PM-1 3 1/2"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

PRECISIONSM RESOLUTION TARGETS

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services Branch

Direction des acquisitions et
des services bibliographiques

395 Wellington Street
Ottawa, Ontario
K1A 0N4

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file - Votre référence

Our file - Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

Compressed Video in Integrated Services Frame Relay Networks

by

C. M. Sharon, BSc.
Royal Roads Military College

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
December 15, 1994
© copyright
1994, C.M. Sharon

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

Your file Votre référence

Our file Notre 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 ACCORDE UNE LICENCE
IRREVOCABLE ET NON EXCLUSIVE
PERMETTANT A LA BIBLIOTHEQUE
NATIONALE DU CANADA DE
REPRODUIRE, PRETER, DISTRIBUER
OU VENDRE DES COPIES DE SA
THESE DE QUELQUE MANIERE ET
SOUS QUELQUE FORME QUE CE SOIT
POUR METTRE DES EXEMPLAIRES DE
CETTE THESE A LA DISPOSITION DES
PERSONNE INTERESSEES.

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 PROPRIETE
DU DROIT D'AUTEUR QUI PROTEGE
SA THESE. NI LA THESE NI DES
EXTRAITS SUBSTANTIELS DE CELLE-
CI NE DOIVENT ETRE IMPRIMES OU
AUTREMENT REPRODUITS SANS SON
AUTORISATION.

ISBN 0-315-98607-7

Canada

Name E. Michael Sharon

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.

Compressed Video in Integrated Services
Frame Relay Networks

SUBJECT TERM

0544

SUBJECT CODE

U-M-I

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 0413
 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 0-27
 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
 Biostatistics 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 0368
 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 0609
 Nuclear 0610
 Optics 0752
 Radiation 0756
 Solid State 0611
 Statistics 0463

Mathematics

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 0609
 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**

Compressed Video in Integrated Services Frame Relay Networks

submitted by C.M. Sharon, BSc.

**in partial fulfilment of the requirements for
the degree of Master of Engineering**

Thesis Supervisors

Chair, Department of Systems and

Computer Engineering

Carleton University

December 13, 1994

Abstract

This thesis proposes a simple and effective methodology for modeling H.261 video traffic using Transform Expand Sample techniques. It is shown that “frame level” source models do not adequately capture the size variability (burstiness) of sub-frame sized data packets. New approaches are developed to capture the bit rate variations which occur *within* each video frame, and to simulate the traffic output from a variable bit rate H.261 codec.

A frame relay congestion control strategy is proposed, such that H.261 rate controls respond to explicit network congestion notifications. Factors which significantly influence the performance of this mechanism are identified, using factorial analysis, and optimized using mean field annealing. Backward explicit congestion notification is found to regulate the output of H.261 video sources effectively in a mixed traffic (video and inter-LAN data) frame relay network.

The relative performance of variable quantization rate control and variable frame rate control is analyzed. Variable quantization rate control is found to offer significant performance advantages over variable frame rate source control.

Acknowledgements

I would like to thank Dr. M. Devetsikiotis and my thesis supervisors, Prof. A.R. Kaye and Prof I. Lambadaris. Their helpful suggestions, endless patience and quiet encouragement has been deeply appreciated.

I gratefully acknowledge the funding support provided by the Telecommunications Research Institute of Ontario and the National Science and Engineering Research Council of Canada. I wish to also thank the Dedicated Technology Corporation, Ottawa, Canada, for providing many of the video sequences used in this research.

I would like to thank my friend and mentor, Mr. P.C. Engstad, for affording me the opportunity to complete this thesis. Finally, I would like to thank my wife Janet, for her constant support and many valuable suggestions.

Table of Contents

Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	ix
List of Figures	xi
List of Acronyms	xii
List of Symbols	xiv
1 Introduction	1
1.1 Background	1
1.2 Previous Research	3
1.3 Thesis Objective	4
1.4 Thesis Organization	5
2 H.261, Frame Relay and Transform Expand Sample	7
2.1 Digital Video	7
2.1.1 Introduction	7
2.1.2 Image Compression	8

2.2	Recommendation H.261	12
2.2.1	H.261 Video Format	12
2.2.2	H.261 Compression	16
2.3	Frame Relay Bearer Service	19
2.3.1	Frame Format	20
2.3.2	Congestion Control	21
2.4	Transform Expand Sample (TES)	23
2.4.1	TES Background Sequences	24
2.4.2	Stitching Function	25
2.4.3	TES Foreground Sequences	27
2.4.4	Summary	28
3	Source Characteristics and Modeling	29
3.1	Empirical Video Data	30
3.1.1	Packetization of Video Data	30
3.1.2	Peak-to-Mean Ratio	35
3.1.3	Probability Distribution	37
3.1.4	Autocorrelation	39
3.2	Video Source Model	42
3.2.1	The GOB Ratio Model	43
3.2.2	Modeling Variable Quantizer Rate Control	45
3.3	InterLAN Data Model	49
3.4	Summary	50

4	Network Analysis	52
4.1	Network Simulation	52
4.1.1	Network Topology	52
4.1.2	Traffic Models	54
4.1.3	Network Delays	57
4.1.4	Reference Connection and Source Rate Control	58
4.1.5	Packet Loss and Error Recovery	59
4.2	Simulation Methodology	60
4.2.1	Transient Time	60
4.2.2	Number of Runs	62
4.2.3	Cost Function	63
4.3	Simulation Results	65
4.3.1	Experimental Design	65
4.3.2	Minimum Data Rate	68
4.3.3	BECN Delay	71
4.3.4	Design Optimization	73
4.3.5	Comparison of Rate Control Schemes	79
4.4	Comparison of Empirical Data, TES GOB and TES Frame Models .	82
4.5	Summary	85
5	Conclusions and Recommendations for Further Study	87
5.1	Conclusions	87
5.1.1	Video Source Models	87
5.1.2	Congestion Control	89

5.2	Recommendations for Further Study	91
5.2.1	Rate Control	91
5.2.2	Discard Eligibility	92
5.2.3	Video Adaptation Layer	93
5.2.4	Improved Traffic Model and Congestion Control	93
5.2.5	Adaptive Buildout Delay	94
	References	95

List of Tables

2.1	H.261 image formats	13
3.1	GOB sizes of empirical VBR video sequences in H.261 NTSC CIF format	32
3.2	GOB sizes of empirical VBR video sequences in H.261 QCIF format .	34
3.3	Peak to mean ratios, H.261 VBR video	36
3.4	Peak to mean ratios: comparison of H.261 CBR and VBR images . .	36
4.1	Generic 3 factor design analysis	65
4.2	Factors and levels selected for design analysis	67
4.3	Parameters and bounds used in MFA optimization	75
4.4	Near optimal values for network parameters determined by MFA . . .	76
4.5	Ranges of near-optimal values for network parameters	77
4.6	Queue length statistics at NODE2 TQ1 for three control conditions .	80

List of Figures

2.1	H.261 GOB grouping	14
2.2	H.261 data stream	15
2.3	H.261 coding algorithm	16
2.4	H.261 intraframe coding process	18
2.5	Frame relay frame format	21
2.6	Probability Density of Innovation Sequence V_n	25
2.7	Migration of correlated background sequence on unit circle	26
2.8	Probability Density of Innovation Sequence V_m	26
2.9	TES stitching transformation S_ξ	27
3.1	Bike sequence: CBR frame data, mean frame size of 4615 octets . . .	38
3.2	Bike sequence: VBR (Q=4) frame data, mean frame size of 4615 octets	38
3.3	Distribution of frame size for two different quantization tables	39
3.4	GOB autocorrelation for NTSC video (10 GOBs/frame), Q=8 and Q=12	41
3.5	Frame autocorrelations for NTSC video coded at Q=8 and Q=12 . .	42
3.6	GOB autocorrelation for QCIF sequence "M3": real and modeled . .	44
3.7	Distribution of GOB size for QCIF sequence "M3": real and modeled	45
3.8	Average interframe size (normalized at Q=10) vs quantization	46

3.9	GOB correlation for VBR sequence “Tennis” in QCIF format (Q=12):	
	comparison of real and modeled data	48
3.10	Distribution of empirical and modeled QCIF video	49
3.11	Histogram of ethernet packet size (total packets = 4089)	50
4.1	Frame relay network simulation	53
4.2	Average packet delay vs simulation run time	61
4.3	Confidence intervals (95%) for primary and secondary effects	68
4.4	Floor (F) vs cost function with 95% confidence intervals	70
4.5	Autocorrelation of buffer occupancy as a function of time	72
4.6	Cost as a function of BECN delay	73
4.7	Congestion threshold θ_1 vs cost function using MFA	77
4.8	Damping (ζ) vs cost function using MFA	78
4.9	Cost function vs link3 bandwidth: with and without BECN rate control	78
4.10	Queue length vs time - no rate control	80
4.11	Queue length vs time - variable frame rate control	81
4.12	Queue length vs time - variable quantization rate control	82
4.13	Maximum end-to-end delay: TES models vs empirical data	84

List of Acronyms

AC	Alternating Current
ATM	Asynchronous Transfer Mode
BECN	Backward Explicit Congestion Notification
BISDN	Broadband Integrated Services Digital Network
CBR	Constant Bit Rate
CCD	Charge Coupled Device
CCITT	Consultative Committee for International Telegraphy and Telephony
CIF	Common Intermediate Format
CLLM	Consolidated Link Layer Management
CODEC	Coder/Decoder
C-Plane	Control Plane
CPU	Central Processing Unit
C/R	Command/Response
DC	Direct Current
DE	Discard Eligibility
DLCI	Data Link Connection Identifier
DPCM	Differential Pulse Code Modulation
FAD	Frame Assembler Dissassembler
FECN	Forward Explicit Congestion Notification
FPS	Frames per Second
FR	Frame Relay
FT1	Fractional T1 Supporting Nx64 Kbps connectivity ($N \leq 24$)
GOB	Group of Blocks
HDLC	High-Level Data Link Control
HVS	Human Visual System
IDN	Integrated Digital Network
IEEE	Institute of Electrical and Electronic Engineers
IP	Internet Protocol
ISDN	Integrated Services Digital Network
LAN	Local Area Network
LAPD	Link Access Protocol D
MB	Macroblock
MFA	Mean Field Annealing
MMP	Markov Modulated Process
MPEG	Motion Picture Experts Group
NTSC	National Television System Committee

OSI	Open Systems Interconnection
PAL	Phase Alternating Line
PBX	Private Branch Exchange
PDU	Protocol Data Unit
PMR	Peak to Mean Ratio
PVC	Permanent Virtual Circuit
QCIF	Quarter Common Intermediate Format
RBOC	Regional Based Carriers
RQ	Receive Queue
SECAM	Sequential Couleura Memoire
SNA	Systems Network Architecture
T1(DS1)	Point-to-point dedicated service supporting 1.544 Mbps aggregate connectivity or twenty-four 64 Kbps subchannels (also DS0)
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing
TES	Transform Expand Sample
TQ	Transmit Queue
U-Plane	User Plane
VBR	Variable Bit Rate
WAN	Wide Area Network
YCbCr	24 bit digital conversion standard (CCIR 601) for YIQ/YUV colour samples
YIQ	NTSC video sampling format
YUV	PAL/SECAM video sampling format

List of Symbols

B_C	Maximum Committed Burst Rate
B_E	Maximum Excess Burst Rate
C	Cost Function
D	BECN Delay
E_1	Empirical Video Data
F_1	TES Frame Model
G_*	TES GOB Model
ζ	Damping Constant
L	Percentage Packet Loss
P	Actual Throughput as Percentage of R_2
R_1	Minimum source data rate
R_2	Preferred source data rate
RQ	Receive Queue
θ_1	Congestion Threshold
θ_2	Absolute Congestion Threshold
t_0	Simulation start time
t_1	Transient time
T_{FAD}	FR Frame Assembly Disassembly Delay
T_{PROP}	Propagation Delay
T_{SW}	FR Switching Delay
T_{TRANS}	Transmission Delay
TQ	Transmit Queue
$U_n^\pm$	TES Background Process
U	Node 2 TQ1 Server Utilization
$X_n^\pm$	TES Foreground Process
$\rho_X(k)$	Autocorrelation of the k^{th} lag

Chapter 1

Introduction

1.1 Background

Voice, data and broadcast video information has historically been transmitted over dedicated and distinctly separate communication networks. During the past two decades, however, a phenomenal increase in demand has occurred for high speed data communications and, more recently, for video and multimedia services. Consumer demand, and the technical means to provide increasingly sophisticated and diverse services, has revolutionized the communications industry and given birth to the concept of a global “information highway”. When implemented, the information highway will consolidate and enhance the functions of existing networks to provide economical voice, data and video communication services over an integrated digital network.

The heart of the information highway will comprise a web of interconnected Broadband Integrated Services Digital Networks (BISDN). The Asynchronous Transfer Mode (ATM) protocol will be used to transport data over optical trunks at several hundreds of megabits per second. Local and regional networks will operate at lower speeds on the periphery of the BISDNs and will use BISDN services for interworking over longer distances. To support efficient utilization of available bandwidth in the

presence of "bursty" traffic, BISDNs will not allocate fixed channel widths to individual connections. Rather, multiple logical connections will be statistically multiplexed over the same physical link so that each connection has access to the full available bandwidth "on demand" [1]. In the case of narrower band ISDNs, fixed channel widths are allocated to individual connections but statistical multiplexing still occurs at higher levels.

Frame Relay (FR) is a variable length packet bearer service based on the ISDN Q.922 standard [2] and currently specified for access line speeds up to 2 Mb/s. Although designed for ISDN applications, the advantages of FR can be realized on both ISDN and non-ISDN networks. FR is protocol transparent above layer 2 of the Open Systems Interconnection (OSI) model and supports switching of X.25, SNA, TCP/IP or any other HDLC-like traffic with equal ease [3]. Due, in part, to its inherent flexibility and modest cost, FR is expected to play a significant role in future corporate networks by providing Local Area Network (LAN) interconnections between dispersed business sites [4].

There are currently over 30 FR service providers in Canada and the United States, including all seven US regional based carriers (RBOCs), and FR services are now available on European ISDN networks [5]. Industry standards for FR internetworking with ATM services are well advanced [6, 7, 8, 9] and FR technology is therefore well postured to become a major force on the periphery of the global information highway.

Market trends indicate a growing government and corporate demand for low bit rate motion video services to support conventional applications, such as teleconferencing, and recent innovations such as remote sensing and surveillance. Economical

implementation of enhanced communication services frequently hinges on the successful adaptation of existing network infrastructures, such as FR, to new applications.

1.2 Previous Research

The FR protocol was not designed to support delay sensitive traffic such as voice and video. Provision for a cumulative delay stamp mechanism, which could be used by the network to monitor the progress of time sensitive traffic, requires OSI level 3 processing at intermediate nodes, as in the packet voice protocol [10] and runs contrary to the fundamental precepts of FR design [11]. However, previous research [12, 13, 14, 15] has shown that appropriate mechanisms can be implemented to achieve high quality transmission of compressed voice over FR networks. In essence, variable transit delays are absorbed at the destination in a receive buffer, where voice traffic is held for a short period prior to commencing the replay. Additional research [16, 17] has established methodologies for controlling FR network congestion in the presence of mixed voice and data traffic. Equivalent FR applications involving video information have not been adequately investigated.

CCITT recommendation H.261 [18] provides guidelines for the compression and transmission of constant bit rate (CBR) video information at integer multiples of 64 kb/s. Considerable research [19, 20] has been conducted on propagating H.261 images over ATM and TCP/IP networks. To take full advantage of the "bandwidth on demand" characteristics of ISDN and BISDN networks, the requirement for CBR video could be relaxed to allow the video bit rate to vary naturally in relation to scene activity. The normal process of limiting video bit rates during peak periods could therefore be abandoned in favour of a more constant image quality. The feasibility of

adapting CBR H.261 coders to variable bit rate (VBR) applications (constant image quality) has been satisfactorily demonstrated in several studies [21, 22].

Traditional approaches to video source modeling, such as Markov modulated processes (MMP) and autoregressive techniques, are computationally extensive and are typically developed to reflect the characteristics of a particular video sequence. Classical approaches of this sort are primarily designed to accurately fit the empirical autocorrelation function but they cannot generally fit the empirical marginal distribution. Transform Expand Sample (TES) represents a fresh approach to video modeling which can accurately match both the marginal distribution and the empirical autocorrelation function of a wide variety of video sources [23]. Computationally simple [24] and inherently flexible, TES is capable of modeling a wide range of video traffic conditions with extreme accuracy. Having been automated [25] to the extent that TES software can operate interactively with a network traffic model, TES based simulations can be structured to produce generalized results of greater relevance to actual network environments.

1.3 Thesis Objective

The aim of this thesis is to:

- i Develop and validate a simple approach to generating high fidelity, TES-based models, of H.261 video sequences.
- ii Develop and validate a mixed traffic (data/video) source model for use in this and future studies.

- iii **Develop a Frame Relay network simulation to validate the TES video models by comparing the relative end-to-end performance of simulated and empirical video source data.**
- iv **Explore, through simulation, the feasibility of using network congestion notification messages to initiate video source rate reduction measures.**
- v **Identify network and source parameters affecting the efficiency of the proposed network congestion control strategy and optimize these parameters for improved performance.**

1.4 Thesis Organization

The remaining chapters of this thesis are organized as follows:

Chapter 2: Reviews the fundamentals of digital video compression, CCITT recommendation H.261, Transform Expand Sample (TES) modeling and the Frame Relay Bearer Service.

Chapter 3: Investigates the characteristics of empirical H.261 VBR video sequences and identifies the underlying size characteristics of frame and sub-frame sized video data packets. Video and LAN traffic modeling techniques are developed and shown to capture accurately the bit rate characteristics of empirical data. Various options for the parsing and packetization of H.261 video information are explored and a preferred approach is identified with supporting rationale.

Chapter 4: The structure of the FR network simulation is described. The feasibility of controlling FR network congestion by introducing network feedback

as an input to video source rate control is explored. Network and source parameters which significantly affect the performance of the proposed congestion control mechanism are identified and optimized. A comparison is performed of two source rate control schemes: variable quantization and variable frame rate. Finally, the influence of medium fidelity and high fidelity video source models on end-to-end network performance is analyzed.

Chapter 5: Conclusions and Recommendations for Further Study.

Chapter 2

H.261, Frame Relay and Transform Expand Sample

2.1 Digital Video

2.1.1 Introduction

The formation of a digital image first requires that the analog source information be sampled at discrete points in space. To achieve this, an optical lens is used to focus the source image onto an array of light sensitive material. The values returned by individual array elements are referred to as “picture elements” (or pixels) and normally have a linear relationship to the radiant energy at each sample location. For motion video, temporal variations are recorded by sequentially capturing fixed images at discrete points in time. Monochrome images can be produced by sampling each pixel location in a single spectral band. Colour images require multispectral sampling at frequencies corresponding to the three subtractive primaries: red, yellow and blue. In this latter case, the luminance (brightness) information is typically isolated from the chrominance, or hue. The resulting sample data is then expressed in YCbCr digital format, where the Y, Cb and Cr bands represent luminance, yellow-

red difference and yellow-green difference values respectively [26]. Under CCIR 601, each of the three (YCbCr) pixel sample components are represented as unsigned 8 bit values.

2.1.2 Image Compression

Moving pictures comprise sequences of still images which are viewed sequentially at the specified frame rate. For each of the YCbCr components, a high degree of correlation exists between the values for neighbouring pixels within the same still image (spatial correlation) and between corresponding pixels on sequential still images (temporal correlation). Video compression is achieved by decorrelating the video data, removing less important information and efficiently encoding the remaining output. The entire compression process therefore comprises three distinct phases:

- i **Reversible Compression:** The video signal is decorrelated by applying one or more common techniques, thus reducing the amount of redundant information that is transmitted. Compression techniques which exploit the spatial correlations that exist within a particular still image are described as *intraframe* techniques. Methodologies that also exploit the temporal correlations that exist between successive still images in a video scene are described as *interframe* techniques. As the name implies, operations in this class are *lossless*, or reversible, since only redundant information is deleted.
- ii **Irreversible Compression:** Irreversible processes reduce the amount of transmitted data by judiciously removing video information which is less readily perceived by the HVS. All processes in this class, such as quantization, are necessarily *lossy* because the data, once discarded, cannot be recovered.

- iii **Data Encoding:** Data encoding techniques, such as run length coding and Huffman coding, perform various operations on the quantized data stream to convey the video information as concisely as possible.

The following is a summary of common image compression methodologies.

Predictive Coding: DPCM and Motion Compensation

Differential Pulse Code Modulation (DPCM) is a predictive coding algorithm which is frequently used to decorrelate interframe and intraframe information. The DPCM method involves forming a linear prediction of a particular pixel value from the observed values of previous pixels. The “true” value can then be expressed in terms of the prediction error, which is normally very small for a typical video sequence. The variance and autocorrelation of the error signal is generally reduced by at least an order of magnitude over that of the true image [27]. The vast majority of error values, can subsequently be accurately represented by relatively few discrete cosine transform (DCT) coefficients, which is a quality that lends itself to improved coding efficiency.

Interframe DPCM coding techniques predict the value of a particular pixel using the known values of the corresponding pixel in previous still images. However, due to image movement or camera motion, the location of corresponding pixels in subsequent frames will tend to migrate across the display. Motion compensation is a technique which seeks to improve the DPCM prediction by producing an estimate of the direction and magnitude of image motion between successive still frames. The size of the error image produced by the DPCM algorithm can be substantially reduced when pixel value predictions are compensated for image motion.

Transform Coding: Discrete Cosine Transform

Transform techniques are commonly applied to intraframe (DPCM) error images to reduce the spatial correlations that occur within any particular still frame. This is accomplished by dividing each image into several smaller $n \times n$ pixel blocks and transforming the domain space of each block individually. The result of a successful transform is a collection of values, with low variance, which can be described by a relatively small set of transform coefficients.

The two-dimensional (2D) discrete cosine transform (DCT) can be applied to a small block of pixels to isolate the spatial frequency components of the pixel values. For an 8×8 pixel block, the resulting output from the 2D DCT is a set of 64 transform coefficients, organized into an 8×8 matrix, which can be interpreted as the amplitudes of the 64 spatial frequency components of the original pixel block. The DC component represents the average luminance of all pixels in the original block and is located in the top left corner of the matrix. The 63 AC coefficients are arranged so that lower spatial frequencies are closer to the DC coefficient and higher spatial frequencies are farther away.

Quantization

Prior to encoding, the floating point spatial frequency coefficients produced by the transform process are approximated (truncated) to the nearest of some finite number of quantization levels. Eight bit precision yields 256 uniform steps whereas 4 bit precision yields 16 uniform steps. Judicious selection of the quantization step size, which is normally specified individually for each of the 64 DCT coefficients, provides

a mechanism for the controlled loss of unnecessary information. For example, because the human eye is far less sensitive to high spatial frequency than low spatial frequency, the high frequency DCT coefficients can be coarsely quantized with no perceivable degradation to the reconstructed image quality. Due to this, and the fact that most images are dominated by relatively slow changes in spatial characteristics, many of the high frequency transform coefficients reduce to zero after quantization. The quantization process typically results in long runs of zeros when the quantized transform coefficients are subsequently arranged in ascending order of frequency. Organization of the coefficients in this way lends itself to further compression through run length coding.

Data Encoding: Run Length and Huffman Coding

Run Length encoding refers to the process of expressing long runs of zeros or ones as a single value. For example, 25 consecutive zeros, which might otherwise be transmitted as 25 individual data bits, can be condensed into a single, five bit, binary number (11001). In this simple example, an incremental compression ratio of 5:1 has been realized.

Variable Length Coding (VLC) is a general term used to describe encoding techniques which achieve lossless compression by exploiting the statistical properties of digital data. The number of bits required to transmit a fixed amount of information is reduced by assigning the shortest code (binary zero or one) to the data value with the highest probability of occurring. Huffman coding defines one of several methods for assigning variable length codes to quantized sample values.

2.2 Recommendation H.261

CCITT recommendation H.261 is a motion picture coding specification primarily intended for low bit rate video applications [18]. The recommendation establishes a common approach to implementing constant bit-rate (CBR) video services on ISDN networks at integer multiples ($P = 1, 2, 3 \dots 30$) of 64 kbps. North American ISDN rates are constrained to 1.544 Mbps with the value of “P” ranging from 1-24. Several commercial implementations of recommendation H.261 have been developed and are generically referred to as “P64” video codecs.

2.2.1 H.261 Video Format

Recommendation H.261 specifies two video formats for Phase Alternating Line (PAL) style inputs, the Common Intermediate Format (CIF) and the Quarter Common Intermediate Format (QCIF). For National Television System Committee (NTSC) style input, many P64 coders allow a modified CIF format which is sometimes referred to as MPEG SIF 525 active sampled stream. NTSC-style QCIF formats are generally not supported because there is no complete “Group of Block” (GOB) structure for the bottom portion of the image. Three spectral components (Y,Cb,Cr) are defined and conversion from the PAL or NTSC video sampling formats to YCbCr is accomplished in accordance with CCIR Recommendation 601 Specifications. The pixel dimensions associated with each P64 video format are shown in Table 2.1. The number of effective pixels (in parenthesis) is truncated to the nearest multiple of 8 to conform to an 8×8 block coding scheme.

Because the human visual system is more sensitive to variations in light intensity (luminance) than color (chrominance), the resolution requirements for chromi-

COMPONENT	PAL CIF		PAL QCIF		NTSC CIF (SIF)	
	LINES/FRAME	PIXELS/LINE	LINES/FRAME	PIXELS/LINE	LINES/FRAME	PIXELS/LINE
Y	288	360 (352)	144	180 (176)	240	352
Cb and Cr (combined)	144	180 (176)	72	90 (88)	120	176

Table 2.1: H.261 image formats

nance and luminance information are not the same. Recommendation H.261 specifies a 4:1:1 sampling ratio of luminance (Y) to chrominance (Cb and Cr) information. If a source pixel element represents each band (YCbCr) with 8 bits (24-bit colour), only the luminance information will be retained for every pixel. Information for each of the two chrominances (Cb and Cr) will be retained only once for every four pixel elements. The average information per pixel is therefore reduced from 24 bits/pixel to 12 bits/pixel to achieve an immediate compression ratio of 2:1.

Image Partitioning

Recommendation H.261 defines a 4 layer image partitioning hierarchy. The 8×8 pixel “block” is the smallest partition and represents the pixel area upon which discrete cosine transforms are performed. A “Macroblock” (MB) comprises four blocks (16×16 pixels) and represents the pixel area upon which motion compensation operations are performed. A “Group of Blocks” (GOB) comprises 33 Macroblocks arranged in 3 rows of 11 Macroblocks each. The pixel dimensions of a GOB are 176×48 . A QCIF frame comprises 3 GOBs organized in a single column. A PAL CIF frame comprises

Figure 2.1: H.261 GOB grouping

12 GOBs arranged in two side by side columns of 6. An NTSC CIF frame comprises 10 GOBs arranged in two side by side columns of 5. The GOB structure and number scheme for the three H.261 video formats is illustrated in Figure 2.1

Data Format

The four level image partitioning hierarchy is reflected in the H.261 data format. As shown in Figure 2.2, each level of information is progressively encapsulated by the next higher level. The picture header, the GOB header and the end of block (EOB) field are fixed length and are always transmitted. All other fields are variable length coded and are transmitted only when necessary. Within a particular GOB, for example, Macroblock information is only included for those MBs in which significant change has occurred. Similarly, the MB quantization table is only specified when it differs from the quantization table currently in use.

Figure 2.2: H.261 data stream

Rate Control

The H.261 Codec is designed for CBR operation at T1 or fractional T1 (FT1) rates which, in North America, may comprise 1 to 24 time division multiplexed (TDM) 64 kbps channels. Encoded video information is deposited into an output buffer which absorbs momentary bit rate fluctuations and which is emptied at a constant rate. The output buffer contents are depleted at the standard TDM sample rate of 8 KHz for each 64 kbps channel. Each TDM slot has sufficient capacity for 1 octet of data.

As shown in Figure 2.3, H.261 bit rate control is maintained by a feedback loop from the output buffer to the quantizer. The quantization step size(s) are adjusted to regulate the flow of information into the output buffer and to maintain a time averaged constant bit rate. During periods of peak scene activity, the escalating data rate is controlled by increasing the quantization step size, thus averting a buffer overflow. The coder must also guard against buffer *underflows*. To maintain synchronization

Figure 2.3: H.261 coding algorithm

with the decoder, sufficient video information must be available in the output buffer to fill all assigned TDM slots. If, during periods of very low scene activity, the output buffer occupancy becomes dangerously low, the P64 coder will increase the data rate by including stuffing bits in the macroblock information or fill bits in the error correction sequence, or both.

Recommendation H.261 also defines the maximum number of bits that may be used to represent a single frame (still image) of video information. The maximum data rate for CIF and QCIF formats is 256 kbit/frame and 64 kbit/frame respectively. These values do not include error correction bits, fill bits or parity bits.

2.2.2 H.261 Compression

CCITT Recommendation H.261 specifies both intraframe and interframe compression techniques. The coding method decision is made at the macroblock level on the basis of bandwidth economy, with the result that most H.261 frames contain a mixture of intraframe and interframe information. An additional criterion is that each MB must be intraframe coded at least once every 132 transmitted frames. This

constraint is necessary to limit the gradual image degradation that can occur through the accumulation of difference errors.

Intraframe Coding

Intraframe coding does not reference previous frames and yields lower average image compression than interframe coding. The intraframe coding process begins with a DCT acting sequentially on the four 8×8 pixel blocks comprising a macroblock. Thirty-one quantization tables ($Q = 1, 2, \dots, 31$) are specified in H.261 and are assigned on an MB or GOB basis to maintain the desired image resolution and bit rate. The quantized transform coefficients from each pixel block are then subjected to a “zig-zag” reordering scheme which produces a 1×64 vector. The trailing coefficients, which correspond to high spatial frequency components, typically evaluate to zero and are compressed using run length coding. Only the 63 AC components are subjected to run length encoding.

The first value of each 1×64 quantization vector represents the DC component of the pixel block. After quantization, the DC component is treated differently from the AC components. The high correlation between sequential DC components is exploited by applying a 1st order DPCM to the DC components only. Essentially, the value of the previous DC component is used to predict the value of the current DC component. The resulting string of values, DPCM encoded DC and run length encoded AC, are then Huffman coded. The complete intraframe coding process is illustrated in Figure 2.4.

Figure 2.4: H.261 intraframe coding process

Interframe Coding

Interframe coding begins by generating an error image for a particular macroblock, which represents the difference between the true pixel values and a 1st order DPCM prediction of the pixel values. The DPCM prediction may be improved by a motion compensation mechanism which is optional in H.261. The motion estimation algorithm compares the luminance data for the macroblock throughout a small search area of the previously transmitted image. The range of comparison is ± 15 pixels, or one MB. The displacement with the smallest absolute difference is considered the motion estimation vector for that macroblock, where the chrominance motion vector is the luminance motion vector divided in half. The DPCM prediction is then shifted to account for the estimated image motion. The resulting macroblock difference image is subsequently processed as an intraframe. The transformation, quantization and encoding process is identical to that previously described for intraframes.

2.3 Frame Relay Bearer Service

Frame relay is a connection oriented, data link control protocol designed to support the switching of bursty traffic through dynamic bandwidth allocation. In what is essentially a streamlined version of X.25 link access protocol D (LAPD), FR takes advantage of improved transmission quality and intelligent end user devices to minimize the amount of processing conducted by the network. Streamlining is achieved by stripping away the X.25 network layer (OSI layer 3), adding a statistical multiplexing capability to the data link layer (OSI layer 2) and by migrating flow control and error correction procedures to the network edges. In FR, only the following core functions are offered by the network [28]:

- frame delimiting, alignment and transparency provided by the use of HDLC flags and zero bit insertion/extraction;
- virtual circuit multiplexing/demultiplexing using the address field of the frame;
- checking the frame to ensure that it consists of an integer number of octets prior to zero bit insertion or following zero bit extraction;
- checking the frame to ensure that it is not too long or too short;
- detection of transmission errors and discarding of errored frames;
- congestion control functions.

FR was developed as an ISDN packet bearer service and includes provisions for a logically separate user plane (U-plane) and control plane (C-plane). The C-plane is intended to support management of switched virtual circuits (SVCs) using

out-of-band signaling on a separate channel. However, currently evolving frame relay services do not support either in-band or out-of-band call set up procedures [29]. Existing FR applications are therefore limited to permanent virtual circuits (PVCs) which are pre-defined by the network manager. Multiple PVCs may be statistically multiplexed on the same (unchannelized) physical connection. Packets associated with a particular data stream (PVC) are uniquely identified by a data link control identifier (DLCI) which is contained in the packet header and inspected by the network switches. Associated with each PVC are “elastic bandwidths” which establish the average bit rate, maximum committed burst rate (B_C) and maximum excess burst rate (B_E) that the network will attempt to support. Within these bounds, each PVC has access to the full capacity of the physical connection “on-demand”.

2.3.1 Frame Format

An FR frame is assembled by the end user equipment and is interpreted by the network switches. The FR frame format is shown in Figure 2.5. Each frame begins and ends with the one octet HDLC flag sequence (01111110). The address field has a default length of two octets but may be extended to 3 or 4 octets to support longer address fields. Each octet of the address field is delimited by an extended address (EA) bit. The EA bit is set to 1 in the final octet and is set to zero in all previous octets. The Data Link Control Identifier (DLCI) has a default length of 10 bits and is used to identify the PVC with which a frame is associated. DLCIs are assigned locally (on a hop by hop basis) and are inspected by the network switches to determine which output port the received packet should be routed to. The command/response (CR) bit is used to implement end-to-end command response protocols and has no bearing

Figure 2.5: Frame relay frame format

on network functions. The FECN, BECN and DE bits have congestion notification and control functions which are now discussed in more detail.

2.3.2 Congestion Control

Congestion, which affects the throughput, delay and loss characteristics of end-to-end traffic, occurs when one or more of the following network resources become depleted:

- i Bandwidth, or frame stream congestion, occurs when the available bandwidth along a particular physical path is inadequate to support demand. Frame stream congestion is detected by monitoring the occupancy of every transmit queue at each network node (switch).
- ii Memory, or nodal congestion, occurs when the total instantaneous traffic buffered at a particular node threatens to exceed the global memory capacity of the switch. Nodal congestion is detected by monitoring the global memory occupancy level at each switch.
- iii Processor CPU congestion typically occurs when network traffic is dominated by short packets and the packet arrival rate exceeds the switching capacity

of nodal CPUs. This type of congestion is detected by monitoring the CPU utilization at each network switch.

Congestion control procedures in FR networks comprise congestion avoidance actions, which are implemented immediately following detection, and congestion recovery actions which are initiated if avoidance measures prove inadequate.

Congestion Avoidance

Congestion notification is accomplished, upon detection of a mild congestion condition, by setting the BECN or FECN bits in the header of FR packets passing through the congested network node. BECN bits are set by the congested node on packets traveling in the opposite direction to traffic experiencing congestion. A BECN notification is intended to stimulate rate reduction procedures by the transmitting unit(s) at the traffic source(s). BECN notifications may also be used to trigger the setting of DE bits by the originating unit(s) on selected packets. A set DE bit indicates that the FR packet contains user data of lesser importance to the end user and should be discarded in favour of other packets if necessary to alleviate congestion.

FECN bits are set by the congested node on packets experiencing congestion enroute to their destination. The FECN bit is intended to notify a receiving end user that arriving traffic has encountered network congestion. This information can be used to regulate the output of destination controlled transmitters.

The consolidated link layer management (CLLM) message represents an alternative to the BECN approach. CLLM messages can be generated by the network and travel over the U-plane to source destinations contributing to network congestion.

The generation and transport of CLLM messages by the network is optional [11] and CLLM procedures are not currently implemented in many network devices.

Congestion Recovery

Congestion recovery is initiated by FR networks if the source response to explicit congestion notifications (FECN, BECN, CLLM) proves inadequate to prevent the development of severe congestion. A severe congestion condition, which is detected in the same manner as mild congestion, triggers congestion recovery procedures within the network. When severe congestion is detected, the network discards FR packets until buffer occupancies and/or CPU utilization return to acceptable levels. Limited protection is afforded to FR packets which are *not* marked as eligible for discard (DE bit not set) but no packets enjoy unconditional exemption from intentional discard when congestion recovery is implemented. High packet loss rates due to congestion recovery are detected by end-to-end protocols (OSI level 3 and above) and provide an implicit notification to end users of severe network congestion. Implicit congestion notification provides end users with additional indications of the requirement to implement effective rate control measures at the traffic source(s).

2.4 Transform Expand Sample (TES)

TES processes constitute a class of stationary stochastic processes that can accurately model the marginal probability distribution and the autocorrelation function of a wide variety of empirical video data. A TES model comprises two stochastic processes: a background sequence U_n , which influences the autocorrelation characteristics of the

resulting TES model, and a foreground sequence X_n which independently influences the distribution of the resulting TES model.

2.4.1 TES Background Sequences

TES background sequences are denoted as either $\{U_n^+\}_{n=0}^\infty$ or $\{U_n^-\}_{n=0}^\infty$ where:

$$U_n^+ = \begin{cases} U_0 & n = 0 \\ \langle U_{n-1}^+ + V_n \rangle & n > 0 \end{cases} \quad U_n^- = \begin{cases} U_n^+ & n \text{ even} \\ 1 - U_n^+ & n \text{ odd} \end{cases}$$

The brackets $\langle \rangle$, represent the fractional part (modulo-1) operator, such that for real X , $\langle X \rangle = X - [X]$, where $[X] = \max \{n \text{ integer: } n \leq X\}$. The sequence U_0 is distributed uniformly on $[0, 1)$ and $V = \{V_n\}_{n=1}^\infty$ is a sequence of iid random variables, independent of U_0 , called the *innovation sequence*. The superscript notation $U_n^\pm$, reflects the fact that $\{U_n^+\}$ and $\{U_n^-\}$ generate lag-1 autocorrelations in the range $[0, 1]$ and $[-1, 0]$ respectively.

A key property of TES background sequences is that they are stationary Markovian and uniformly distributed on $[0, 1)$, regardless of the probability distribution of the selected innovations $\{V_n\}$ [30]. However, the autocorrelation of a TES background sequence is influenced by the innovation density function f_V , which can be varied to achieve a wide variety of autocorrelation characteristics.

The modulo-1 arithmetic applied to TES background processes allows them to be interpreted geometrically as random walks on a circle of circumference 1. For example, the innovation density function illustrated in Figure 2.6 can be viewed as causing the background sequence U_n^+ to migrate slowly and consistently around the circle in a clockwise direction (Figure 2.7). A background sequence of this type will

Figure 2.6: Probability Density of Innovation Sequence V_n

exhibit strong positive autocorrelations over several lags. As well, a periodic component can be introduced into the background sequence autocorrelation by shifting the density function in Figure 2.6 away from the y-axis.

Figure 2.8 illustrates a different innovation density function, such that the value of V_n varies uniformly on $[-0.5, 0.5]$. In this case, the corresponding background sequence moves arbitrarily on the circle and is therefore uncorrelated.

2.4.2 Stitching Function

Referring once again to Figure 2.7, it is evident that a discontinuity will occur in the sample path of a TES background sequence whenever the sequence traverses the 0/1 boundary on the circle. This discontinuity can be alleviated by applying a linear *smoothing* (or stitching) transformation to the background process, as illustrated in Figure 2.9. The stitching parameter ξ , assumes values from 0 to 1, with the greatest

Figure 2.7: Migration of correlated background sequence on unit circle

Figure 2.8: Probability Density of Innovation Sequence V_m

Figure 2.9: TES stitching transformation S_ξ

smoothing occurring in the range $0.4 < \xi < 0.6$. It is useful to note that because linear transformations preserve uniformity, all stitched background sequences, $S_\xi(U_n)$, are also uniformly distributed on $[0, 1)$.

2.4.3 TES Foreground Sequences

A TES foreground sequence $X_n^\pm$ is generated by performing an inverse transform on the stitched background sequence $S_\xi(U_n^\pm)$, such that:

$$X_n^+ = \hat{H}_Y^{-1}[S_\xi(U_n^+)], \quad X_n^- = \hat{H}_Y^{-1}[S_\xi(U_n^-)]$$

where $\hat{H}_Y$ typically represents the distribution (histogram) of the empirical data being modeled and S_ξ is the stitching function.

The *inversion method* [31] makes possible the transformation of any uniform variate (ie. any stitched background sequence) to another sequence with arbitrary

distribution $\hat{H}_Y$. Also, the inverse transform $\hat{H}_Y^{-1}$ does not affect the autocorrelation characteristics of the stitched background sequence $S_\xi(U_n^\pm)$. The autocorrelation characteristics of the background sequence are therefore preserved in the foreground sequence $X_n^\pm$, which also has a marginal distribution identical to the empirical data.

To express the two stage transformation (stitching and inverse transform) of background sequences to foreground sequences more succinctly, we define a composite transformation of the form:

$$D_{Y,\xi}(U_n^\pm) = \hat{H}_Y^{-1}[S_\xi(U_n^\pm)], \quad U_n^\pm \in [0, 1)$$

where $D_{Y,\xi}$ is known as the *distortion*, the inner transformation S_ξ is the stitching function and the outer transformation $\hat{H}_Y^{-1}$ is the inverse of an empirical distribution (histogram). We may then rewrite the relationship between background and foreground sequences as:

$$X_n^+ = D_{Y,\xi}(U_n^+), \quad X_n^- = D_{Y,\xi}(U_n^-)$$

2.4.4 Summary

The TES modeling methodology has the attractive property of *decoupling* the distribution of the foreground process from the autocorrelation function. Because the distribution of $X_n^\pm$ is guaranteed to match the empirical distribution F_Y , the model need only be optimized to fit the empirical autocorrelation function. Fitting the TES model to the empirical autocorrelation function is accomplished by varying the innovation density, f_v and the stitching parameter, ξ . The fitting process has been automated in the TEStool software utility [30], which employs optimal search methods to match the autocorrelation characteristics of TES models to empirical statistics.

Chapter 3

Source Characteristics and Modeling

The primary purpose of this chapter is to develop H.261 video source models, using the TEstool software utility, which faithfully represent the size characteristics of empirical data. To identify how H.261 video information should be parsed for encapsulation into FR packets, an analysis of the H.261 data structure is first conducted. The video frame and GOB size characteristics of H.261 empirical data are then investigated, to determine if the size statistics of FR packets containing GOB sized PDUs can be approximated by source models which only capture the video *frame size* statistics. A TES based modeling technique is developed and shown to capture the GOB characteristics of empirical VBR traffic accurately when the rate feedback loop in the H.261 codec is disabled and the same quantization table is applied to every macroblock. The model is then expanded to consider an H.261 video codec, which generates VBR traffic but reacts to network congestion by applying coarser quantization values to reduce the mean data rate.

The secondary purpose of this chapter is to introduce a TES based LAN traffic model, which accurately reproduces the first order size statistics (histogram)

of ethernet packets. A slow decaying correlation function is imposed on the size characteristics of sequentially arriving ethernet packets, to capture the inherently *bursty* nature of LAN traffic.

3.1 Empirical Video Data

A public domain H.261 video codec was used to analyze several, constant bit rate (CBR) H.261 video sequences acquired over the Internet. The CBR video sequences were decoded into uncompressed binary files and recoded as variable bit rate (VBR) video sequences. VBR encoding was accomplished by disabling the rate control feedback loop in the H.261 software codec and specifying a fixed quantization table for the entire video sequence.

Additional VBR video sequences were produced by a proprietary H.261 hardware codec supplied by the Dedicated Technologies Corporation (DTC), Ottawa, Canada. The DTC approach to generating H.261 VBR video is identical to that previously described for the public domain software codec. The compatibility of images generated by the two different codecs was demonstrated by using the software codec to replay video sequences produced by the hardware codec. Motion compensation, which is an optional H.261 feature, was enabled at all times in both the software and the hardware codec.

3.1.1 Packetization of Video Data

Parsing of H.261 binary data into variable length PDUs, suitable for FR encapsulation and transport, is most logically accomplished by partitioning the H.261 bit stream at the video frame or GOB boundaries. Partitioning of the video bit stream

into smaller segments, along H.261 macroblock (MB) boundaries, is not considered feasible as it causes increased network overhead without improving resilience to cell loss. This latter condition is because macroblock address is difference coded, relative to the address of the last transmitted macroblock, for all macroblocks within a particular group of blocks. In essence, loss of any one macroblock will render useless all subsequent information pertaining to the same group of blocks. Conversely, frame and GOB level data is self contained and can be decoded without reference to preceding information.

The maximum H.261 frame sizes for NTSC/PAL CIF and QCIF are 256 kbits and 64 kbits respectively [18], from which an approximate maximum GOB size of 46.1 kbits (5760 octets) can be inferred. This is done by calculating the average distribution of empirical GOB sizes as a percentage of the total frame size, and observing that a particular NTSC CIF GOB rarely comprises more than 18% of the total frame size. NTSC CIF sequences were chosen for this analysis because the largest GOBs are expected to occur in NTSC format, which has the same maximum frame size as PAL CIF but contains two fewer GOBs. As shown in table 3.1, empirical GOB size data collected for two high activity H.261 VBR sequences, coded in NTSC CIF format, fall well within the theoretical maximum value of 5760 octets. It is also apparent that an inverse relationship exists between the quantization table number, Q , hereafter referred to as the quantization number, and the GOB size. Even for sequences in which significant motion occurs, it is evident that maximum GOB sizes are likely to occur only at high resolutions, where the quantization number is very small ($Q < 5$). Therefore, a maximum GOB size estimate of 5760 octets, which assumes high resolution sequences in NTSC format, is overly pessimistic for low bit

SEQUENCE	FORMAT	QUANT LEVEL	MEAN GOB (OCTETS)	MAX GOB (OCTETS)	THEORETICAL MAX (OCTETS)
TENNIS	NTSC	10	539	3914	5760
TENNIS	NTSC	4	1357	3944	5760
BIKE	NTSC	10	242	1517	5760
BIKE	NTSC	4	461	2636	5760

Table 3.1: GOB sizes of empirical VBR video sequences in H.261 NTSC CIF format rate FR video applications, such as video telephony and surveillance.

To determine a typical range of quantization numbers for *realistic* FR video applications, a subjective analysis of several H.261 VBR sequences was conducted. It was found that for quantization numbers greater than 15 ($Q > 15$), the quantization is sufficiently coarse to produce visible artifacts in the reconstructed image. Conversely, sequences encoded at lower quantization numbers ($Q \leq 15$) were found to yield reconstructed images of acceptable quality, when viewed with the naked eye. When quantization numbers of less than 5 were used ($Q \leq 5$), the mean frame and GOB size was observed to escalate rapidly with no perceivable improvement to image quality. H.261 quantization numbers in the range ($5 \leq Q \leq 15$) are therefore considered adequate to produce video sequences with sufficient resolution for typical FR video applications.

Regarding video formats, it is evident that video applications running on FR networks will be sufficiently constrained in bandwidth to preclude the transmission of broadcast quality sequences. For example, sequence "Tennis" encoded at ($Q = 10$) and transmitted at 30 frames per second, produces a mean bit rate of 1.29 Mbps and a peak rate of 9.39 Mbps, which exceeds the maximum FR bandwidth by a considerable margin. In general, if the CIF or NTSC CIF format is preserved, very significant

decreases in image resolution (quantization) and frame rate are required to achieve the necessary reduction in data rate, resulting in a blurred and unsteady (jerky) video sequence. A more useful approach for most applications is to adopt the QCIF format, which allows better image quality to be maintained over a smaller area. It is therefore useful to consider the characteristics of *realistic* FR video sequences in more detail, which are now defined as H.261 sequences in QCIF format with quantization numbers in the range ($5 \leq Q \leq 15$).

A survey of four empirical video sequences was conducted to determine the GOB size statistics of VBR H.261 information, encoded in QCIF format. Sequences M1, M3 and M5 each comprise approximately 400 QCIF frames and were produced with the DTC codec. A fourth sequence ("QHEAD") was produced with the software codec and contains 150 frames of video data. The following is a brief summary of relevant scene characteristics:

M1 - is a talking head sequence which features a consistently high degree of body and hand motion as the speaker holds various objects up to the camera for viewing.

M3 - is a medium activity scene, similar to M1 but with less relative motion.

M5 - is a very *bursty* scene caused by a camera which is intermittently panned across an office work area. Scene activity is extremely high when the camera is panned but is low when the camera is at rest.

(QHEAD) - is a very benign talking head sequence of an official reading a press release. Body motion and background activity are minimal.

SEQUENCE	FORMAT	QUANT LEVEL	MEAN GOB (OCTETS)	MAX GOB (OCTETS)	THEORETICAL MAX (OCTETS)
M1	QCIF	5	786	1635	4000
M3	QCIF	5	535	1477	4000
M5	QCIF	5	407	1925	4000
QHEAD	QCIF	10	154	1403	4000

Table 3.2: GOB sizes of empirical VBR video sequences in H.261 QCIF format

It was found that the maximum GOB size of any particular video frame, rarely comprises more than 50% of the video frame size, from which a maximum GOB size of 32 Kbits, or 4000 octets, can be inferred. The maximum empirical GOB sizes were determined for each empirical sequence, which were encoded with quantization values in the *typical* range, ($5 \leq Q \leq 15$). As shown in table 3.2, the maximum empirical GOB sizes, which are consistent with previous results from similar studies [24, 20], are significantly smaller than the maximum inferred GOB size of 4000 octets. Once again, it is evident that the maximum inferred GOB size is a rather pessimistic estimate of the maximum GOB size associated with *realistic* FR video sequences. In practice, the approximate maximum frame and GOB sizes associated with *realistic* FR video sequences are 6000 octets and 2000 octets respectively.

GOB sized PDUs associated with *realistic* FR video applications have an approximate mean and maximum size of 500 octets and 2000 octets respectively. Frame relay packets containing GOB sized PDUs therefore comply with industry specifications for FR networks [32] and can generally migrate through high speed WANs and LANs without the need for further partitioning. For example, the maximum PDU size specified for FDDI LANs is 4500 octets and the maximum PDU size for an SMDS WAN is 9188 octets. Conversely, FR packets containing frame sized PDUs

are too long for practical applications and exceed the maximum specified frame size for many networks. For these reasons, partitioning of H.261 video information along GOB boundaries is considered the most viable parsing scheme for the transmission of H.261 video information over FR networks. A clear implication of this result is that FR network simulations designed to study video traffic phenomena may require a video source model which faithfully represents the GOB statistics. It is prudent to determine first if significant differences exist between the frame and GOB size characteristics of H.261 empirical data.

3.1.2 Peak-to-Mean Ratio

The peak-to-mean ratio (PMR), which serves as a first order measure of “burstiness”, was calculated for several VBR video sequences at the frame and GOB level. The results, which appear in Table 3.3, indicate a frame PMR of 2:1 to 4:1 which is consistent with previous studies [32, 19]. At both the frame and GOB level, the PMR is observed to be proportional to scene activity. Accordingly, the least bursty image is the relatively static “talking head” and the most bursty image is the multi-scene sequence. Of greater interest is the difference that exists between the frame PMR and the GOB PMR for every image. In all cases, the GOB size data is observed to be significantly more variable than the associated frame size data. This result is attributable to the statistical smoothing that occurs when several GOBs are combined into a single, frame-sized unit. The implication, however, is that frame level modeling of video source data will generally underestimate the actual burstiness of GOB sized packets. GOB level modeling of the source is therefore necessary to capture accurately the true size variability of network video packets carrying GOB sized PDUs.

IMAGE	FRAME MEAN	FRAME MAX	RATIO	GOB MEAN	GOB MAX	RATIO
MULTI SCENE	5797	22579	3.89	461	3999	8.67
TENNIS 10	5391	16843	3.12	539	3914	7.26
BIKE 4	4615	13013	2.82	461	2636	5.72
HEAD M3	1604	3402	2.12	535	1477	2.76

Table 3.3: Peak to mean ratios, H.261 VBR video

IMAGE	MEAN	MAX	STD DEV	RATIO
VBR BIKE FR	4615	13013	2006	2.82
CBR BIKE FR	4615	7952	1069	1.72
VBR BIKE GOB	461	2636	429	5.72
CBR BIKE GOB	469	2104	395	4.49

Table 3.4: Peak to mean ratios: comparison of H.261 CBR and VBR images

Variable quantization rate control is used by the H.261 coder to maintain a nominally constant bit rate (CBR) binary output. The effect of variable quantization rate control on the burstiness of a source image is illustrated by comparing the VBR (fixed quantization) and CBR (variable quantization) output for the same video sequence. Referring to Table 3.4 the CBR bit stream is observed to be less bursty than the VBR bit stream at both the frame and GOB level. The reduction in burstiness is, however, far more pronounced at the frame level and is a subtle reminder that CBR video is only constant when averaged over time intervals of approximately 1 second using the smoothing buffer. To understand this limitation, it is useful to recall that the H.261 coder achieves a CBR bit stream by varying the quantization step size at either the MB or GOB level. Because of this, a nominally constant bit rate can only be achieved over significantly longer time intervals. The implication is that, even when implemented at the source, rate control is only effective in controlling bursts of reasonably long duration. As illustrated in Figures 3.1 and 3.2, short duration bursts of the order of a frame interval exist in both CBR and VBR data streams. Rate control schemes which are triggered by network feedback mechanisms, such as the Frame Relay BECN notification, will certainly be even less effective in controlling short duration bit rate fluctuations.

3.1.3 Probability Distribution

Probability distributions of frame size and GOB size were developed for several VBR video sequences. The distributions were observed to be highly image dependent with more “bursty” video sequences typically being characterized by a bimodal, or multimodal distribution. As illustrated in figure 3.3, the shape of the distribution for a

Figure 3.1: Bike sequence: CBR frame data, mean frame size of 4615 octets

Figure 3.2: Bike sequence: VBR (Q=4) frame data, mean frame size of 4615 octets

Figure 3.3: Distribution of frame size for two different quantization tables

particular video sequence is preserved when the same image is encoded with a different quantization table. The effect, on the distribution, of changing the quantization step size is equivalent to multiplying the size of each frame or each GOB by a scalar value.

3.1.4 Autocorrelation

The autocovariance of a *stationary* random sequence, $X(n)$ is defined as:

$$C_X(k) = E[X(n)X(n+k)] - E[X(n)]E[X(n+k)] \quad (3.1)$$

for $(n = 1, 2, \dots)$ and $(k = 0, 1, \dots)$.

When considering the binary output from a video coder, $X(n)$ comprises a finite sequence $(n = 1, 2, \dots, N+k)$ of real numbers which represent the size (in

octets) of sequentially transmitted GOBs or frames. The autocovariance of such a sequence can be estimated as:

$$\hat{C}_X(k) = \frac{1}{N} \sum_{n=1}^N X(n)X(n+k) - \left(\frac{1}{N} \sum_{n=1}^N X(n) \right) \left(\frac{1}{N} \sum_{n=k}^{N+k} X(n) \right) \quad (3.2)$$

where $k \ll N$ and $\hat{C}_X(k)$ converges to $C_X(k)$ as $N \rightarrow \infty$.

The k th-lag autocorrelation coefficient, which is a measure of the similarity between sequence values that are separated by k lags [34], is defined as the normalized autocovariance:

$$\rho_X(k) = \frac{\hat{C}_X(k)}{(\sigma_X)^2} \quad (3.3)$$

where σ_X is the standard deviation of $X(n)$, ($n = 1, 2, \dots, N+k$). It is useful to note that because the autocorrelation coefficient is normalized, it is not affected by scalar multiplication.

Frame size information was tabulated for several different VBR video samples as a sequence of real numbers and the autocorrelation for each sequence was computed as a function of the lag measured in number of frames. Similar computations were performed on the GOB size information for several different empirical video samples.

At the frame level, strong positive correlations are observed to occur and are attributed to the similarities (temporal redundancy) that exist between still images recorded on sequential video frames. At the GOB level, pronounced but distinctly

Figure 3.4: GOB autocorrelation for NTSC video (10 GOBs/frame), $Q=8$ and $Q=12$

different, correlations are evident with periodicities at both the spatial GOB scan rate and the temporal frame rate. Figures 3.4 and 3.5 illustrate the frame and GOB correlation characteristics for a VBR video sequence (“Tennis”) in NTSC CIF format. The sequence is of two people playing table tennis and comprises 150 frames (1500 GOBs). In figure 3.4, the complex GOB correlation characteristics are clearly visible and each of the 10 GOBs per frame are easily distinguished. However, in figure 3.5 it is evident that spatial (GOB) correlations are lost in the aggregate frame statistics and cannot, therefore, be captured by a frame level source model.

Figures 3.4 and 3.5 also provide a comparison of the autocorrelation for the same video sequence encoded with two different quantization numbers. The frame and GOB autocorrelation was observed to remain constant as the quantization number is varied. It is therefore apparent that a change in the H.261 quantization step size (quantization number) induces a linear shift in the distribution function and has no

Figure 3.5: Frame autocorrelations for NTSC video coded at $Q=8$ and $Q=12$

effect on the autocorrelation of the frame size or GOB size. As scalar multiplication influences the first and second order statistics in precisely the same manner, it is possible to infer frame and GOB size statistics for a wide range of H.261 quantization numbers from a single, fixed quantization number source model.

3.2 Video Source Model

The marginal distribution and autocorrelation properties of empirical bit rate/frame data (frame-level data) can be accurately and easily modeled with the TESTool utility. However, the slow decay properties and periodic behaviour of typical GOB-level autocorrelations can not be directly modeled using TES or autoregressive techniques [24]. Because GOB size statistics are required in many network simulations to determine the size of individual PDUs, a crude estimate of GOB size is commonly given by dividing the frame size by the number of GOBs per frame. Such “frame level”

source models fail to capture the variability of GOB size within each frame and, as illustrated in table 3.3, underestimate the actual “burstiness” of the GOB bit-rate (and FDU size) data.

Accurate TES modeling of H.261 GOB statistics was demonstrated in [24] by removing the deterministic periodic components from the GOB bit-rate data through periodogram analysis [35]. The residual sequence is then modeled using TEStool, and the periodic components are subsequently reintroduced to produce the final source model. However, because periodogram analysis is time consuming and computationally extensive, a more streamlined approach to GOB-level source modeling is frequently useful. An alternative approach to GOB-level source modeling was therefore developed and is presented in the following section.

3.2.1 The GOB Ratio Model

If we consider a QCIF format video sequence, which contains three GOBs per frame, the measured average size of each GOB can be easily determined as a percentage of the frame size. Similar computations for PAL CIF format (12 GOBs per frame) and NTSC CIF format (10 GOBs per frame) are equally trivial. We refer to the corresponding vector of fractions as a GOB ratio vector, G_R , which contains either 3 (QCIF), 10 (NTSC CIF) or 12 (PAL CIF) elements. Thus obtained, the GOB ratio vector is applied to a frame-level source model to provide an estimate of the empirical GOB bit rate characteristics. Figures 3.6 and 3.7 illustrate the result of this very simple approach in capturing the GOB-level statistics of a the QCIF “talking head” sequence “M3”, for which $G_R = [0.380, 0.195, 0.425]$. The autocorrelation of the simulated data is consistently higher than that of the empirical data, due

Figure 3.6: GOB autocorrelation for QCIF sequence “M3”: real and modeled

to the slightly reduced GOB size variability induced by averaging the GOB-to-frame percentages across the entire video sequence. This “averaging error”, which introduces modeling inaccuracies in the GOB-level autocorrelation and distribution function, becomes more noticeable as scene activity increases. Averaging error is, however, easily controlled by computing a different GOB ratio vector, $[G_R(1), G_R(2), \text{etc}]$, for different scene segments. The frame-level source model is then sequentially modulated with the different $G_R(*)$ to produce the final, aggregate, bit stream. The percentage of time allocated to each GOB ratio vector is determined by the distribution of the different scene segments in the original sequence. This modeling approach, which is subsequently referred to as “GOB ratio modeling”, was successfully applied to several empirical video samples. Experience reveals that three GOB ratio vectors are typically sufficient to produce good agreement between the empirical and reconstructed GOB-level statistics.

Figure 3.7: Distribution of GOB size for QCIF sequence “M3”: real and modeled

3.2.2 Modeling Variable Quantizer Rate Control

VBR video output is typically produced in an H.261 codec by disabling the rate control feedback loop from the codec output buffer to the quantizer. The original CBR codec is designed to maintain a constant output bit rate by varying the quantization number applied to individual macroblocks and selecting it in response to rate information received via the output buffer feedback loop. However, VBR video can be produced if the output buffer feedback loop is disabled and the normal feedback information is replaced with a constant signal. Thus modified, a constant quantization number is maintained.

Despite the many advantages of implementing VBR network traffic, control of the mean output of a VBR source may be necessary to avoid network congestion or to honour peak bandwidth allocations negotiated for a particular PVC or SVC. Such control, which could be triggered by feedback received from the network, can be

Figure 3.8: Average interframe size (normalized at $Q=10$) vs quantization

achieved by retaining some logic to allow the video codec to select the quantization number most appropriate for current conditions. To explore these issues, which is the goal of the succeeding chapter, a variable quantizer modeling capability is required.

Figure 3.8 illustrates the relationship between bit rate/frame and quantization number (1-31) for two different video sequences. For ease of comparison, the data is normalized to a frame size ratio of 1 for $Q=10$. The small differences observed between the two profiles are attributed to the following two factors:

- i Each of the H.261 quantization numbers (1-31) specify a quantization step size for each of the 64 transform coefficients that result from the DCT of an 8×8 pixel block. Within a particular quantization table (number) the quantization step size is not necessarily uniform for all 64 transform coefficients. Coarser step sizes are typically assigned to the higher numbered transform coefficients to take advantage of the relative insensitivity of the HVS to high spatial frequencies.

Progressing through the H.261 quantization numbers from $Q=1$ to $Q=31$, the proportional change in quantization step sizes is larger for high numbered transform coefficients than for low numbered transform coefficients. Accordingly, a video sequence containing an above average amount of high spatial frequency information will suffer a larger than average reduction in aggregate bit rate as the quantization table number increases.

- ii An H.261 frame will typically comprise a mixture of intracoded and intercoded macroblocks. The decision to employ absolute or difference coding for a particular macroblock is made “on the fly” on the basis of bandwidth economy. For example, when a particular region of an image experiences a large change, coding of the large difference values often requires more binary information than does absolute coding of the pixel values. The first (DC) coefficient of each intrablock is always quantized with a fixed step size of 8 and no dead zone [18], regardless of the quantization number in use. The relative mix of intracoded and inter-coded macroblocks, which is scene dependent, will therefore have a small but significant effect on the bandwidth reduction achieved as progressively larger quantization numbers are specified.

The effect of image unique characteristics on the relationship between normalized bandwidth and the specified H.261 quantization number is small, but not insignificant. Nevertheless, for a particular video sequence or group of sequences, the mean bandwidth associated with a particular quantization number can be predicted with considerable accuracy. The implication is that a GOB level model need only be developed for one of the 31 possible H.261 quantization numbers.

Figure 3.9: GOB correlation for VBR sequence “Tennis” in QCIF format ($Q=12$): comparison of real and modeled data

Specifying a different quantization number to the H.261 codec is simulated in the GOB source model by applying a scalar multiple to each value generated. Recall from our previous analysis of empirical data that changes to the quantization number affect the distribution and autocorrelation function in precisely the same way as scalar multiplication. In both cases, the shape of the distribution function is preserved, but shifted, and the autocorrelation function is not affected. To demonstrate the utility of this approach, a GOB Ratio Model, denoted $G_R(8) = [(0.304, 0.304, 0.392), (0.257, 0.234, 0.509), (0.432, 0.104, 0.464)]$, was developed from the empirical VBR sequence “Tennis”, encoded in QCIF format at $Q=8$. Because the “Tennis” sequence is a high activity scene, it was necessary to split the sequence into three equal scene segments and define individual ratio vectors for each segment. A second empirical file was generated by recoding the same sequence at $Q=12$. A model of the $Q=12$ sequence was then produced by applying a scaling factor to the model,

Figure 3.10: Distribution of empirical and modeled QCIF video

$G_R(8)$. As illustrated in figures 3.9 and 3.10, good agreement is achieved between the marginal distribution and autocorrelation functions of the empirical and modeled data.

3.3 InterLAN Data Model

A histogram of Ethernet packet sizes was compiled from empirical data retrieved from a 10 Mbps research network comprising 70-80 engineering work stations. The empirical histogram of packet size, illustrated in Figure 3.11, indicates that the network traffic is dominated by packets with the minimum length (68 octets) or maximum length (1522 octets) permitted by the IEEE 802.3 standard [36]. The empirical histogram was reconstructed with the TESTool utility and a TES model was developed to match the empirical distribution (histogram). A slow decaying exponential autocorrelation function was imposed on the data model to enhance the probability of

Figure 3.11: Histogram of ethernet packet size (total packets = 4089)

producing long bursts of large packets, a condition which is particularly conducive to network congestion.

3.4 Summary

An analysis of the H.261 data structure reveals that H.261 frame and GOB level data is self contained and can be decoded without reference to preceding information. However, because frame sized blocks of data are too large for practical network applications, FR video PDUs should be constructed by parsing H.261 video information along GOB boundaries. The approximate average and maximum GOB sizes for H.261 video in QCIF format, which is the most practical format for FR video applications, are 500 octets and 2000 octets respectively. Video PDUs of this size can be transmitted through high speed LANs, WANs and FR networks without further partitioning.

To simulate network traffic caused by the transmission of H.261 video information accurately, it is necessary to model the GOB size statistics of H.261 data. We find that frame level models of video source data do not capture the variability of GOB size within each frame and consistently underestimate the actual “burstiness” of the video data stream.

The slow decay properties and periodic behaviour of typical GOB level autocorrelations can not be directly modeled using TES techniques. However, frame level TES models can be modulated with an empirically derived GOB ratio vector, which describes the average distribution of GOB size within a video frame. In this manner, the GOB size distribution and autocorrelation characteristics of empirical video sequences can be accurately captured. The GOB ratio modeling approach is computationally trivial and can be easily and quickly applied to empirical data. As well, the accuracy of the model is proportional to its complexity and can be varied to meet specific requirements.

The normalized data rates associated with a particular quantization table (number) are found to remain relatively constant for a wide range of scene types. Due to this property, a single video traffic model can be developed for a fixed quantization number and subsequently used to model traffic characteristics at different quantization numbers. To achieve this, a scaling factor is applied to the original model, which influences the distribution and autocorrelation statistics in a manner analogous to changing the quantization number. A simple and accurate method is therefore available to simulate the traffic output from a VBR codec, which varies the quantization number in response to external stimulus, such as congestion notifications received from the network.

Chapter 4

Network Analysis

The purpose of this chapter is to explore the feasibility of reducing frame relay network congestion by introducing network feedback as an input to video source rate controls. Network and source parameters which significantly affect the performance of the proposed congestion control mechanism are identified and optimized. A comparison of two source rate control schemes, variable quantization and variable frame rate, is performed. Finally, an analysis of the influence of improved (GOB-level) source models on end to end network performance is presented.

4.1 Network Simulation

A software simulation of a mixed traffic (LAN/Video) frame relay enterprise network was developed in a UNIX environment with the SIMSCRIPT II.5 discrete event simulation language.

4.1.1 Network Topology

The network model includes two corporate branch offices and a headquarters office, each equipped with a private branch exchange (PBX) and each connected to a cor-

Figure 4.1: Frame relay network simulation

porate frame relay network over full duplex, fractional T1 (FT1) leased access links. Additional (notional) branch offices exist to provide background traffic over the network. Frame relay assembly and disassembly (FAD) functions are integral to each PBX.

The frame relay network comprises two network nodes (switches) which are joined together by a full duplex T1 trunk. The access links and trunk line are unchanneled and the entire bandwidth is available to any logical connection "on demand". Each network switch comprises a single receive queue (RQ) to the switch CPU and separate transmit queues (TQ) for each physical connection. Congestion notification is initiated by monitoring the occupancy of a selected transmit queue and setting the BECN bits on reverse traffic whenever a predetermined threshold, θ_1 , is exceeded. An absolute threshold, θ_2 , is also defined for the transmit queue, at which point all received packets are discarded.

Two permanent virtual circuits (PVC1, PVC2) are established between the first branch office and the corporate headquarters. PVC1 carries multi-source video information (video1/video2) in H.261 QCIF format from the branch office to the headquarters. PVC2 carries bi-directional interLAN ethernet traffic between the first branch office and another (notional) branch office. The average data rate associated with video1, video2 and the LAN traffic is nominally equal, so that each of the three sources comprise approximately one third of the total offered load on the access link (link 1). A third PVC (PVC3) carries single source video information (video4), in NTSC CIF format, from the second branch office to the corporate headquarters. Return traffic over PVC1 and PVC3 comprises messages which are used to send control instructions and status queries to remote cameras and may be used by the network to carry BECN notifications. Additional traffic (LAN2, LAN3, video5, video6) is generated to load the network switches but the total number of sources is intentionally kept low, to avoid statistical smoothing of the aggregate bit rate. The combined traffic is sufficient to provide a nominal 70% network utilization, as measured at each output queue server.

4.1.2 Traffic Models

GOB Ratio models, the details of which appear below, were developed for the primary video sources, video1, video2 and video4. Frame sized bursts of digital video information arrive at the source PBX receive queue at time intervals corresponding to the inverse of the frame rate. The video data is parsed into GOB sized PDUs, assembled into frame relay packets with the appropriate address (DLCI) and switched to the PBX transmit queue. The following video traffic characteristics are specified:

Video1 - is a GOB level model of a multiscene QCIF sequence comprising three scene segments (M2,M3,M5), each of which is based on empirical data generated with the DTC codec. Individual GOB ratio models were developed for the three sets of empirical data (M2,M3,M5). Each GOB model is selected by a video generator algorithm, which switches sequentially between the three scene segments. The duration of each scene is exponentially distributed with a minimum scene duration of 1 second and a maximum scene duration of 8 seconds. The talking head segment "M2" is a low activity scene with a mean GOB size of 168 octets and a GOB ratio of $G_R(M2) = [0.453, 0.236, 0.311]$. The segment "M3" is a higher activity talking head video with a mean GOB size of 535 octets and a GOB ratio of $G_R(M3) = [0.380, 0.195, 0.425]$. The third segment "M5", is a relatively *bursty* scene caused by a camera which is intermittently panned across a crowded work area. The mean GOB size is 407 octets and the GOB ratio is $G_R(M5) = [0.480, 0.245, 0.275]$. The combined video sequence operates at an initial frame rate of 6 fps and an initial quantization number of ($Q = 10$).

Video2 - is a GOB level model of a single scene QCIF talking head sequence with a mean GOB size of 154 octets and GOB ratio $G_R(QHEAD) = [0.257, 0.447, 0.296]$. The initial frame rate is 15 fps and the initial quantization number is ($Q = 10$).

Video3 - is the empirical data upon which the video1 model is based. Video3 is generated by reading the empirical data for M2, M3 and M5 from individual ASCII files. Video3 can be used to replace video1 in the network simulation, to allow a comparison between the end-to-end performance of empirical and

modeled video traffic. The initial frame rate is 6 fps and the initial quantization number is ($Q = 10$).

Video4 - is a GOB level model of the QCIF sequence "Tennis". The mean GOB size is 875 and the GOB ratio is $G_R(QTEN) = [0.371, 0.343, 0.286]$. The initial frame rate is 5 fps and the initial quantization number is ($Q = 10$).

Video5 - is a *frame level* model of the NTSC sequence "Tennis". The frame rate and quantization number are fixed at 8 fps and ($Q = 10$) respectively. Video5 is not subject to rate control and produces background traffic which is used to load the network switches and the trunk line (LINK2) in the forward direction (same direction as Video 1,2,4).

Video6 - is a *frame level* model of a multiscene NTSC sequence. The frame rate and quantization number are fixed at 30 fps and ($Q = 10$) respectively. Video6 is not subject to rate control and produces background traffic which is used to load the network switches and the trunk line (LINK2) in the reverse direction (from node2 to node1).

TES data models were developed for all simulated LAN sources. The LAN packet arrival process at the source PBX receive queue is batch Poisson. The batch size follows a truncated geometric distribution with a mean value of 25. The LAN packets are framed, without segmentation, assigned a DLCI and switched to the PBX transmit queue.

Another common approach to modeling data traffic involves multiple independent ON-OFF sources, with exponentially distributed ON-OFF periods. However, the

batch Poisson approach has been shown [37] to be more appropriate for the modeling of LAN interconnections and was selected for this reason.

4.1.3 Network Delays

Traffic propagating through the network simulation experiences variable queueing delays at the receive and transmit queues associated with each PBX and each network switch. The input queueing delay represents the delay incurred while the packet waits to be served by the central CPU at the PBX or network switch. The output queueing delay represents the additional delay incurred, after the packet has been processed by the central CPU, while it waits for access to the outgoing transmission line. The following fixed delays are also imposed on each packet traveling through the network:

Framing Delay - Appending and stripping of frame relay headers and trailers to each PDU is accomplished by the source/destination PBX. The FAD delay time, T_{FAD} , is imposed on each packet, at both the source and destination, once access is gained to the central CPU.

$$T_{FAD} = 2.0 \text{ msec}$$

Switching Delay - Once the LAN/Video PDU is encapsulated in a frame relay packet and has entered the network, the CPU at each network switch must read the DLCI address, contained in the frame relay header, and assign the packet to the appropriate transmit queue. The process time associated with this task, T_{SW} is imposed on each packet once it gains access to a network switch CPU.

$$T_{SW} = 0.25 \text{ msec}$$

Propagation Delay - The propagation delay, T_{PROP} is calculated for each hop and is proportional to the distance.

$$T_{PROP} = 2.3 * 10^8 \text{ m/s}$$

Transmission Time - The transmission time, T_{TRANS} , is the time required to transmit the packet once it reaches the head of the output queue. The value for T_{TRANS} is determined by dividing the packet length by the line speed.

4.1.4 Reference Connection and Source Rate Control

The reference connection comprises video1, video2 and video4, which originate in two different nodes, but share a common destination and a common path through the network. Congestion on the reference connection is achieved by constricting the bandwidth over the final link (link 3). The occupancy of the transmit queue to the final link (NODE2 TQ1) is monitored by the NODE2 CPU to determine when the congestion threshold θ_1 and the absolute threshold θ_2 are exceeded. When θ_1 is exceeded, the BECN bit is set on all traffic returning to either branch office. When a BECN bit is received by a branch office PBX, rate control is initiated for all *video sources* originating from that branch. Throughout this work, LAN and background video sources are not subject to flow control.

Rate control at the video sources (video1, video2, video4) is achieved by either reducing the frame rate or increasing the quantization number, but not both. Starting from the desired bit rate R_2 , the rate is decreased incrementally with each received BECN notification, until congestion subsides or a specified minimum bit rate, R_1 ,

is achieved. Since the average packet arrival rate at each branch office is 100 packets/second, several congestion notifications are typically received in rapid succession and roll back to the minimum bit rate occurs very rapidly.

To begin the process of restoring all video sources to their R_2 rate(s), a predetermined number of clear BECN bits must first be received. The number of clear bits that must be received equates roughly to a damping time constant and is denoted ζ , with units (*ms*). The translation from “number of bits” to units of time is achieved by recognizing that the average packet arrival rate at each branch office is 100 packets/second. Nineteen clear bits therefore equates to an average time interval of $19/100$ sec, or 190 *ms*. The rate restoration process, once initiated, proceeds incrementally with dwell periods of magnitude ζ imposed at each rate increment. Receipt of another congestion notification (BECN bit set) will immediately halt the restoration process and re-initiate rate reduction procedures. This implementation, which permits rapid rate reductions and relatively slow rate increases, allows rapid convergence to optimal source rates during periods of congestion.

4.1.5 Packet Loss and Error Recovery

The end-to-end transit time of all packets on the reference connection is monitored. A maximum acceptable end-to-end time of *0.35 seconds* is specified for the reference connection, which corresponds to the maximum cell delay tolerance for videoconferencing applications [5]. Packets with excessive transit times are discarded at the destination PBX. Packet loss on the reference connection can therefore occur in two ways:

- i Packets (video1, video2, video4) can be discarded at the second network switch (NODE 2) if the occupancy of the transmit queue (TQ2) exceeds the absolute threshold θ_2 .
- ii Packets can be discarded at the destination PBX (PBX2) if the end-to-end transit time exceeds the specified acceptable end-to-end delay time.

If a packet is lost, an error is detected by the transport layer when the next arriving GOB is observed to have the wrong sequence number. Whenever this condition is observed, a return message is transmitted to the source requesting the transmission of a new intraframe.

4.2 Simulation Methodology

The network simulation was tested under a variety of conditions to verify that the model was correctly implemented in software. End-to-end packet delays were monitored under light traffic conditions (no queueing delays) and observed to approach the aggregate end-to-end fixed delay value of 41 ms, for an *average sized* 500 octet video packet. Under heavier traffic loads, mean queue lengths were monitored and the expected average end-to-end delay was computed as the sum of the fixed delays and the mean queueing delays. The expected delay value was found to agree with the average end-to-end delays generated by the network simulation.

4.2.1 Transient Time

If a simulation begins at some arbitrary time, t_0 , the first packet through the simulated network will not experience queueing delays and the observed end-to-end transit time

Figure 4.2: Average packet delay vs simulation run time

for the first packet will not be indicative of *typical* network performance. Clearly, some finite period of time must elapse before the switch buffers become populated and the simulation reaches *steady state*. We define a steady state system as one which is in statistical equilibrium, such that the probability that the system is in a given state is not time dependent. The amount of time that elapses before the simulated system reaches steady state is the *transient time*, t_1 .

To determine the value of t_1 , the average end-to-end delay for video packets was computed at 1 second intervals and plotted, as shown in figure 4.2. The plot suggests that average end-to-end delay values collected during the first 4 seconds of simulated time are downward biased. To prevent network performance statistics from being skewed by initial (transient) conditions, all simulations were allowed to run for 5 seconds of simulated time prior to commencing data collection.

4.2.2 Number of Runs

A fixed run length of 30 simulated seconds was selected and the number of replications (simulation runs) necessary to estimate the average end-to-end delay with an accuracy of 5.0 ms, and 95% confidence, was determined. Using the procedure outlined in [31], we find that the half length (h.l) of a $100(1 - \alpha)\%$ confidence interval for a mean θ , based on the Student's t -distribution, is given by:

$$h.l = t_{\alpha/2, R-1} \hat{\sigma}(\hat{\theta}) \quad (4.1)$$

where $\hat{\sigma}(\hat{\theta}) = S/\sqrt{R}$, S is the sample standard deviation, and R is the number of replications (simulation runs).

To meet an end-to-end delay half length criterion of $\epsilon = 5.0$ ms, a sample size R must be chosen such that:

$$h.l = \frac{t_{\alpha/2, R-1} S}{\sqrt{R}} \leq \epsilon \quad (4.2)$$

Solving the inequality for R , it is possible to determine the minimum number of simulation runs that must be executed to estimate the average end-to-end delay with the required accuracy ($\epsilon = 5.0$ ms) and confidence (95%). The necessary number of simulation runs is therefore given by:

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

The average end-to-end delay for video packets was collected from sixteen independent simulations, each of 30 simulated seconds duration. The sample standard deviation was computed to be $S = 0.0077$. For a 95% confidence interval and a required accuracy of 5.0 ms, we have $t_{0.025,15} = 2.13$ and $\epsilon = 0.005$. The required number of simulations is therefore:

$$\begin{aligned} R &\geq \left(\frac{2.13 \times 0.0077}{0.005} \right)^2 \\ &\geq 10.76 \end{aligned} \quad (4.4)$$

It is therefore apparent that $R = 16$ satisfies the inequality and provides an excellent estimate of end-to-end performance with a high level of confidence and a high degree of accuracy. Accordingly, a minimum of 16 replications, each of 30 simulated seconds duration, were performed for all network simulations conducted in support of this thesis.

4.2.3 Cost Function

A common expression for network performance is necessary to reconcile various and conflicting criteria (low loss, high throughput, acceptable delay) into a unified measure of global performance. A weighted *cost function* was developed for this purpose and is defined by the following expression:

$$C = \begin{cases} \frac{L + (1.0 - P) + (1.0 - U)}{2.4} & \text{if } L \leq 0.01 \\ \frac{20 \times L + (1.0 - P) + (1.0 - U)}{2.4} & \text{if } .01 < L \leq 0.02 \\ 1.0 & \text{if } L > 0.02 \end{cases}$$

where L is the aggregate packet loss for video 1/2/4, P is the actual throughput as a fraction of the *target* throughput R_2 for video 1/2/4, and U is the transmit queue server utilization measured at NODE2 TQ1.

When the total packet loss on the reference connection is less than 1 percent, equal weight is assigned to loss, target data rate and network utilization. However, as loss increases, the throughput and network utilization will also increase and, without biasing, the overall cost function will decrease. To avoid this situation, a bias is added to the cost function when packet loss on the reference connection rises above 1 percent. We observe that for loss rates in the range $[0.0 \leq L \leq 0.02]$, the maximum value that can be assumed for the numerator is 2.4, although in practise the numerator value is rarely greater than 2.0. A constant denominator of 2.4 is therefore included to limit the range of the cost function to $[0.0 \leq C \leq 1.0]$ for aggregate loss rates not greater than 2 percent ($L \leq 0.02$). A packet loss rate in excess of 2 percent is considered unacceptable and, under these circumstances, the loss function defaults to its maximum value of 1.0.

It is useful to recall that congestion is induced in the network simulation by reducing the bandwidth on the final link (link3), which increases T_{TRANS} and the average end-to-end transmission time. The cost function is therefore sensitive to the link3 bandwidth, since packet discard due to maximum end-to-end delay will increase as the bandwidth is constrained. In the following sections, comparisons are made between various network parameters, source rate control techniques and traffic models. In every case, however, cost function observations for different conditions are recorded at a fixed congestion level (link3 bandwidth), thus ensuring a valid basis for direct comparison.

DESIGN POINT	FACTOR			SIMULATION RESULT
	1	2	3	
1	+	+	+	C1
2	-	+	+	C2
3	+	-	+	C3
4	-	-	+	C4
5	+	+	-	C5
6	-	+	-	C6
7	+	-	-	C7
8	-	-	-	C8

Table 4.1: Generic 3 factor design analysis

4.3 Simulation Results

4.3.1 Experimental Design

Preliminary experimentation with the network model provided evidence that changes to the congestion threshold θ_1 , the absolute threshold θ_2 , the damping constant ζ and the BECN delay (D) would affect network performance, as measured by the cost function. To determine which parameters (or combinations of parameters) have the most significant influence on network performance, a 2^k factorial analysis [38] was performed.

Factorial analysis supports the characterization of several design parameters (factors) with reference to a standard system response measure, the cost function. Two levels (discrete values or ranges) are selected for each of the k factors. Simulations are then run for each of the 2^k possible factor-level combinations, from which the effect of each factor and combinations of factors can be determined.

Table 4.1 illustrates a notional experiment involving three factors (F_1, F_2, F_3). The two levels associated with each factor are denoted “+” (high) and “-” (low). Eight different combinations of factors and levels are therefore possible and each

combination defines a *design point*. The simulation result for each design point is expressed as a cost function value, $(C_1, C_2, \dots, C_8)$. The *main effect* for a particular factor is determined by computing the average difference in simulation results when the factor level is high and low, and the levels for all other factors remain fixed. For example, the main effect e_1 , for the first factor F_1 , is given by:

$$e_1 = \frac{(C_1 - C_2) + (C_3 - C_4) + (C_5 - C_6) + (C_7 - C_8)}{4} \quad (4.5)$$

Similarly, the *combined effect* $e_{1,2}$ of two factors F_1, F_2 is defined as *half* the difference between the average effect of factor F_1 when factor F_2 is at its "+" level (and all factors other than F_1 and F_2 are held constant) and the average effect of F_1 when F_2 is at its "-" level. Referring to table 4.1, the combined effect for factors F_1 and F_2 is given by:

$$e_{1,2} = 1/2 \times \left\{ \left[\frac{(C_1 - C_2) + (C_5 - C_6)}{2} \right] - \left[\frac{(C_3 - C_4) + (C_7 - C_8)}{2} \right] \right\} \quad (4.6)$$

The remaining main and combined effects can be computed in the same way.

When several simulation results are collected for each design point, the main effects $(e_1, e_2, \dots, e_k)$ and combined effects $(e_{12}, e_{13}, \text{etc})$ can be calculated as confidence intervals, which are influenced by the variability of the measured effect and the number of simulations performed for each factor-level combination. A completely positive or negative confidence interval indicates that the factor, or combination of factors, has a statistically significant influence on the performance of the system and is a good candidate for optimization efforts. The factorial analysis results were, in fact, applied

to a stochastic optimization process known as Mean Field Annealing (MFA), which is discussed in the latter part of this chapter.

FACTOR	SYMBOL	LEVEL 1	LEVEL 2
D	E1	0.0 s	0.5 s
θ_1	E2	35000 bits	55000 bits
θ_2	E3	65000 bits	85000 bits
ζ	E4	50 ms	500 ms

Table 4.2: Factors and levels selected for design analysis

Table 4.2 illustrates the 4 factors and the levels for each factor that were selected for design analysis. Sixteen independent, steady-state simulations were performed for each of the 16 factor-level combinations. The level of network congestion (link3 bandwidth) was maintained at a constant value for all simulations. The rate control method was variable frame rate with a minimum value of 2 fps defined for all video sources. Ninety five percent confidence intervals for each of the primary and secondary effects were computed and are illustrated in figure 4.3. These results indicate that the values of the BECN delay (E1), the congestion threshold (E2) and the damping constant (E4) have significant bearing on the network performance, as determined from the cost function. Second order effects (combinations of factors) and the absolute congestion threshold (E3) have no observed influence on the cost function. The insensitivity of the cost function to the absolute congestion threshold is indicative of the effectiveness of the rate control feedback mechanism. In essence, if throughput is successfully regulated, buffer overflows should not occur.

Figure 4.3: Confidence intervals (95%) for primary and secondary effects

A second factorial analysis was conducted, which was identical to the first except that variable quantization rate control was used in place of variable frame rate control. The results from this second experiment were very similar to the first. In particular, primary factors E1 (BECN delay), E2 (congestion threshold) and E4 (damping constant) were observed to significantly influence the value of the cost function. Once again, the value of the absolute congestion threshold (E3) had no observed influence on overall network performance.

4.3.2 Minimum Data Rate

The sensitivity of the cost function to the minimum source rate, R_1 , was investigated by performing a 2^k factorial analysis with $k = 1$. The *floor* value (F) is defined as the ratio (R_1/R_2 , where R_1 is the minimum source rate and R_2 is the target rate. Two values for F were selected:

$$F = 0.2$$

$$F = 0.5$$

Sixty four sets of independent data points were collected using variable frame rate source control. Each set of points comprised two measurements of the cost function, one each for the two F values. Within a particular set of data points, identical traffic conditions were achieved for both cost function measurements by initiating both simulation runs with the same seed values. The difference between the two cost function measurements was computed for each set of points. The average difference was determined at 95 percent confidence with the following result:

$$\begin{aligned} \Delta C &= C(F = 0.5) - C(F = 0.2) \\ &= 0.0619 \pm 0.0558 \end{aligned} \tag{4.7}$$

The experiment was repeated for variable quantization rate control and similar (statistically significant) results were observed. It is therefore evident that the congestion control mechanism is sensitive to the value of F .

To explore further the effect of F on network performance, the average cost function was tabulated for several different values of F . The result is illustrated in figure 4.4 and suggests that an F value in the range $(0.3 \leq R_1 \leq 0.6)$ is optimal for video sources subject to variable quantization or variable frame rate control. The vertical bars represent the 95 percent confidence interval for each data point. If the floor is set too high (greater than 60% of the target value), the *dynamic range* of

Figure 4.4: Floor (F) vs cost function with 95% confidence intervals

the source rate is insufficient and congestion can not always be effectively controlled. Packet loss rates escalate and the value of the cost function is occasionally set to unity, to indicate a loss rate of greater than 2 percent. Because the transition from a normal cost function ($L \leq 0.02$) to unity ($L > 0.02$) is not smooth, the variance increases for a cost values collected in this region. The large confidence intervals on the right side of figure 4.4 are due to this phenomenon.

When the floor value is set too low ($F < 0.3$), the rate control mechanism has too much dynamic range and will essentially over-react to network congestion notifications. Decreasing the data rate to approximately 50 percent of the target rate will effectively alleviate network congestion and reduce loss rates to acceptable levels. Further improvements to the packet loss rate, achieved by reducing the source rate below the optimal range ($0.3 \leq F \leq 0.6$), appear minimal and may not adequately mitigate the adverse effects of lower link utilization and data throughput.

4.3.3 BECN Delay

BECN delay (D) is defined as the time elapsed between when congestion is detected at NODE2 and the first BECN notification arrives at the source PBX. If BECN delay becomes sufficiently large, congestion information received at the source will not accurately reflect current conditions in the network and the effectiveness of rate control measures will be degraded. To assess the amount of BECN delay that can be tolerated, the temporal variations of the NODE2 TQ1 buffer occupancy were characterized by computing the autocorrelation of queue length as a function of time lag. Referring to figure 4.5 the queue length correlation remains above 0.8 when ($D \leq 0.05\text{sec}$) and above 0.7 when ($D \leq 0.20\text{sec}$). This information suggests that congestion notifications which experience delays of less than 0.2 seconds will approximately reflect current network conditions and will result in appropriate rate adjustments at the source PBX. BECN delays of 0.05 seconds or less are clearly preferred.

Under typical network conditions (70% utilization) the average BECN delay was determined to be 0.0652 seconds with a 95% confidence interval of $[0.0634 \leq D \leq 0.0670]$. To further verify that BECN delays of this magnitude are acceptable, the network simulation was modified to allow a *fixed* BECN delay to be specified. A plot of BECN delay vs cost function was developed and is depicted in Figure 4.6. This figure, which illustrates the relationship between BECN delay and network performance, reveals a somewhat intuitive trend. Moreover, the quantitative results depicted in Figure 4.6 confirm that network performance achieved with BECN delays of approximately 0.065 seconds is comparable to the performance achieved with zero BECN delay. This latter result is consistent with the previous autocorrelation re-

Figure 4.5: Autocorrelation of buffer occupancy as a function of time

formation, which also suggests that very little degradation occurs for average BECN delays in the 95% confidence interval $[0.0634 \leq D \leq 0.0670]$.

Referring once again to figure 4.5, the periodic behaviour observed at lags of 0.2 seconds is attributed to the influence of video4, which operates at 5 fps and comprises 50 percent of the total traffic through NODE2 TQ1. We are thus reminded that all traffic sources on the reference connection operate at reasonably low frame rates and are more tolerant to long BECN delays than video sources operating at broadcast quality frame rates (30 fps). For example, the aggregate frame rate from all sources on the reference connection is 26 fps. Assuming a BECN delay of 0.05 seconds, an average of 1.3 “uncontrolled” frames will be transmitted after congestion is detected and before source rate control is implemented. Conversely, for three broadcast quality video sources each operating at 30 fps (90 fps aggregate), a BECN delay of 0.014 seconds is required to limit the average number of uncontrolled frames to the same

Figure 4.6: Cost as a function of BECN delay

value. Nevertheless, networks capable of supporting broadcast quality video, which has a maximum end-to-end delay tolerance of 0.040 seconds [5], are sufficiently “fast” that an average BECN delay value of 0.014 seconds is not unrealistic. It is therefore probable that BECN initiated rate control is applicable to networks supporting a broad range of video applications and frame rates.

4.3.4 Design Optimization

Optimization of network performance can be achieved by locating the global minimum of the cost function, which has a multi-dimensional solution surface. The optimization effort involves computing the value of the cost function at several points on the solution surface and then applying some methodology, such as gradient descent, to locate the minimum value. In this study, a closed form solution for the cost function is not available and the true value of the cost function must be approximated by statistical

estimates, derived through Monte Carlo simulation. Unfortunately, conventional optimization techniques (Gradient descent , Fletcher-Powell algorithm) do not perform satisfactorily when applied to inherently “noisy” estimates of deterministic values [39, 40].

Mean Field Annealing (MFA), a variant of standard simulated annealing, is a statistically based technique which is well suited to the general problem of optimizing stochastic cost functions over a multi-dimensional parameter space [41]. In particular, MFA is resistant to variations in the cost estimate from simulation, is not susceptible to local minima in the solution surface and converges rapidly to near-optimal solutions. The MFA algorithm randomly selects a dimension, N , in the parameter space $[S(M, N, P, \text{etc})]$ and generates estimates of the cost function C , at m evenly spaced increments (of size q) over the interval $[N_{min}, N_{max}]$. A new value, a_N , is selected for the N axis by computing the following weighted average of the cost function estimates:

$$a_N = \frac{\sum_{j=0}^{m-1} (N_{min} + jq) \exp(\frac{-C_j}{T})}{\sum_{j=0}^{m-1} \exp(\frac{-C_j}{T})} \quad (4.8)$$

where terms with larger values of the cost C_j are assigned less weight than terms with smaller values. This process is repeated iteratively for each dimension while the value for all other dimensions remain fixed, and until equilibrium is reached.

The parameter T , which is commonly referred to as the *temperature*, provides a mechanism for escaping from local minima in the solution surface. While equation 4.8 guarantees that a_N is, *on average* weighted in favour of the minimum value of the cost estimate, a move in the direction of increased cost can occasionally be taken.

PARAMETER	MIN VALUE	MAX VALUE	STEPSIZE
D (s)	0.0	0.5	0.05
θ (bits)	10000	60000	5000
F	0.1	1.0	0.1
ζ (ms)	50	550	50

Table 4.3: Parameters and bounds used in MFA optimization

The probability of an uphill move of magnitude ΔC is given by $P_{UPHILL}(\Delta C) = \exp(-\frac{\Delta C}{T})$. By gradually lowering the temperature over time, the probability of an uphill move is slowly decreased and the biasing applied to the weighted averaging is proportionally increased. This combination of effects allows the MFA algorithm to avoid local minima and to gracefully relax to an optimal or near optimal solution.

A four dimensional MFA algorithm was implemented and linked for interactive execution with the frame relay network simulation. The response surface for the cost function was sampled at discrete intervals over the critical parameter space, as shown in table 4.3. Sixteen independent simulation runs, each of 30 simulated seconds duration, were conducted at every sample point in each of the four dimensions. A minimum of three complete passes was made in each of the four dimensions before the MFA temperature value was permitted to decrease. The initial MFA temperature value was 0.7, which is comparable to the maximum cost function value. The final MFA temperature value was 0.04. The initial temperature was decremented geometrically by a factor of 0.9 until the final temperature value of 0.04 was reached. The global minimum of the cost function solution surface was successfully located by the MFA algorithm, which converged to the near-optimal solutions in table 4.4.

The values for D and F are clearly consistent with earlier results, in which the response surface cross-sections for D and F were analyzed in isolation. These previous results are expected to remain valid, as the factorial design analysis indicates no significant second order (combined) effects. Figure 4.7 and figure 4.8 illustrate the response surface cross-sections for θ_1 and ζ respectively, which were recorded during execution of the MFA algorithm. Once again, the one dimensional global minima shown in these figures agree closely with the near optimal values determined by the MFA algorithm. It is also useful to note that the solutions for θ_1 and ζ are consistent with the factorial design result, which correctly predicts better network performance at $\theta_1 = 35000$ than at $\theta_1 = 55000$ and at $\zeta = 500$ versus $\zeta = 50$.

Although MFA converges to optimal solutions, it does not provide a mechanism to identify *ranges* of near optimal values. In Figure 4.7, for example, it is apparent that good network performance is achieved for any threshold value in the range ($30000 \leq \theta_1 \leq 50000$). Ranges of critical parameter values over which good end-to-end performance can be achieved, were determined by examining response surface cross-sections for each of the four critical parameters. These ranges are shown in table 4.5.

PARAMETER	NEAR OPTIMAL VALUES DERIVED WITH MFA
D	0.05 s
θ_1	36600 bits
F	0.50
ζ	190 ms

Table 4.4: Near optimal values for network parameters determined by MFA

Figure 4.7: Congestion threshold θ_1 vs cost function using MFA

The performance of an uncontrolled network was compared to that of a controlled network, using variable frame rate source control and near optimal parameters. The bandwidth of the final link (link 3) was incrementally reduced, to induce increasingly severe congestion at the final network node (NODE2 TQ1). The performance of the controlled network is vastly superior, as illustrated in figure 4.9.

PARAMETER	NEAR OPTIMAL RANGE
D	0.0 - 0.1 s
θ_1	30000 - 50000 bits
F	0.3 - 0.5
ζ	150 - 200 ms

Table 4.5: Ranges of near-optimal values for network parameters

Figure 4.8: Damping (ζ) vs cost function using MFA

Figure 4.9: Cost function vs link3 bandwidth: with and without BECN rate control

4.3.5 Comparison of Rate Control Schemes

The effect of variable frame rate source control and variable quantization source control on end-to-end network performance was compared. Sets of data were collected for each of 64 independent test points. The data set associated with each test point comprised two measurements of the cost function C , the packet loss L and the normalized video throughput P , one measurement of each parameter for both source rate control methods (frame rate and quantization). For each of the 64 test points, identical traffic conditions were achieved for both sets of measurements by initiating both simulation runs with the same seed values. A direct comparison of the two source control methods was made by calculating difference values for the cost function, utilization and source throughput measurements collected at each test point. The following 95 percent confidence intervals were computed by subtracting the values for variable frame rate control from the values for variable quantization rate control:

$$\Delta C = -1.66 \times 10^{-2} \pm 0.688 \times 10^{-2}$$

$$\Delta L = -4.80 \times 10^{-4} \pm 7.25 \times 10^{-4}$$

$$\Delta P = 1.17 \times 10^{-2} \pm 0.623 \times 10^{-2}$$

The 95 percent confidence interval for ΔL contains zero and the result is therefore inconclusive. There appears to be no significant difference between the two rate control methods in terms of packet loss. A significant difference does, however, exist between the two rate control schemes when throughput or the cost function are used as the performance criteria. These results indicate that variable quantization source

CONTROL METHOD	MEAN LEVEL NODE2 TQ1 (BITS)	MAX LEVEL NODE2 TQ1 (BITS)	STD
NONE	24504	92399	26050
FRAME	9336	75832	11143
QUANT	10736	60135	11762

Table 4.6: Queue length statistics at NODE2 TQ1 for three control conditions

Figure 4.10: Queue length vs time - no rate control

control yields significantly better throughput than variable frame rate source control, while maintaining equivalent packet loss figures. Variable quantization rate control therefore offers significantly better end-to-end performance in terms of throughput and the overall cost function.

To further illustrate the advantages of variable quantization rate control over variable frame rate control, the NODE2 TQ1 queue length was monitored as a function of time. Temporal plots of the NODE2 TQ1 queue length are given in Figure 4.10

2 of/de 2

PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

PRECISIONSM RESOLUTION TARGETS

Figure 4.11: Queue length vs time - variable frame rate control

(no control), Figure 4.11 (variable frame rate) and Figure 4.12 (variable quantization), while Table 4.6 gives the mean queue length, maximum queue length and standard deviation for the same three conditions. Referring to Table 4.6, it is apparent that both control schemes offer significant improvements over an uncontrolled network, and that the two control mechanisms are roughly equivalent in terms of mean queue length and standard deviation. Also, having previously observed that the packet loss statistics are similar for variable frame rate and variable quantization rate control, it is evident that both control mechanisms can effectively limit the maximum queue length, which is a principle cause of packet loss due to excessive delay. However, variable quantization rate control achieves effective rate control more *efficiently*. As shown in Table 4.6, the maximum queue length for variable quantization rate control is substantially smaller than for variable frame rate control, although the average throughput is significantly higher.

Figure 4.12: Queue length vs time - variable quantization rate control

4.4 Comparison of Empirical Data, TES GOB and TES Frame Models

To this point, all network simulations cited in this thesis have used the traffic models described in section 4.1.2. GOB level models have therefore been used for all video sources on the reference connection (video1, video2, video4). The rationale for employing GOB level models is to provide improved resolution of temporal bit rate fluctuations over short time intervals, equivalent to the GOB scan rate. Referring to Table 4.6 however, the average queue length at NODE2 TQ1 is approximately 10000 bits (1250 octets), which corresponds to roughly 3 GOB sized FR packets. Since a QCIF frame comprises 3 GOBs, it is reasonable to expect that intraframe fluctuations captured by the GOB source model are subsequently lost due to *smoothing* in the network buffers. The video1 traffic generator in the FR network simulation was modified to allow a direct comparison between the end-to-end performance charac-

teristics for empirical traffic (E1), a TES frame model (F1) and the original video1 GOB ratio model (G1). The network statistics module was also modified so that end-to-end performance measures (cost function, maximum delay) were collected only for video1, rather than for all video sources on the reference connection.

A three position software switch was installed in the video1 traffic generator. When the first switch position is selected, empirical GOB size data is read from three ASCII files, one each for M2, M3 and M5. The second switch position causes the video1 generator to produce a frame level traffic simulation by replacing the three original GOB ratio vectors with $G_R = [0.33, 0.33, 0.33]$. The final switch position causes the video1 generator to revert to the original GOB ratio model, as defined in section 4.1.2.

Performance information for the empirical video1 traffic (E1), TES frame model (F1) and TES GOB model (G1) were collected at each of 80 independent test points. Within any given test point, the same seed values were used to initiate each of the three simulation runs (E1,F1,G1). Traffic conditions for each of the three simulations were therefore identical, with the exception of the video1 traffic. Difference values for the video1 cost function and maximum end-to-end delay were calculated for all three video1 traffic sources. The following 95% confidence intervals were computed:

$$\Delta C(E_1 - F_1) = 0.0319 \pm 0.0780$$

$$\Delta C(E_1 - G_1) = 0.0422 \pm 0.0732$$

$$\Delta C(F_1 - G_1) = 0.0104 \pm 0.0234$$

Figure 4.13: Maximum end-to-end delay: TES models vs empirical data

$$\Delta D_{max}(E_1 - F_1) = 0.0054 \pm 0.0188$$

$$\Delta D_{max}(E_1 - G_1) = -0.0044 \pm 0.0182$$

$$\Delta D_{max}(F_1 - G_2) = 0.0010 \pm 0.01$$

This result, graphically illustrated in Figure 4.13, strongly suggests that no significant difference exists between end-to-end performance data generated using empirical source data, a TES frame model or a TES GOB model. While it is clear that TES models provide an excellent representation of the empirical source data, it is evident that GOB level models are not necessarily required for the study of end-to-end performance. It is useful to observe that end-to-end network performance is primarily influenced by events such as video scene changes or bursts of LAN traffic, that generally occur over substantially longer time intervals than the GOB scan rate. Significant network congestion is not induced by minor fluctuations in the size distribution of GOB sized packets, which occur over short durations and are subsequently

lost due to smoothing in the network buffers. GOB statistics are therefore expected to have little impact on network congestion phenomena and need not be accurately modeled to achieve reliable simulation results.

4.5 Summary

A mixed traffic (video/LAN) FR network simulation was developed and used to explore the feasibility of reducing congestion by introducing network feedback as an input to video source rate controls. A BECN notification is generated by the network whenever the length of a network output queue exceeds the prescribed *congestion threshold*. Factorial analysis of the network design was performed to identify network parameters which significantly influence the performance of BECN congestion control. The congestion threshold, the BECN transit delay, the source damping constant and the minimum video source rate (floor) were each determined to affect network performance. No combined effects were observed.

A stochastic optimization technique (MFA) was applied to the network simulation to identify optimal design parameters. The following ranges of values were determined to yield near optimal end-to-end traffic characteristics:

$$\begin{array}{lll} D & = & 0.0 \text{ to } 0.1 \quad s \\ \theta_1 & = & 30000 \text{ to } 50000 \quad bits \\ F & = & 0.30 \text{ to } 0.50 \\ \zeta & = & 150 \text{ to } 200 \quad ms \end{array}$$

The effectiveness of BECN initiated rate control was demonstrated by comparing the performance of an uncontrolled network to the performance of a controlled network, using near optimal parameters. It is evident that BECN initiated source

rate control provides an effective mechanism for reducing the output of video sources in response to incipient network congestion.

The relative utility of variable quantization and variable frame rate source control was determined. Both schemes were found to control network congestion effectively but the quantization approach is able to regulate peak buffer occupancies *more effectively* than the variable frame rate control mechanism while maintaining a higher average throughput than the frame rate control method.

A comparison of the end-to-end traffic characteristics of empirical source data, a TES frame model and a TES GOB model was performed. No statistically significant differences were observed between the end-to-end characteristics of any two of the three traffic sources. This result demonstrates that TES models provide an excellent representation of empirical video bit streams. However, it is also evident that short duration bit rate fluctuations, which are accurately modeled at the GOB level, have no appreciable affect on network congestion and end-to-end traffic characteristics. In fact, the average length of network switch queues is sufficient that GOB level bit rate fluctuations are *smoothed* when multiple packets accumulate in a buffer, and effectively become invisible to the network. It is therefore evident that GOB level video source models are not necessarily required for the study of end-to-end network performance issues.

Chapter 5

Conclusions and Recommendations for Further Study

5.1 Conclusions

5.1.1 Video Source Models

We find that H.261 GOB level data is self contained and can be decoded without reference to preceding information. The approximate average and maximum GOB sizes for H.261 video in QCIF format, which is the most practical format for FR video applications, are 500 and 2000 octets respectively. FR packets containing GOB sized PDUs can migrate through many high speed LANs, WANs and FR networks without further partitioning, and the reconstructed image is inherently resilient to cell loss. It is therefore concluded that FR video PDUs should be assembled by parsing the H.261 bit stream along GOB boundaries.

Frame level models of video source data do not capture the variability of GOB size within each video frame and consistently underestimate the true "burstiness" of the empirical bit stream. The slow decay properties and periodic behaviour of typi-

cal GOB level autocorrelations can not be directly modeling using TES techniques. However, frame level TES models can be modulated with an empirically derived GOB ratio vector, which describes the average distribution of GOB size within a video frame. In this manner, the GOB size distribution and autocorrelation characteristics of empirical video sequences can be accurately reproduced.

The effect of variable quantization on the bit rate statistics of a video sequence can be truthfully replicated by applying a linear scaling factor to the output of a single GOB ratio model. Because the TESstool utility will operate interactively with C-based simulations, the model parameters can be varied during execution to emulate non-stationary traffic behaviour. GOB ratio models are therefore well suited for traffic perturbation analysis and generalized simulation studies, which require variations over a broad spectrum of source characteristics.

End-to-end network statistics generated by TES frame and TES GOB (ratio) models are indistinguishable, in a statistical sense, from similar statistics collected from empirical data. We find that while TES models provide an excellent representation of empirical video bit streams, short duration bit rate fluctuations at the GOB level have no appreciable affect on network congestion and end-to-end traffic characteristics. We observe that network congestion is primarily influenced by relatively long duration events, such as video scene changes, and that GOB level bit rate fluctuations are effectively *smoothed* in the network switch buffers. High resolution (GOB level) video source models are therefore not necessarily required for the study of end-to-end network performance issues.

5.1.2 Congestion Control

Under typical loads and moderate frame rates, explicit network congestion notifications are received at video transmitter sites in sufficient time to elicit appropriate and useful flow control responses. We demonstrate that for video telephony and video teleconferencing applications, maximum delay criteria (350 ms) and acceptable loss rates ($L \leq 2\%$) can be achieved by implementing BECN initiated rate control at one of the two network nodes. It is reasonable to expect that much better performance can be achieved if BECN rate control is implemented at all network nodes and for all traffic sources. Moreover, we observe that the network simulation in this thesis represents a reasonable *worst case* scenario, since the network is dominated by a relatively small number of very bursty traffic sources. We conclude that this result, which confirms the utility of BECN initiated source rate control under very *bursty* traffic conditions, can be generalized and is broadly applicable to generic cell relay networks.

Design optimization and stochastic optimization techniques are useful tools for defining and solving the multi-dimensional response surface for congestion control. Network performance, using optimal congestion control parameters, is vastly superior to the performance of an uncontrolled network.

The data rate associated with individual sources must have sufficient *dynamic range* to allow the full benefit of BECN initiated rate control to be realized. A minimum value of between 30% and 60% of the average output was found to be adequate in our study, which simulated a reasonable “worst case” condition. We also find that network performance is far more sensitive to floor values that are set too high ($floor > 60\%$) than to floor values that are set too low. It is therefore concluded

that a floor value of 30% to 50% of the average source output is a useful benchmark for generic cell relay networks and traffic conditions.

We observe that a damping constant can be applied to the source rate control mechanism to prevent rate constraints from being prematurely rescinded, in response to one or two *clear* BECN bits. Intermittent congestion notifications, which occur when the buffer occupancy is near the congestion threshold, can be avoided by maintaining reduced source rate(s) until the buffer occupancy is well below the congestion threshold. The rate of buffer depletion is defined by the ratio of traffic arrival rate to service rate, which is relatively constant for any queue with an average utilization of 70% and a rate reduction floor of approximately 50%. Simulation results indicate a damping constant of approximately 150-200 msec provides near optimal performance, while damping constants as large as 400 msec yield very good cost function values. Accordingly, a damping constant of approximately 200 msec represents a useful benchmark value for generic network applications.

We find that, when variable quantization source control is employed, the peak buffer occupancy (60135) is less than twice the congestion threshold (36600). It is therefore possible to establish congestion thresholds which are tailored to a maximum acceptable buffer occupancy, or equivalently, to a maximum acceptable queueing delay. A similar, although less stringent, rule of thumb can be applied if variable frame rate control is implemented.

Variable quantization rate control is more effective than variable frame rate control in preventing and alleviating network congestion. Variable quantization rate control also simplifies the process of reconstructing the video sequence at the receiving end. Despite random transit delays, an implicit replay rate is defined by the video

frame rate, which is constant when variable quantization rate control is implemented. The receiver is therefore able to absorb the random transit delays in a receiver buffer and feed video PDUs to the H.261 decoder at the implicit frame rate. However, when variable frame rate control is in effect, the average video packet arrival rate is not constant and the appropriate replay rate is not known by the receiver *a priori*. When video PDUs are channeled to the H.261 decoder too quickly, the codec will attempt to display the frames at the correct rate as indicated by the frame number in the H.261 frame header. Limited frame buffering at the codec will, however, quickly require that frames be displayed at the wrong rate to prevent memory overflow. If video PDUs are channeled to the H.261 decoder too slowly, the codec will have no option but to display images too slowly. In this latter case, the length of the receive buffer will begin to grow without bound.

5.2 Recommendations for Further Study

5.2.1 Rate Control

In addition to variable quantization and variable frame rate, the H.261 motion detection algorithm offers a third potential mechanism for regulating the output bit rate. The H.261 codec performs motion detection independently, on every 16×16 pixel macroblock. The magnitude of the resulting motion vector is compared to a predefined *motion threshold*, which determines whether the MB is *sufficiently different* from the corresponding block in the previous frame to warrant encoding and transmission. Adjusting the motion threshold will, therefore, influence the number of MBs that are transmitted, along with the aggregate bit rate.

5.2.2 Discard Eligibility

The DE bit in the FR header can be set by the source to identify packets that should be jettisoned in preference to others in the event of network congestion. This thesis has not addressed the issue of how H.261 data may be organized and parsed into high and low priority FR packets. Despite its four level coding hierarchy, H.261 GOBs contain a mixture of intracoded and intercoded MBs, most of which contain essential (intracoded) information. The feasibility and utility of modifying the H.261 coding algorithm to segregate GOBs into intracoded (essential) GOBs and intercoded (non-essential) GOBs should be investigated. The DE bit can then be set on FR packets containing intercoded GOB sized PDUs.

An alternative approach to the DE bit issue may be to modify the H.261 algorithm to support two-level sub-band coding. The coder would make one encoding pass, using coarse quantization levels or a high motion detection threshold, to capture essential scene information at low resolution. A second encoding pass would then be made, using a smaller quantization step size or a lower motion threshold, to capture finer image detail. The DE bit could be set on FR packets carrying fine detail information acquired during the second pass. Although coding delay will increase, a layered coding structure would support more graceful image degradation during network congestion by allowing a more intelligent delineation between high and low priority traffic. Motion detection rate control and layer coding strategies can both be explored using existing, public domain, H.261 software codecs.

5.2.3 Video Adaptation Layer

It is evident that some additional information about the video signal must be included at the beginning of each FR PDU, for interpretation by a higher level protocol at the receiving end. A sequence number, for example, is required to detect cell loss/discard and to initiate loss recovery procedures. When variable frame rate control is in effect, frame rate information must be included to allow playback at the correct cadence. If GOB sized PDUs prove too large for efficient network operation, MB sized PDUs may be necessary. Packet loss resilience in MB sized PDUs can be enhanced by including absolute values for the MB number in the video adaptation layer [42, 19]. Also, variable length coded H.261 GOB and MB information does not necessarily comprise an integer number of octets, which is a constraint imposed on PDUs by the FR (and ATM) protocol. Byte alignment of H.261 GOB and MB binary information may be achieved by invoking existing H.261 *bit stuffing* provisions at the adaptation layer.

A video adaptation layer format has been proposed for H.261 applications on TCP/IP networks and represents a useful starting point for similar FR considerations. FR video adaptation layer requirements should be investigated in detail and a proposed implementation should be validated through simulation. Particular emphasis should be placed on resilience to cell loss and synchronization of video and audio information.

5.2.4 Improved Traffic Model and Congestion Control

The traffic model used in this thesis should be expanded to consider the following possibilities:

- a Two way video communications required to support teleconferencing:

b Mixed LAN, video and packet voice communications.

c Fractal data traffic.

In this thesis, BECN congestion notification was only implemented at the 2nd network switch (NODE2), and not for all traffic. A comprehensive congestion notification model should be developed to include controls for all traffic types on all connections. Optimal control schemes for data traffic (such as leaky bucket) should be identified along with practical limitations for data packet size and burst duration.

5.2.5 Adaptive Buildout Delay

An adaptive buildout delay study has been completed for packet voice traffic in FR networks [14]. A similar study should be completed to determine if an adaptive buildout delay implementation is feasible and desirable for video traffic. A video replay capability should be added to the existing network simulation to assist in the subjective analysis of reconstructed video quality. Public domain codecs and video conferencing software are available and may be suitable for adaptation to FR applications.

References

- [1] R. Haendel and M. Huber. *“Integrated Broadband Networks”*. New York: Addison-Wesley, 1991.
- [2] International Telegraph and Telephone Consultative Committee. *CCITT Recommendation Q.922: “ISDN Data Link Layer Specifications for Frame Relay”*, Jan. 1991.
- [3] W. Stallings. *“Advances in ISDN and Broadband ISDN”*. Los Alamitos, CA: IEEE Computer Society Press, 1992.
- [4] R. Pretty and M. Morin. “Frame Relay Interworking with Asynchronous Transfer Mode”. In *Proc. IEEE Global Telecommun. Conf., Globecom '93*, pages 1854–1860, Houston, TX, Nov. 1993.
- [5] G. C. Kessler. *“ISDN: Concepts, Facilities and Services”*. McGraw-Hill, Inc, 1993.
- [6] P. Brackett, R. Damodaram, N. Ransom, and D. Spears. “Interworking Broadband Networks and Services Using ATM”. In *Proc. IEEE Global Telecommun. Conf., Globecom '93*, pages 1849–1853, Houston, TX, Nov. 1993.

- [7] International Telegraph and Telephone Consultative Committee. *CCITT Draft Recommendation I.555: "Frame Relaying Bearer Service Interworking"*, Jan. 1993. Temporary Document 18 (XVIII).
- [8] International Telegraph and Telephone Consultative Committee. *CCITT Recommendation I.365.1: "Frame Relaying Service Specific Convergence Sublayer (FR-SSCS)"*, Jan. 1993. Temporary Document 11 (XVIII).
- [9] International Telegraph and Telephone Consultative Committee. *CCITT Recommendation I.363, Section 6: "ATM AAL Type 5"*, Jan. 1993. Temporary Document 10 (XVIII).
- [10] International Telegraph and Telephone Consultative Committee. *CCITT Draft Recommendation G.764, "Voice Packetization - Packetized Voice Protocols"*, Aug. 1990.
- [11] American National Standards Institute, Inc. *"Integrated Services Digital Network (ISDN) Core Aspects of Frame Protocol" for use with Frame Relay Bearer Service*, Jun. 1991. ANSI T1.618.
- [12] A.R. Kaye, L. Dong, and S.A. Mahmoud. "Adaptive Buildout Delay Mechanism for Transmission of Compressed Voice over Integrated Services Frame Relay Networks". Technical Report SCE-94-25, Dep. of Systems and Computer Engineering, Carleton University, Ottawa, Canada, Sep. 1994.
- [13] L. Dong, A.R. Kaye, and S.A. Mahmoud. "Priority Transmission of Compressed Voice over Integrated Services Frame Relay Networks". Technical Report SCE-

94-26, Dep. of Systems and Computer Engineering, Carleton University, Ottawa, Canada, Sep. 1994.

- [14] L. Dong, A.R. Kaye, and S.A. Mahmoud. "Transmission of Compressed Voice over Integrated Services Frame Relay Networks: Priority Service and Adaptive Buildout Delay". *IEE Proc. Commun.*, 141(4):265-274, Aug. 1994.
- [15] L. Dong. "Compressed Voice in Integrated Services Frame Relay Networks". Master's thesis, Carleton University, Ottawa, Canada, Dec. 1992.
- [16] A.R. Kaye and S. Zhang. "Congestion Control in Integrated Voice-Data Frame Relay Networks and the Case for Embedded Coding". In *Proc. Globecom*, 1994. To be published.
- [17] S. Zhang. "Congestion Control in Integrated Service Frame Relay Networks with Compressed Voice and Data Traffic". Master's thesis, Carleton University, Ottawa, Canada, Jul. 1993.
- [18] International Telegraph and Telephone Consultative Committee. *CCITT Recommendation H.261: "Video Codec for Audiovisual Services at $p \times 64$ kbits/sec"*, Jul. 1990. WPXV/1.
- [19] M. Ghanbari and C. Hughes. "Packing Coded Video Signals into ATM Cells". *IEEE Trans. Net.*, Vol. 1(5):505-509, Oct. 1993.
- [20] T. Turetti. "H.261 Software Codec for Videoconferencing Over the Internet". Technical Report 1834, Institut National de Recherche en Informatique et en Automatique, France, 1993.

- [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] J.C. Bolot and T. Turletti. "A Rate Control Mechanism for Packet Video in the Internet". In *Proc. IEEE Conf. Comp. Commun., Infocom '94*, pages 1216–1223, Toronto, ON, Canada, 1994.
- [23] B. Melamed and B. Sengupta. "TES Modeling of Video Traffic". *IEICE Trans. Commun*, E75-B(12):1292–1300, 1992.
- [24] 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.
- [25] B. Melamed and D. Geist. "TESTool: An Environment for Visual Interactive Modeling of Autocorrelated Traffic". In *Proc. IEEE Intern. Conf. Commun., ICC '92*, pages 1285–1289, 1992.
- [26] P. Ang, P. Ruetz, and D. Auld. "Video Compression Makes Big Gains". *IEEE Spectrum*, pages 16–19, Oct. 1991.
- [27] J.B. O'Neal. "Entropy Coding in Speech and Television PCM Systems". *IEEE Trans. Info. Theory*, IT-17:758–761, 1971.
- [28] American National Standards Institute, Inc. "*Integrated Services Digital Network (ISDN) - Architectural Framework and Service Description for Frame-Relaying Bearer Service*", Oct. 1990. T1.606.

- [29] Daniel Minoli. *"Enterprise Networking"*. Boston: Artech House, 1993.
- [30] P. Jelenkovic and B. Melamed. "Automated TES Modeling of Compressed Video". Preprint, 1994.
- [31] Jerry Banks and John S. Carson II. *"Discrete-Event System Simulation"*. Prentice-Hall, Inc, 1984. pages 294-312.
- [32] Newbridge Networks Corporation, Kanata, Ontario, Canada. *"Newbridge Frame Relay Interface Specification"*, 1991. version 1.0.
- [33] H. Chin, J. Goodge, R. Griffiths, and D. Parish. "Performance of Video Signals for Viewphone-Type Pictures". *IEEE Jour. Sel. Areas in Commun.*, Vol. 7(5):826-832, Jun. 1989.
- [34] Alberto Leon-Garcia. *"Probability and Random Processes for Electrical Engineering"*. New York: Addison-Wesley, 1989.
- [35] U. Grenander and M. Rosenblatt. *"Statistical Analysis of Stationary Time Series"*. Wiley, 1957.
- [36] Jean Walrand. *"Communication Networks: A First Course"*. Boston: Richard D. Irwin Inc., 1991.
- [37] 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.
- [38] A. Law and D. Kelton. *"Simulation Modeling and Analysis"*. New York: McGraw-Hill, Inc, 1991.

- [39] R. Nobakht, D. Van Den Bout, K. Townsend, and S. Ardalan. "Optimization of Transmitter and Receiver Filters for Digital Communication Systems Using Mean Field" Annealing. *IEEE Jour. Sel. Areas in Commun.*, Vol. 8(8):1472–1480, Oct. 1990.
- [40] G. Bilbro, R. Mann, T. Miller III, W. Snyder, D. E. Van den Bout and M. W. White. "Mean Field Annealing and Neural Networks". In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, San Mateo, CA, 1989.
- [41] M. Devetsikiotis and K. Townsend. "Statistical Optimization of Dynamic Importance Sampling Parameters for Efficient Simulation of Communication Networks". *IEEE ACM. Trans. Net.*, Vol. 1(3):293–305, Jun. 1993.
- [42] N. Ohta. *"Packet Video Modeling and Signal Processing"*. Boston: Artech House, 1994.

END

31-05-95

FIN