-PDUs (Protocol Data Unit) from the higher layers. It appends a header to the user-PDU and adds 0-3 pad octets to build it out to an integral multiple of 32 bits. After appending the trailer, the data unit is referred to as a CS-PDU which is a collection of the HEADER, USER-PDU, PAD and TRAILER fields. Next, the CS-PDU is passed to the SAR sublayer for segmentation and further transmission. Upon reception, the Convergence Sublayer validates the collection of SAR-PDU payloads received from the SAR sublayer by using the information contained in the CS-PDU header and the trailer. It removes the PAD, if any, and presents the validated CS-PDU to the higher layers.

The Segmentation and Reassembly sublayer is responsible of segmentation of higher layer information into a size suitable for the information field of an ATM cell, and reassembly of the contents of ATM cell information fields into higher layer information.

The AAL maps user information into ATM cells and accounts for transmission errors. It may also transport timing information so the destination can regenerate time dependent signals. The information transported by the ATM layer is divided into four AAL classes currently defined by ITU-T Recommendation I.362. Table 2.1 gives the properties of these service classes, labeled A through D.

Four AAL services are defined to match up with the four ATM/B-ISDN service classes:

Traffic Classes	Timing Requirement	Bit Rate	Connection Mode
Class A	Time Dependent	CBR	Connection Oriented
Class B	Time Dependent	VBR	Connection Oriented
Class C	Time Independent	VBR	Connection Oriented
Class D	Time Independent	VBR	Connectionless

Table 2.1: Properties of ATM/B-ISDN service classes.

AAL1

AAL type 1 is used for CBR services that require a strict timing relationship between the end points of connections. Telephony, high-quality audio and video are the examples of such services.

AAL2

AAL type 2 specifies ATM transportation of connection-oriented circuit and VBR video and audio. The standardization of AAL2 has not yet been completed by the ITU-T [8, 23, 26]. On the other hand the general structure of AAL2 is defined in [27]. Based on this structure an AAL2 protocol for VBR MPEG video traffic has been proposed in [26]. Figure 2.2 depicts the ATM cell payload with AAL2 headers and trailers. The structure, given in [26], is used in our study for the calculation of the size of the AAL layer overhead.

The header in each MPEG frame contains timing fields that encode time stamps for the presentation of that frame at the decoder. Hence, the CS does not have to provide timing information for MPEG-PDUs, i.e., slices. The CRC function is assumed to be performed at the SAR layer. Therefore, the proposed AAL protocol does not consider any CS layer overhead.

Figure 2.2: ATM cell payload

At the SAR sublayer, a four-octet overhead consists of a two-octet header and a two-octet trailer. Following is the explanation of the fields that are shown in Figure 2.2:

- **SN (Serial Number)** is an eight-bit field that is used for detecting cell losses during transmission.
- **B/E (Begin and End Tag)** is a two-bit code that is used to inform the receiving node about the boundaries of a particular slice.
- **RES (Reserved)** field
- **PAD (Padding bits)** field is used to fill the cell payload in order to complement it to 48 octets.
- **LI (Length Indicator)** informs the receiver about the useful content of the cell when the cell is partially filled.

- CRC (Cyclical Redundancy Code) is a field used to protect the SAR-PDU from bit errors.

In our simulation study, we use AAL2 as the AAL layer. The assumptions that are explained here are used with regard to the calculation of the cell overhead.

AAL3-4

AAL3 and AAL4 are combined into a single common part AAL3-4 in support of VBR traffic for both connection oriented and connectionless service. Support for connectionless service is provided at the Service Specific Convergence Sublayer level. The functions of AAL3-4 includes segmentation, reassembly and error control.

AAL5

AAL type 5 supports VBR traffic for both connection oriented and connectionless service. This protocol is defined to improve the efficiency of the adaptation layer. The main difference between the AAL2 or AAL3-4 and AAL5 is that the CRC is performed at the CP level instead of SAR level. Implementing the CRC before the user-PDU is divided into cells reduces the AAL layer overhead substantially.

2.1.2 Operation and Maintenance of the B-ISDN UNI at ATM Level

Before proceeding further into the specific OAM functions for ATM layer a brief explanation of the OAM philosophy is in order. *Operations* involve the day-to-day, and often minute-to-minute monitoring of the network in order to ensure that it is fulfilling its design purpose. *Monitoring* involves watching for faults and invoking

Functions	Parameters
Alarm Surveillance (VP)	OAM cells
Connectivity Verification (VP)	OAM cells
Invalid VPI/VCI Detection	VPI/VCI

Table 2.2: ATM Layer OAM Functions at the UNI

maintenance actions to repair them. Operations coordinate actions between administration and maintenance throughout all the phases of the ATM connection's life. *Administration* includes a set of activities such as network design, processing orders, assigning addresses, and tracking usage. *Maintenance* involves the unavoidable circumstances when it is necessary to diagnose a problem and repair it. Performing a routine maintenance so that the performance degradations and faults are less likely to happen is an important maintenance activity.

OAM functions of the ATM layer supported at the UNI are classified in three function groups as shown in Table 2.2. ATM cells with invalid VPI/VCI labels are discarded and Layer Management is informed. In order to exchange *operations* information for the performance of alarm surveillance and connectivity verification, OAM cells are used. OAM flows are defined for the exchange of information between nodes. The ATM layer contains two OAM levels in the form of OAM flows: F4, virtual path level flow; F5, virtual channel level flow [28]. The Virtual Path Connection (VPC) operational information is carried via the F4 flow OAM cells. These cells have the same Virtual Path Identifier (VPI) as the user cells but are identified by pre-assigned Virtual Channel Identifier (VCI). The Virtual Channel Connection (VCC) operation information, on the other hand, is carried via the F5 flow OAM cells. The identifier for these cells is the preassigned code in the Payload Type (PT) field of the OAM cell

structure. The VPI/VCI values are the same as user cells for F4 flow OAM cells. The OAM cell flow used for end-to-end operations functions may be carried transparently through the private ATM switch and made available to the user [29]. This concept, i.e., using RM cells for rate control purposes, was proposed and accepted in ATM Forum [30].

The proposed congestion notification mechanism in this study suggests the use of OAM cells using either OAM flow. The details of the mechanism will be presented in Section 3.4.

2.2 MPEG Compression

Standardization of video compression techniques has become a high priority after the development of digital video technology for two decades. The Moving Picture Expert Group (MPEG), a joint international committee for image sequence compression, has started the effort to develop a standard for video and associated audio on digital storage media, such as CD-ROM, tape drives, as well as telecommunication channels [31].

The MPEG algorithm encodes video frames in a sequence using three different algorithms. A DCT-based algorithm first encodes intraframes (I frames) reducing the spatial redundancy. Then, two prediction techniques are used to exploit temporal redundancy between frames. One encodes predicted frames, P frames, with reference to a previous frame with motion compensation. The other uses bidirectional or interpolated prediction to encode B frames¹. Interpolated prediction uses a previous or a future frame (I or P) to code a B frame achieving the highest amount of compression.

¹The words frame and picture are used here interchangeably.

There is a trade-off between the coding gain provided by the interpolated prediction and the cost associated with coding interpolated prediction. An increasing number of B-pictures improves the coding efficiency but increases the encoding delay because of the buffering required for frame reordering. Also, an increasing number of B frames decreases the correlation of B pictures with the references and the correlation between the references themselves. It is stated in [31] that placing the reference frames at approximately 100 ms apart seemed reasonable for a large class of scenes translating into placing two B frames in between reference frames.

An MPEG coded image sequence consisted of following layers in a top-down fashion[32].

- Video Sequence Layer
- Group of Picture Layer
- Picture Layer
- Slice Layer
- Macroblock Layer
- Block Layer

The video sequence layer contains the signaling information such as the dimensions of the frame, the pel-aspect ratio, the picture rate. The next layer is the GOP layer which consists of an I frame and a number of P and B frames. A GOP structure is defined by M which is the number of B frames that are inserted in two reference (I or P) frames plus one, and N which is the number of reference frames. The total number of frames

a GOP contains is obtained by multiplication of M and N . The picture layer contains all coded information for one picture. It describes the location of the current video frame with respect to the GOP layer, picture type and the order in which the pictures must be displayed at the destination. Pictures are divided into slices which are placed end-to-end in a raster scan across the image. The slice layer consists of macroblocks. A macroblock is the smallest image unit for motion compensation. The composition of a macroblock from an input image depends on the ratio of the luminance and chrominance components of the input image. A ratio of 4:1:1 (4 luminance and 2 chrominance component blocks) is used in the MPEG standard because the human visual system is more sensitive to variations in brightness (luminance) than in colour (chrominance). The block layer, the bottom layer, represents an 8×8 sample array of the image.

The main features of the MPEG coding which distinguishes it from the other coding algorithms such as JPEG or H.261 are the following: First, it uses bidirectional temporal prediction to achieve a high degree of compression. Second, it can compress entertainment quality video. These are the reasons which make the MPEG coding scheme attractive for many video applications [21, 5, 31]. On the other hand, increased compression efficiency leads to even greater burstiness in addition to the inherent burstiness of the VBR video as mentioned in the previous section. This makes the transmission of MPEG video very challenging from the network point of view.

2.3 Variable Bit Rate Coded Video

Traditionally, video has been transmitted using constant bit rate (CBR) channels by using a buffer at the encoder to translate the variable output of the encoder into a CBR

Figure 2.3: (a) Constant and (b) Variable Bit Rate Transmission

stream. This eliminates burstiness, but results in variable picture quality. In order to achieve constant image quality, and to take advantage of statistical multiplexing, VBR codecs can be used instead.

Figure 2.3 illustrates the basic differences between CBR and VBR transmissions. The CBR encoder has a feedback input from the output buffer. In order to prevent the buffer from underflowing or overflowing the encoder changes certain coding parameters such as the quantization index. For example, when there is a scene change or a scene with an active background the frames with large information content fill the output buffer. The feedback, then, informs the encoder about the buffer occupancy. The encoder reacts by increasing the quantization index, resulting in a perceptual quality degradation.

In the case of VBR, on the other hand, the encoder's output is connected to the

transmission channel directly, eliminating the feedback from the buffer. Other than having constant quality, the VBR codec has the following advantages [33]: reduction of total delay by eliminating the buffering at the encoder's output, user control of video quality instead of the fixed rate channel control, and better utilization of network resources when the VBR channels are available. The drawback of using VBR sources in the network is the burstiness which makes it more difficult to predict and control.

Chapter 3

Composite Rate Control Scheme (CRCS)

3.1 Introduction

This chapter presents the Composite Rate Control Scheme for MPEG video traffic. The main idea in this scheme is to incorporate the condition of network elements in the source rate control by means of a feedback mechanism. First, the scheme as a whole is described. Then, the three aspects of CRCS, namely, the quantizer control, the feedback mechanism and the rate controller are explained in detail.

The algorithm used in the rate controller is adapted from the study by Gong and Akyildiz [34] known as *Balanced Mechanism* (BM). The rate calculation method of BM with certain modifications is used in this study to adjust the cell emission rate of the video source. Following the rate control algorithm of Balanced Mechanism, the reasons that lead to modification of the algorithm are discussed briefly. The detailed explanation of modifications is given in two steps in a comparative manner in Section 5.2.

3.2 Overview of Composite Rate Control Scheme

The CRCS rate controller and multiple rate video coder are illustrated in Figure 3.1. The controller continuously observes the network status through *resource management cells* (RM cells). It calculates the target transmission rate and regulates the output of the coder so that it conforms to this rate. The feedback from the network controls the amount of data being encoded. This is done by a rate adaptor located at the user network interface (UNI). The rate adaptor emits RM cells periodically. An RM cell traverses the same path as the source cells do and brings the maximum link utilization along the path back to the rate adaptor.

The rate adaptor calculates the effective transmission rate and enforces it by using a cell spacing mechanism, as shown in Figure 3.1. The cell spacer can be viewed as a server with modulated service rate that serves the source output buffer. The calculated rate is the maximum rate at which the source can submit its cells to the network. If the source rate is greater than the calculated rate for some period of time, the excess cells are buffered in the output buffer of the source.

The occupancy of the source output buffer is controlled by forcing the video codec to select an appropriate quantization index level (QIL). The selection of the QIL is achieved by a local feedback loop from the buffer to the quantizer. If the source rate exceeds the calculated rate momentarily, the excess data is stored in the output buffer of the coder. The coder checks the buffer occupancy periodically, i.e. in the beginning of each video slice [32], and controls the queue length of the buffer by changing the QIL. The buffer occupancy is the means for the rate controller to inform the video coder about the current status of the network.

Figure 3.1: Block diagram of the composite rate control scheme.

Using a buffer to control the output of the MPEG codec is one of the main differences between our scheme and the schemes reviewed in the Section 1.1 which do not use buffering at the source codec. The source rate which is controlled by the buffer occupancy and the quantization level is changed *only* when a threshold value is exceeded in the buffer. As a result of this, sudden changes are smoothed out through the buffer, and quality changes result only when persistent overload occurs in the network.

3.3 Quantizer Control

The excess cells that cannot be transmitted right away because of the cell spacing mechanism in the rate adaptor are buffered in the output buffer of the MPEG codec. The information about the buffer level is conveyed to the quantizer by a local feedback loop as shown in Figure 3.1.

Quantization is the lossy stage in the MPEG coding algorithm. The DCT coefficients of each block are quantized in order to reduce the number of digits required to represent them and to increase the number of zero coefficients. By increasing the Quantization Index Level (QIL), the amount of information in a slice is decreased. Using quantization as a control parameter gives us the ability to achieve graceful degradation in picture quality at the onset of congestion. The QIL is calculated by dividing the buffer occupancy (in cells) by a division factor as in the rate control algorithm of CBR MPEG coders:

$$Q = \frac{L}{D} + Q_{offset} \quad (3.1)$$

where Q is the quantization index level ($3 \leq Q \leq 31$). L is the buffer occupancy level. D is the buffer division factor, and Q_{offset} is the quantization offset which defines the minimum Q level. The division factor, D , is calculated dynamically using a moving average of the cell transmission rate.

3.4 Feedback Mechanism

In a closed loop congestion control scheme the most important element is the way the intermediate nodes notify the sources about their congestion status. Previous studies proposed the EFCI (Explicit Forward Congestion Indication) that has been used in Frame Relay networks successfully. In ATM, it uses a field in the header of the data cell as a single bit indicator of congestion in the forward direction of the connection[16]. It has been also accepted as an optional mechanism for the use of end-systems in the ATM Forum [30]. Recent studies discussed in the ATM Forum have proposed yet another type of notification method: explicit-rate feedback [14, 15].

Figure 3.2: Structure of ATM layer OAM cell.

It is stated that the communication of the explicit-rate can be implemented using the ATM layer OAM flows in the form of ATM RM cells. The OAM flows provide the means for the transportation of monitoring cells to perform management functions. Our scheme makes use of RM cells to provide the rate controller with the feedback information from the network. The field tagged *unused*, shown in Figure 3.2, in the function-specific-field of RM cell can be used to carry the feedback.

The additional traffic that is created by RM cells might be a concern for network management. Since our scheme is based on the utilization of the bottleneck output link as feedback, the information will be the same for all VCs using the same VP (Sources that use the same access and destination node). ATM layer OAM functions are implemented in two levels, namely, the Virtual Path (VP) and the Virtual Channel (VC) levels. In order to reduce the amount of traffic created by VC level RM cells, the VP level could be utilized to carry the feedback. Then, one steady stream of RM cells using the VP level flow will be sufficient for all the sources using the same VP.

3.5 Switch Behavior

In order to provide information about its status, the switch keeps track of the utilization of its output ports. When an RM cell arrives at the switch, its feedback field is read and compared to the output port utilization value by the switch processor. If the port utilization is greater than the one already written in the cell, the new value replaces the previous one indicating that the current switch is more utilized than the previous one. Then, the cell is passed to the next switch.

Since the service time is fixed given the link transmission rate, in order to measure the utilization of a port the switching element simply counts the number of cells transmitted from an output port in a given period, T_s . Then, it multiplies this value with the cell service time and divides it by T_s . The implementation of such a measurement in an ATM switch can be achieved easily with a minor software modification. Since the switch does not have to keep track of the information for each and every single VC, our scheme suggests a practical feedback control for ATM networks.

3.6 Adaptive Rate Control

3.6.1 Balanced Mechanism

The rate control algorithm used in CRCS has been adapted from the *Balanced Mechanism* (BM) proposed by Gong and Akyildiz for non-real-time data traffic sources [34]. They use feedback and traffic prediction in order to adjust the source transmission rate; their scheme is described in some detail in this section. The feedback is obtained from the network via a special cell, called a feedback cell (FC), issued periodically by

Figure 3.3: Operating cycles and important parameters of *Balanced Mechanism* where r_i is the token rate in the average rate policer during the i -th cycle, q_i is the feedback available at the end of the i -th cycle, s_i is the average source rate during the i -th cycle and t_i is the transmission rate during the i -th cycle.

the controller. The controller consists of two leaky bucket UPC's (Usage Parameter Controllers) arranged sequentially to police the peak rate and the average rate. The leaky bucket controllers are called the peak rate policer and the average rate policer, respectively. While the peak rate is kept constant, the average rate is modified based on the feedback.

The FC travels along the virtual connection, and brings back the maximum utilization value of the output ports. The elapsed time between the two FC's is defined as a *cycle*. The FCs are assumed to be given the highest priority in all intermediate nodes; therefore, the cycle is almost constant. The operation of the mechanism and the parameters used in the algorithm are illustrated in Figure 3.3.

The token rate, r_i , is estimated at the beginning of i -th cycle. First of all, r_i

is limited to maximum and minimum token rates, r_m and r_M , respectively.

$$r_m \leq r_i \leq r_M \quad (3.2)$$

Secondly, at the beginning of each cycle the source rate, s_i , and the transmission rate, t_i , are estimated. The smaller is taken as the estimated token rate, $\hat{r}_i$:

$$\hat{r}_i = \min(\hat{s}_i, \hat{t}_i) \quad (3.3)$$

where estimated values are denoted by a hat symbol.

In order to minimize the cell loss, the estimated source rate should be as close to the actual rate as possible. Assuming that the source is not very bursty and the cycle time is sufficiently long, the source rate can be predicted from the source rate measured during the previous cycle:

$$\hat{s}_i = s_{i-1} \quad (3.4)$$

The estimation of the transmission rate is based on the feedback information obtained from the network and the average transmission rate during the last two cycles. These parameters are denoted by q_{i-1} , t_{i-1} and t_{i-2} , as shown in Figure 3.3.

The algorithm compares the feedback value to a target value (q_t) which is decided at the call set-up phase. A feedback value greater than the target indicates the presence of congestion in the network. The functions which are used to calculate the current transmission rate are chosen according to these values as shown in Figure 3.4.

Two main functions are used to calculate the transmission rate: an optimistic function, F_o , and a pessimistic function, F_c , which are given in equation (3.5) and

Figure 3.4: Flow chart of the rate control algorithm used in BM, where q_i is the feedback available at the end of the i -th cycle, t_i is the average rate of traffic submitted to the network at the i -th cycle, and q_t is congestion threshold value. In the equations $f.(q)$ is denoted as f for the sake of simplicity.

Figure 3.5: (a) Optimistic and (b) Conservative rate adjustment functions vs. q (feedback) for various values of $f(q)$ denoted by y .

(3.8), respectively. Selection of these functions depends on the difference between the mean transmission rate during the last two cycles, t_{i-2} and t_{i-1} . The effect of the target utilization, q_t , and the utilization obtained from the network, q , is added to the calculation in the form of another set of functions, given in equations 3.6, 3.7, 3.9 and 3.10. They are called the "degree of confidence" and are defined within the $[C_m, C_M]$ interval. These functions are used as parameters of the main functions mentioned above.

The behavior of the main optimistic and conservative functions are illustrated in Figure 3.5 for various values of "degree of confidence" which are denoted by y in the graphs.

$F_o(q, y)$ is a logarithmic decreasing function of q and y which controls the decreasing speed. Its logarithmic nature helps to maintain a gradual decrease of transmission rate when q increases. In the optimistic function, optimism increases when y increases, and vice versa, optimism decreases when y decreases, as plotted in Figure 3.5. The algorithm illustrates that, the optimistic function is used whenever

the transmission rate in the last cycle is less than the one in the previous cycle. The choice of y depends on the feedback value q .

$$F_o(q, f_o) = (r_M - r_m) \frac{\log(f_o + 1 - qf_o)}{\log(f_o + 1)} \quad (3.5)$$

$$f_{o1}(q) = (C_M - C_m) \frac{1 - q}{1 - q_t} + C_m \quad (3.6)$$

$$f_{o2}(q) = (C_M - C_m) \frac{q_t - q}{q_t} + C_m \quad (3.7)$$

$F_c(q, y)$ is an exponential function which causes a drastic decrease in rate with a small change of q . The rate of decrease depends on the value of y . The greater the y the more conservative the rate calculations become.

$$F_c(q, f_c) = \frac{1}{1 - e^{-f_c}} [(r_M - r_m) e^{-f_c q} + r_m - r_M e^{-f_c}] \quad (3.8)$$

$$f_{c1}(q) = (C_M - C_m) \frac{q - q_t}{1 - q_t} + C_m \quad (3.9)$$

$$f_{c2}(q) = (C_M - C_m) \frac{q}{q_t} + C_m \quad (3.10)$$

One of the advantages of this scheme is that it uses continuous feedback instead of binary feedback as in the case of EFCN (Explicit Forward Congestion Notification). This allows the controller to react to changing conditions of the network in a more precise way. Another reason for choosing this approach is that the controller takes into account the source traffic submitted to the network during the FC return trip

time. The consideration of the previous source rate together with the feedback enables the control scheme to eliminate the negative effect of the large feedback delay. As a result, the controller reacts to a congestion situation more prudently.

3.6.2 Modified Algorithm

We applied the rate adjustment algorithm used in Balanced Mechanism [34] to control MPEG compressed video sources. Instead of controlling the token rate of the leaky bucket, we periodically adjusted the service rate of the output buffer to regulate the source output. Buffering allows the coder to adjust itself to the network conditions and to smooth the variability of the calculated rate providing smoother traffic to the network. This observation will be discussed further in section 5.5.

The transmission rate when calculated using the BM's algorithm of [34] caused certain problems such as cell accumulation in the source buffer, resulting in large end-to-end cell delays. The algorithm is expected to calculate the transmission rate so that most of the cells produced by the source are able to go into the network when there is no overload in the network, i.e. when the maximum utilization reported from the network is always less than the target utilization of the connection.

Although the simulations were conducted in medium network load, the effective transmission rate calculated by the rate adaptor oscillated between minimum and maximum values. Codec output buffer occupancy rose substantially causing violations in the maximum allowable end-to-end delay. This, of course, increased the QIL for the codec resulting in the lowest quality level even when there was no congestion in the network. We believe this is due to the use of a conservative approach even when there is no congestion.

On the basis of these observations the flow diagram and the selection of the functions were altered and resulted in very satisfactory and interesting results. The experimental results obtained before and after the modifications are given in Section 5.2. Before this material is presented, Chapter 4 discusses the simulation study.

Chapter 4

Simulation Study

4.1 Network Configuration

In this section we describe the network configuration and the parameters used in our simulation of rate control algorithms for VBR video sources in ATM networks. A software simulator of mixed video and data traffic for an ATM network has been developed using the OPNET¹ Modeler software package to study the performance of our Composite Rate Control Scheme. OPNET provides an environment for modeling and performance evaluation of communication networks in a hierarchical manner. Two levels are used in our simulator: node level, and process level. The node level consists of the simulated network elements such as buffers and servers, traffic sources, and communication links. The process level provides the finite state machine modeling for the switches and the traffic sources.

The network topology used in the simulation study is illustrated in Figure 4.1. It consists of four switching nodes forming a tandem wide area network. Up to 9 video sources and 2 data sources representing the "cross traffic" are connected to the switches. Since the processing delay in the switch is negligible compared to the

¹OPNET is a registered trademark of MIL 3, Inc.

Figure 4.1: Simulation network topology.

output port buffering delay the ATM switch is characterized by the output queue for each output link. The scheduling policy is assumed to be first-in-first-out (FIFO).

Nodes are assumed to be located 1000 km apart from each other corresponding to a propagation delay of 5 ms (fiber optic cable propagation delay is 5 microsecond per km [8]). The transmission links are assumed to operate at 45 Mbps.

The switch denoted "int2_node" is chosen as the bottleneck by routing all the video and cross traffic through it. The reference connection is defined as the connection between the "v1" and the "dest_host". At the destination node the background traffic is directed to different nodes, which are not shown in the diagram, other than the destination host. Each source starts transmitting at a random time. The lower and upper bounds of the starting time are 0 and 2 sec.

4.2 Source Modeling

4.2.1 Data Source Model

The data sources are modeled using the Batch Poisson Model [35]. Each batch (also referred to as a burst), consisting of multiple ATM cells arrives at the queue instantaneously following a Poisson process with an average rate of λ . The burst size is geometrically distributed with an average length of l cells. The output port queue which the data is destined for has a fixed service rate of C . The load of the queue is defined as :

$$\rho = \frac{\lambda l}{C} \quad (4.1)$$

4.2.2 Video Source Model

The video source model, Figure 4.2, is based on a composite TES (Transform-Expand-Sample) model. The TES process is a non-linear auto-regressive scheme for modeling general autocorrelated traffic streams [36]. The advantage of TES is that it can simultaneously capture the first-order and second-order statistics of the empirical time series.

The 250-second MPEG sequence², "football.mpg" was used as a test sequence for the reference connection in the simulations. The statistics of the sequence are tabulated in Table 4.1. The last column shows the mean cell rate calculated after including the AAL header overhead which is nine bytes per cell. The mean rate of the sequence is 2.7 Mbps (i.e. 6349 cells per second as given in Table 4.1) including

²This sequence was digitized and encoded in the Broadband Network Laboratory, Carleton University from a clip showing a football game

Frame Type	Mean Frame Size (bytes)	Mean Rate (Mbps)	Mean Cell Rate (cells/sec)
I frame	17250	4.14	11765
P frame	15966	3.83	10885
B frame	5826	1.4	3972
Composite	9313	2.23	6349

Table 4.1: Statistics of the sequence "football.mpg"

the AAL layer overhead. Accordingly, the mean offered load by a video source, when the transmission capacity is 45 Mbps, becomes $2.7/45 = 0.06$.

We modeled each frame type (I, P and B) using TESTool³, a visual interactive software environment (TESTool gives the modeler the ability to capture both marginal and autocorrelations of real video stream in a user friendly graphical platform). We incorporated the I, P, B models into OPNET, by dynamically generating the individual frames and interleaving them in the same manner as in the MPEG compression standard [5].

After the generation of a frame, slice level information is produced. Observation from the real video sequences indicate that, during a scene, slice distribution over a frame remains approximately constant [5]. This distribution is affected by the location of the activity in the picture. The higher values indicate the location of the activity while the lower values indicate repetition of details or redundancy. For example, Figure 4.3 shows the slice distribution between frame number 218 and frame number 250 from the MPEG compressed video sequence "football.mpg". The axis shows the slice number, and the ordinate shows the percentage of amount of

³TESTool was developed at the C&C Research Laboratories, NEC USA Inc., Princeton, NJ, U.S.A.

Figure 4.3: Slice content of the frames between frame 218 and frame 250 of "football.mpg" sequence.

information of a frame. The shape of the distribution indicates that there is high activity in the middle of the picture (slice number 4 and number 12) corresponding to the running football players on the field, and low activity everywhere else.

Two scenes are chosen for the model that are typical of a football sequence: i) one with activity on the bottom and on top of a frame (eye would interpret it as background and foreground), corresponding to a scene showing an empty field and the spectators on top, some people on the bottom; ii) one with activity in the middle as explained above.

Since variable bit-rate encoded video is categorized as Class B traffic, AAL Type 2 is chosen for the simulation of AAL. This determines the length of the video information in each ATM cell. Therefore, each cell is assumed to contain 44 octets of video information, 4 octets of Segmentation and Reassembly (SAR) - PDU header, and 5 octets of ATM layer header. After calculation of the slice size (in bytes), the number of ATM cells for each slice is calculated depending on the AAL and ATM

layer header and the trailer sizes.

4.3 Quality of Service Requirements for MPEG Video

The compressed digital video traffic has a number of properties which lead to different quality of service requirements than the other types of traffic. The compression of digital video removes a large amount of redundancy present in the video image. As a result of this, the impact of cell loss on the quality increases. At the same time, the sensitivity of the human visual system to different quality degradations allows a certain amount of cell loss. The tolerable cell loss rate depends very much on the codec type and the error concealment techniques that are used at the destination end. One of the greatest difficulties about doing a simulation study with video models instead of real video data is the lack of criteria about perceptual quality. In order to ensure a good perceptual quality, in the experiments on efficiency, we set the cell loss requirement to zero. We assumed that in the simulation the video cell loss occurs due to the buffer contention in the switch output port buffers and due to exceeding the maximum allowable end-to-end-delay at the destination.

Another factor affecting the quality of the picture is the quantization index level (QIL). The Composite Rate Control Scheme controls the total number of bits per slice by varying the QIL. According to the study ⁴ given in [21], QIL greater than 22 causes some deficiency in the perceptual video quality for broadcast quality video streams. Based on this fact, we set the maximum level allowed for QIL to 22 in the simulations. In order to maintain good quality level for the intracoded (I) and the

⁴The results in this study are gathered using an audience formed by fellow colleagues.

predicted (P) frames, at the same time to keep the source buffer occupancy level low. I and P frames are quantized at the half of the B frame quantization level. Hence, the minimum QIL for I and P frames are set to 3 while it is set to 6 for B frames.

The delay requirements, on the other hand, depend on the video service. Interactive services such as video-telephony and teleconferencing, require bounded cell delay. The maximum end-to-end cell delay for interactive video services is given as 350 ms in [6]. In our simulation, this value has been chosen as the maximum cell delay for video cells. Cells delayed more than this value are assumed to be discarded at the destination.

4.4 Calculation of End-to-End Delay for Video

In VBR transmission the end-to-end delay is the sum of the encoding delay, the packetization delay, network delay, and the decoding delay. The typical values for these delays are given in [37] in terms of frame times ($1/30$ s). We used these values in order to set the the maximum allowable network delay for the simulations.

- **Encoding Delay (D_E):** This is the time required to convert the analogue signal generated by the video source to the digital signal transmitted into the network. This delay depends on the coding algorithm and the software and/or the hardware used in the coding. A typical encoding delay of an MPEG Algorithm is $M + 1.3$ frames where M is the number of B frames that are inserted in two reference frames (I or P) plus one. In the MPEG codec $M = 3$ is typically used. Thus the encoding delay is approximately:

$$D_E = (3 + 1.3) * frame_time = 4.3 * 1/30 s = 143.33 ms$$

- **Decoding Delay (D_D):** It is stated in [37] that a half-frame delay is necessary in the post processor of the decoder to convert the signal into the interlaced one.

Thus the decoding delay is:

$$D_D = 0.5 * 1/30 s = 16.66 ms$$

- **Packetization Delay (D_P):** Assuming that the payload in at least one cell is transmitted for each slice, the worst case packetization delay is one slice time interval. There are 15 slices in each frame, thus the packetization delay :

$$D_P = 1/15 * 1/30 = 2.22 ms$$

The maximum cell delay tolerance, D_{max} , for real time video is 350 ms. Thus, the maximum cell delay tolerance for network delay, D_N , is:

$$D_N = D_{max} - (D_E + D_D + D_P)$$

$$D_N = 350 - (143.33 + 16.66 + 2.22) = 187.8 ms$$

In the simulations, the maximum allowable network delay including codec output buffer delay, queueing delay in the switches and the propagation delay is assumed to be 180 ms allowing 7.8 ms for conservative design. Cells delayed more than this value are discarded at the destination.

4.5 Output Analysis of the Simulation Study

4.5.1 Transient Removal

In steady state simulations the influence of the initial conditions should be reduced. When the simulation starts with an empty and idle network, for a finite period of time called *transient time*, the observed performance statistics will always have a

Figure 4.4: Averaged batch means for end-to-end cell delay. Each batch corresponds to 10 sec.

worsening trend. In order to reduce this trend in steady state simulations, averaged batch means and cumulative averages with different deletion methods have been used [38]. The total queueing delay composed of the delay at the codec's output buffer and the delay due to the buffering for the reference video connection was chosen as the response variable. From Figure 4.4 and Figure 4.5 it is apparent that downward bias is present for the first observation (first batch in Figure 4.4) or for the "no deletion" labeled graph in Figure 4.5. In order to reduce this trend data collection was started after 10 second of simulated time had elapsed. Noticing that in "one deletion" and "two deletions" graphs downward bias seem nonexistent, the transient time due to the initial conditions defined as 10 sec.

Figure 4.5: Cumulative average of end-to-end cell delay. 1: no deletion. 2: one deletion. 3: two deletions

4.5.2 Simulation Run Length and Number of Runs

In order to achieve a 95% confidence interval with an accuracy of 1 ms, a simulation run length of 160 ms and 10 replications were found sufficient. Following are the results obtained from the experiment conducted to determine the simulation parameters:

- Run Length: 150 sec.
- Number of Replications, n : 9
- Variance ($S^2(n)$): 1.57×10^{-6}
- Standard Deviation ($S(n)$): 0.0013

An approximate $100(1 - \alpha)$ percent confidence interval for an unbiased point estimator, $\hat{x}(n)$, of the variable network delay D_{var} is given by:

$$\hat{x}(n) \pm t_{(n-1), \frac{\alpha}{2}} \times \sqrt{\frac{S^2(n)}{n}} \quad (4.2)$$

Substituting numerical values, the interval is:

$$\begin{aligned} 0.0175 \pm t_{8,0.025} \times \sqrt{\frac{1.57 \times 10^{-6}}{9}} \\ 0.0175 \pm 9.6 \times 10^{-4} \end{aligned}$$

For a 95% confidence interval and a specified accuracy of ϵ we can calculate R , the number of replications required according to following formula:

$$R \geq \left(\frac{t_{\frac{\alpha}{2}, R-1} S(n)}{\epsilon} \right)^2 \quad (4.3)$$

resulting in a minimum number of replications.

$$R \geq 8.37 \quad (4.4)$$

Chapter 5

Experimental Results

5.1 Introduction

In this chapter we present the results of the experimental setup for testing the performance of the Composite Rate Control Scheme (C RCS). The performance metrics considered are end-to-end delay, cell loss rate and the quantization index level of the reference connection. The reference connection is composed of the video source denoted by "v1", the transmission links and the ATM switches between the video source and the destination host denoted by *dest_host* in Figure 4.1.

The control mechanism is applied only to the video sources. The offered load for each video source is 6% of the total link capacity, 45 Mbps, which is explained in Section 4.2.2. In the simulations up to nine video sources have been activated producing 54% of the traffic that the network can carry. In order to obtain certain network load, we adjusted the parameters of the background sources which are mean burst length and burst interarrival time.

The feasible values for the dynamic rate control parameters required a fair amount of experimentation and intuition. For example, we set the minimum transmission rate, r_m , to the mean rate of the B frames, 4000 cells/sec, given in Table 4.1

since the B frames are the most compressed frame type in the MPEG compression scheme. The maximum transmission rate, r_M , is the maximum I frame rate measured from the real sequence, 30000 cells/sec. By setting the target utilization, q_t , to a very high value, 0.95, we aim to show that even at this value the algorithm efficiently prevents the network overload from causing violations in cell loss of the reference video source, without a visible degradation in the perceptual quality.

5.2 Comparison of Balanced Mechanism and Modified Rate Control Algorithms

In this section we present the reasons that lead to the modification of the Balanced Mechanism (BM) which was originally proposed for data traffic. The experiments are carried out in order to obtain best results in terms of video quality (using QIL) and total queueing delay of the reference connection.

MODIFICATIONS-I

We tested the BM in medium network load. The network consisted of four video sources and a background data source. The average utilization of the bottleneck switch was approximately 0.5. The target utilization was set to 95%. Observations of the traces from the simulations, given in Figure 5.1, indicated that the algorithm behaved conservatively in the calculation of the transmission rate. The calculated rate was observed to be below the minimum rate, r_m , causing the source to increase its QIL (i.e. reduction of video quality) which remained above 20, during 50% of the observation period. One of the reasons for this type of behavior was found in the BM's flow chart, shown in Figure 3.4. When the measured transmission rate in the current cycle, t_{i-1} , is greater or equal than the one measured in the previous cycle,

Figure 5.1: The traces of calculated transmission rate (left) and QIL (right) using the BM in medium network load.

t_{i-2} , or the network is congested, the calculated rate was chosen as the smaller of t_{i-1} and the output of the chosen function, namely, F_o or F_c . This behavior may be attributed to the bursty nature of the video traffic and the buffering at the output of the codec which was not used in [34]. We changed the formulation so that the value for transmission rate is obtained directly from the optimistic, F_o , or the conservative function, F_c .

With this change, the results improved, but did not fulfill the quality requirements we assumed for the simulation study. Although the simulations were conducted in medium network load, the calculated transmission rate oscillated between the minimum and the maximum values. This oscillation caused unnecessary increases in QIL. The traces of both the transmission rate and the QIL are plotted in Figure 5.2.

MODIFICATIONS-II

The next set of changes were made based on the following observation: In BM, if the bottleneck utilization, q_{i-1} , is less than the target utilization, q_t , then either the

Figure 5.2: The traces of calculated transmission rate (left) and QIL (right) after the first set of modifications in medium network load.

optimistic function or the conservative function was used depending on the difference between the average transmission rates of the last two cycles. If the transmission rate increased, the current transmission rate was calculated by the *conservative function*, anticipating an increase in the amount of submitted traffic. This intuitive estimation works well for the sources with less bursty characteristics as presented by Gong and Akyildiz [34]. However, the MPEG compressed VBR video is much burstier (and therefore less predictable) leading to inadequate estimation of overload using the above approach. In another words, an increase in the transmission rate compared to the previous cycle is not always an indication of the congestion. Since the conservative function is an exponentially decreasing function of the bottleneck utilization, q , the calculation gives quite low values if q is equal or greater than 0.4 even if a small η value is used, as shown in Figure 3.5. Using a conservative function creates sharp unwanted variation (spikes) of the QIL because of an increase in the transmission

Figure 5.3: Flow chart of the rate adjustment algorithm after the second set of modifications.

rate.

Based on this observation, we modified the rate control algorithm of BM by changing the logic of the flow in BM's flow chart into the one shown in Figure 5.3. The conservative function is only used if there is a congestion in the network, and the optimistic function otherwise. The change in the source rate is accommodated by using appropriate $f(\cdot)$ functions.

The simulations gave satisfactory results after the modifications for both heavy load and the medium load situations. The QIL trace in Figure 5.4 shows that, after the second set of modifications the rate adaptor does not apply any control over the transmission rate. The QIL values alternate between 3 and 6, favorably, as in the original MPEG compression scheme. The compression scheme uses QIL value 3 for I

Figure 5.4: The trace of Quantization Index Level of the reference connection after the second set of modifications.

and P frames and value 6 for the B frames.

The end-to-end queueing delay the source cells experience, including the buffering at the source, has also improved as a result of the modification, mainly because of the decrease in the source buffering delay. The comparison of the total queueing delay after the first and second set of modifications is presented in Figure 5.5.

The algorithm was also tested in heavy load situations where the output port utilization of the bottleneck switch was increased to 0.95. The improvements in QIL and end-to-end delay are shown in the form of marginal histograms in Figure 5.6.

5.3 The Effect of Buffer Size on the Performance Metrics

When there is a heavy load in the network, the feedback carries this information back to the video sources. In turn, the sources reduce their transmission rates in order to reduce the likelihood of cell loss during their connection period. The impact of this

Figure 5.5: The trace of total queueing delay and its marginal histogram, above: before the second set of modifications, below: after the second set of modifications

Figure 5.6: The histograms of delay and QIL in heavy load. Above: before the second set of modifications, below: after the second set of modifications

behaviour on the buffer occupancy in the intermediate switches was observed in our simulations. Since both video and data sources are bursty in nature, average values gathered from the simulations do not give a complete picture while peak values may be more indicative of the network conditions during overloads. In order to evaluate the performance of the proposed scheme in a heavily loaded network (99%) we set the size of the switch output port buffers to values between 5000 and 17000 cells and observed the performance metrics for the reference connection. In this set of experiments we chose the background data source parameters as follows: burst length: 90 cells, burst interarrival time: 2.83 ms, and offered load: 30%.

The results indicated that the algorithm was very effective in the case of severe congestion in terms of resource sharing. In the following subsections the implications of these results in terms of chosen performance metrics are investigated in detail.

5.3.1 Cell Loss Rate

The main goal of this experiment was to determine the required buffer sizes for a given cell loss rate for both uncontrolled and controlled cases. For buffer sizes less than 16,000 cells a cell loss rate of 2% was observed in the uncontrolled case as shown in Figure 5.7. In order to decrease this rate we increased the buffer size since the cell loss was due to the buffer contention in this case. It was observed that, the mean cell loss rate for the cells of the reference connection increased exponentially with the increasing buffer size, reaching intolerably high values. The reason for the pronounced loss is the increase in the end-to-end delay of the reference video cells. Large switch buffer sizes cause a rise in buffering delay that increases end-to-end delay of the video cells beyond the maximum tolerable delay (350 ms). The video cells delayed more

Figure 5.7: The effect of switch buffer size on average cell loss rate of reference video cells: (1) without control. (2) with control.

than this value are discarded at the destination host.

In the controlled case, however, practically no cell loss rate occurs, suggesting that the algorithm was insensitive to the change in the switch buffer size. The mean bottleneck buffer occupancy, as shown in Figure 5.8, explains this behavior: the buffer occupancy is bounded in the controlled scheme.

5.3.2 Cell Delay

Figure 5.9 illustrates the mean queueing delay composed of queueing delays in the intermediate switches and the codec output buffer. At greater buffer sizes the bottleneck buffer occupancy reaches values that result in very high delay. The cells with higher end-to-end delay than the maximum allowable delay are discarded at the destination host. The main contribution to the variable network delay comes from the output port buffer occupancy of the bottleneck switch as shown in Figure 5.8.

Figure 5.8: The effect of switch buffer size on average buffer occupancy of the bottleneck switch: (1) without control, (2) with control.

In the controlled case, however, the mean queueing delay including the delay at the codec output buffer as well as the queueing delays at the intermediate switches does not change with the increasing buffer size. The average output port buffer occupancy of the bottleneck switch is kept constant explaining the bounded delay and cell loss rate.

The experiments show that increasing the switch buffer size would not improve the cell loss rate of the reference connection in the uncontrolled case. On the other hand, applying the dynamic rate control to the video sources would keep the average buffer occupancy low. Hence, the switch could use large buffers to accommodate momentary traffic increase and might lower the overall cell loss rate in the network. As a result, we can conclude that the composite control approach offers better use of network resources in severe congestion while keeping the cell loss rate and end-to-end

Figure 5.9: The effect of switch buffer size on average total queuing delay of reference video cells: (1) without control, (2) with control.

delay low compared to the uncontrolled case.

5.4 Efficiency of the Algorithm in Heavily Loaded Networks

The experiment in the previous section demonstrated that the average buffer occupancy has been kept low within the intermediate switches. This indicates that the network can accommodate more sources without violating their QoS requirements when our proposed scheme is used.

In order to explore the efficiency of the scheme we have conducted a group of experiments. The parameters for the data sources have been set to following values: burst length: 90 cells, burst interarrival time: 2.8 ms, and offered load: 30%. Figure 5.10 shows the number of video sources that the network can accommodate at different

Figure 5.10: Number of video sources that can be accommodated by the network given that no cell loss has occurred, versus the utilization of the output port buffer in the bottleneck switch. (1): with control. (2): without control. x-axis is used to show bottleneck utilization, although the controlled variable was the number video of sources, to provide better visualization of the efficiency of the scheme.

utilization values practically without any cell loss in the reference connection. When there is no control the network can accommodate only four sources with no cell loss. When the fifth source starts transmission the cell losses happen frequently. Since the QoS requirement for video cell loss rate is assumed to be zero, the curve for the uncontrolled case in Figure 5.10 is not plotted for the number of sources greater than four. On the other hand, the network can accommodate more than twice as many sources (up to nine) without violating the QoS requirements of the reference video source in the controlled case.

Our scheme achieves high efficiency in terms of resource sharing by controlling

Figure 5.11: Mean QIL vs. number of video sources. (1): without control. (2): with control.

the rate of the video sources. In order to determine whether this efficiency causes any perceptual quality degradation, we observed the mean QIL change with the increasing number of video sources, and plotted in Figure 5.11. The curve designated as "uncontrolled" shows the QIL values used in real MPEG compression scheme. MPEG scheme quantizes the I and P frames at QIL value 3, and the B frames at 6. According to the study given in [21] the change in the QIL is not perceptible as long as QIL value remains below 20. The difference of the mean QIL values between the controlled case and the uncontrolled case found to be small indicating that the change in the quality is insignificant. The cumulative distribution function of the QIL for the limiting case (network with 9 video sources), is also presented in Figure 5.12 for I-P frames and B frames, respectively. According to these plots, 96% of the I and P frames are quantized at QIL equal to 10 and less and 96% of the B frames are quantized at 14 and less.

In view of the above observations, we can conclude that the proposed scheme increases network efficiency, allowing more video sources for a given bandwidth, without causing any perceptual degradation in the video quality in heavily loaded networks.

5.5 Implications of the Algorithm on Queueing Delay Distribution

Figure 5.10 clearly shows that in heavily loaded networks the control scheme achieves high efficiency compared to the uncontrolled case. On the other hand, it is still important to show that the algorithm does not cause unnecessary delay when the network is not overloaded assuming that the network will work under normal conditions most

Figure 5.12: CDF of Quantization Index Level of (a) I-P frames, (b) B frames.

of the time. The total queueing delay with four video sources in the presence of background sources is observed.

The histograms and the CDFs for both controlled and uncontrolled cases are plotted in Figure 5.13. The shape of the histogram in the controlled case is smoother, and the peak values decay exponentially with the delay. The average delay is equal or less than 20 ms with 0.9 probability in the controlled case. For the uncontrolled case the same value can be guaranteed with only 0.7 probability.

The result of this experiment indicates the important role our algorithm can play in terms of the source behaviour. Using this algorithm the traffic becomes smoother before being transmitted to the network: consequently, a more manageable traffic is provided to the network resulting in smaller and less variable cell delays.

Figure 5.13: The histogram and the CDF of total queueing delay. Above: Uncontrolled Case. Below: Controlled Case

Figure 5.14: The effect of propagation delay on (a) mean cell loss rate and (b) mean quantization index level of the reference connection.

5.6 The Effect of Propagation Delay on Performance Metrics

One of the concerns about using feedback to control congestion in ATM networks is the large propagation delay relative to transmission speed. Propagation delay might cause a decrease in the efficiency of a reactive rate control algorithm because the time it takes to react the congestion increases with the propagation delay. A set of experiments was conducted in order to determine the impact of the propagation delay on the performance of the proposed scheme. Nine video sources and two background data sources with the same parameters are used as in the Section 5.1. The results of the experiment are plotted in Figures 5.14 - 5.15 where on the horizontal axis we depict the end-to-end propagation delay for the reference connection.

As illustrated in Figure 5.14, increasing propagation delay does not have an impact on the cell loss rate for a range of 45 ms, corresponding to a distance of 9000 km. The cell loss occurs when the propagation delay becomes the main component of the

total network delay, and causes a violation in the end-to-end delay requirement of the reference connection. In such a case, the cell loss does not occur due to the buffer contention but due to excess delay violating video QoS requirements.

The total queueing delay, illustrated in Figure 5.15, is affected by the increasing propagation delay. The fact that the mean QIL is unchanged throughout the experiment, implies that the source buffer occupancy is not affected by the propagation delay. Therefore, the only reason for the increase in total queueing delay is the queueing delay at the intermediate switches, especially of the one at the bottleneck switch. The mean bottleneck buffer occupancy, plotted in Figure 5.15, also proves this observation.

As a result of this experiment, we can conclude that the proposed algorithm is quite robust for connections with distances up to as large as 9000 *km*. Because of this our algorithm is applicable to a variety of networks ranging from local area to wide area networks.

5.7 The Effect of Background Source Traffic on the Performance Metrics

One of the main factors that plays an important role in the congestion control problem is the effect of the burstiness of the background traffic. In this section we investigate the effect of the mean burst length of the background data sources on the quality of the reference video connection. The simulation is run with the mean burst lengths of both background sources ranging from 45-360 cells. In order to keep the offered load of the data sources fixed at 30%, we adjust the mean burst arrival time accordingly. Table 5.1 shows the data source traffic parameters.

Figure 5.15: The effect of total propagation delay on (a) mean total queuing delay of the reference connection cells and (b) mean buffer occupancy of the bottleneck switch.

Mean Burst Length (cells)	Mean Burst Interarrival Time (ms)
45	1.11
90	2.83
180	5.66
240	7.5
300	9.4
360	11.3

Table 5.1: Data source traffic parameters

Figure 5.16: The effect of mean burst length on mean cell loss rate.

The effect of the increasing mean burst length on the behaviour of the control scheme can be seen through the change in the cell loss rate and the QIL of the reference video connection. The results are shown in Figures 5.16 and 5.17. In these experiments the number of video sources was kept constant at 9. The CRCS is used as a control scheme throughout these experiments. The results from the uncontrolled case are not shown since the network is unable to support 9 videos sources along with the background sources as discussed in section 5.4.

Increasing the mean burst length beyond 100 causes cell loss. On the other hand we observe only a slight increase in mean QIL which suggests that the scheme does not allow a great degradation to occur in the perceptual quality of the video connection. The difference between the QIL trace and the histograms with mean burst length 90 and 360 are shown in Figure 5.18. The effect of the increasing burst

Figure 5.17: The effect of mean burst length on mean QIL.

length on the video quality is represented with the higher density of the greater QIL values, indicating a slight but certainly not drastic reduction of video quality as the burstiness of the background sources increases.

Figure 5.18: The comparison of QIL traces and the histograms with the mean burst length: Above: 90 cells and Below: 360 cells.

Chapter 6

Conclusions and Recommendations for Future Research

6.1 Conclusions

This thesis has presented a performance study of MPEG coded VBR video transmission over ATM networks under a number of network loads varying from light to heavy. A dynamic rate control algorithm proposed for available bit rate data traffic has been modified for the real-time video traffic. The overall goal for applying this control methodology was to adapt the output rate of MPEG compressed video sources to varying network conditions in order to achieve an acceptable video quality and higher utilization of network resources.

We proposed an integrated approach for the rate control of the MPEG coded VBR video sources. This approach combines both the codec and the network by means of a feedback loop with a goal to alleviate congestion. The service rate of the codec's output buffer is adjusted periodically to regulate the source output. The experimental results have shown that this buffering allows the coder to adjust itself

to the network conditions acting as a damping mechanism for the control scheme.

The modification of the Balanced Mechanism's rate control algorithm, shown to be necessary to achieve the target QoS for the real time MPEG video, has been attained by changing the formulation and the flow of the rate control algorithm presented in [34]. After the modifications, our scheme has demonstrated an appropriate behaviour in order to be used as an effective rate control mechanism.

In the study that followed, we showed that the Composite Rate Control Scheme keeps the average buffer occupancy low, compared to the uncontrolled case. As a result, network switches and multiplexers can use large buffers in order to obtain low cell loss rate for the uncontrolled background sources without causing intolerable network delays for the real-time video traffic.

Our simulation results show that the network can accommodate more sources without violating the QoS requirements of real-time video connections in the controlled case. Instead of keeping the video quality constant, our scheme lowers the transmission rate with the increase of the network load. However, in general, the network resources are used more efficiently ensuring the best perceptual image quality for a given bandwidth.

Finally, we have tested the spatial scalability of our algorithm by changing the distance between the switches. The results show that the performance is not affected for distances as large as 9000 km. The increasing burstiness of the uncontrolled background sources also does not affect the visual quality of the video sources despite some cell loss. The implication is that, the composite rate control scheme is able to handle increasing traffic burstiness gracefully, resulting in a controlled reduction of video quality.

6.2 Recommendations for Future Research

In this study we use the quantization index in order to evaluate the quality of the video connection. Instead, a real MPEG sequence with a codec could be used in order to play out the sequence at the destination to evaluate the *perceptual* quality of the connection.

In addition to variable quantization rate control, the MPEG compression scheme offers another potential mechanism for providing minimal subjective affects as bandwidth is restricted. The MPEG-2 coding scheme has an ability to produce low and high priority streams. An ATM switch with a pushout buffer mechanism could utilize this mechanism by discarding low priority cells in the time of congestion. Incorporation of a priority control mechanism such as the one suggested in [5] into CRCS could increase the utilization of the network with a little sacrifice in the perceptual video quality.

Here we force zero cell loss for the controlled sources since we do not have a device for evaluating the perceptual video quality by playing out the real video sequences at the destination end. Such a device would provide the ability to set some cell loss probability to the video connection for an acceptable quality. Furthermore, an increase in the number of controlled sources could be achieved by incorporating an error concealment mechanism such as the one given in [39] and setting an appropriate cell loss probability for the video connection.

We have applied the CRCS to video sources only. Instead, the background sources could be controlled in the same manner which should increase the network performance further. The performance of the scheme could be tested in the presence of different background data source models such as self similar traffic.

References

- [1] International Telecommunication Union. "*CCITT Recommendation I.150. B-ISDN Asynchronous Transfer Mode Functional Characteristics*". Nov. 1993.
- [2] D. Heyman, A. Tabatabai and T. Lakshman. "Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM Networks". *IEEE Trans. Circ. Syst. Video Tech.*, 2(1):49-59, Mar. 1992.
- [3] S. Lam, S. Chow, and D.K.Y. Yau. "An Algorithm for Lossless Smoothing of MPEG Video". In *Proc. ACM Sigcomm '94*, pages 281-293, London, England, UK, 1994.
- [4] L. Dittmann and S.B. Jacobsen. "Statistical Multiplexing of Identical Bursty Sources in an ATM Network". In *Proc. Global Telecommun. Conf., Globecom '88*, pages 1293-1296, Hollywood, Florida, USA, 1988.
- [5] R. Ismail, I. Lambadaris, M. Devetsikiotis, and A. R. Kaye. "Modeling Prioritized MPEG Video Using TES and a Frame Spreading Strategy for Transmission in ATM Networks". In *Proc. INFOCOM'95*, pages 762-769, Boston, Massachusetts, April 1995.

- [6] G. C. Kessler. "*ISDN: Concepts, Facilities and Services*". McGraw Hill, Inc. 1993.
- [7] D. Minoli. "*Enterprise Networking*". Boston: Artech House, 1993.
- [8] R. Onvural. "*Asynchronous Transfer Mode Networks: Performance Issues*". Boston: Artech House, Inc. 1994.
- [9] A. Atai. "A Rate Based Feedback Traffic Controller for ATM Networks". In *Proc. IEEE Supercomm, ICC'94*, volume 3, pages 1605-1615. New Orleans, Louisiana, USA, 1994.
- [10] J. J. Bae and T. Suda. "Intelligent Congestion Control for ABR Service in ATM Networks". *Proc. of IEEE*, 79(2):170-189, Feb. 1991.
- [11] L. Orozco-Barbosa and N. Hirzalla. "A Rate Control Mechanism for LAN/MAN Interworking Systems". In *High Performance Networking, HPN '94*, pages 357-371. Grenoble, France, 1994.
- [12] V. Bemmell and M. Ilyas. "A Unified Congestion Control Strategy in ATM Networks". In *Proc. IEEE Supercomm, ICC'94*, volume 3, pages 1600-1601. New Orleans, Louisiana, USA, 1994.
- [13] S.J. Golestani. "Congestion-Free Communication in High Speed Packet Networks". *IEEE Trans. Commun.*, 39(12):1802-1812, Dec. 1992.
- [14] K. Y. Siu and H. Y. Tzeng. "Intelligent Congestion Control for ABR Service in ATM Networks". *Computer Communication Review, ACM SIGCOMM*, 24(5):81-106, Oct. 1994.

- [15] F. Bonomi and K. W. Fendick. "The Rate Based Flow Control for the Available Bit Rate ATM Service". In *IEEE Network Magazine*, pages 25-39. Mar./Apr. 1995.
- [16] N. Yin and M. Hluchyj. "On Closed Loop Congestion Control Scheme for Real-Time Packet Video Transfer". In *Proc. IEEE Conf. Comp. Commun., Infocom '94*, pages 99-108. Toronto, Ontario, Canada. 1994.
- [17] B.A. Makrucki. "Explicit Forward Congestion Notification in ATM Networks". In *Proceedings of TriComm '92: High Speed Communications Network*, pages 73-96. Raleigh, North Carolina, USA. 1992.
- [18] V. Jacobson. "Congestion Avoidance and Control". In *Computer Communication Review: Proc. ACM Sigcomm Conf.*, pages 314-329. 1988.
- [19] Q. Li and Z. Wang. "A Comparison Study of Congestion Control for ABR Service in ATM Networks". In *Proc. IFIP Conference on High Speed Networking*, pages 7/1-7/10. Bradford, UK. 1995.
- [20] C. M. Sharon, I. Lambadaris, M. Devetsikiotis, and A. R. Kaye. "Modeling and Control of VBR H.261 Video Transmission over Frame Relay Networks". Accepted for publication in the *IEEE Transactions on Circuits and Systems for Video Technology*.
- [21] H. Kanakia, P. Mishra, and A. Reibman. "An Adaptive Congestion Control Scheme for Real-Time Packet Video Transport". In *Proc. ACM Sigcomm '93*, pages 20-31. Ithaca, NY. 1993.

- [22] S. Karademir, I. Lambadaris, M. Devetsikiotis, and A.R. Kaye. "Dynamic Rate Control of VBR MPEG Video Transmission over ATM Networks". In *Proc IEEE Global Telecommun. Conf., Globecom '96*, London, UK, 1996.
- [23] D. E. McDysan and D. L. Spohn. "ATM Theory and Application". McGraw Hill, Inc. 1994. pages 243-246.
- [24] International Telecommunication Union, Geneva. "CCITT Recommendation I.1321. B-ISDN Protocol Reference Model and Its Applications", 1991.
- [25] R. Haendel and M. Huber. "Integrated Broadband Networks". New York: Addison-Wesley, 1991.
- [26] R. Ismail. "Statistical Modeling of VBR-MPEG Video and Its Transmission in ATM Networks". Master's thesis, Carleton University, Ottawa, Canada, 1995.
- [27] International Telecommunication Union, Geneva. "CCITT Recommendation I.1363. B-ISDN ATM Adaptation Layer (AAL) Specification", 1991.
- [28] International Telecommunication Union, Geneva. "CCITT Recommendation I.610. OAM Principles of B-ISDN Access", 1991.
- [29] ATM Forum Technical Committee. "ATM User-Network Interface Specification, Version 3.0". New Jersey: PTR Prentice Hall, 1993.
- [30] ATM Forum Technical Committee. "Traffic Management Specification, Version 4.0". Jun. 1995.
- [31] D. Le Gall. "MPEG: A Video Compression Standard for Multimedia Applications". *Commun. ACM*, 34(4):47-58, Apr. 1991.

- [32] International Standards Organization, ISO/MPEG. "*Standard ISO1172. Coding of Moving Pictures and Associated Audio*", Nov. 1991.
- [33] M. Nomura, T. Fujii, and N. Ohta. "Basic Characteristics of Variable Rate Video Coding for ATM and Broadcast Applications". *IEEE Jour. Sel. Areas in Commun.*, 7(5):752-760, Jun. 1989.
- [34] Y. Gong and I. Akyildiz. "Dynamic Traffic Control Using Feedback and Traffic Prediction in ATM Networks". In *Proc. IEEE Conf. Comp. Commun., Infocom '94*, pages 75-84, Florence, Italy, May 1992.
- [35] N. Yin and M. Hluchyj. "Models for Statistically Multiplexed Data Traffic in Cell Relay Networks". In *Proc. IEEE Global Telecommun. Conf., Globecom '93*, pages 824-829, Houston, TX, 1993.
- [36] B. Melamed, D. Raychaudhuri, B. Sengupta, and J. Zdepski. "TES-Based Modeling For Performance Evaluation of Integrated Networks". In *Proc. IEEE Conf. Comp. Commun., Infocom '92*, pages 75-84, Florence, Italy, May 1992.
- [37] M. Kawashima, C.T. Chen, and S. Singhal. "Adaptation of the MPEG Video Coding Algorithm to Network Applications". *IEEE Trans. Circ. Syst. Video Tech.*, 3(4), Aug. 1993.
- [38] J. Banks and J. S. Carson. "*Discrete-Event System Simulation*". Prentice-Hall, Inc, 1984. pages 407-442.
- [39] T. Han and L. Orozco-Barbosa. "Performance Requirements for the Transport of MPEG Video Streams over ATM Networks". In *IEEE International Conference on Communications, ICC '95*, pages 221-225, Seattle, Washington, USA, 1995.

END

1 | 8 - 0 | 6 - 9 | 7

FIN