PERFORMANCE EVALUATION AND APPLICATIONS OF ATM NETWORKS

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

PERFORMANCE EVALUATION AND APPLICATIONS OF ATM NETWORKS

edited by

Demetres Kouvatsos
University of Bradford, United Kingdom

KLUWER ACADEMIC PUBLISHERS
New York, Boston, Dordrecht, London, Moscow

eBook ISBN: 0-306-47023-3
Print ISBN: 0-792-37851-2

©2002 Kluwer Academic Publishers
New York, Boston, Dordrecht, London, Moscow

Print ©2000 Kluwer Academic Publishers

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

To Mihalis and Maria

CONTENTS

Preface

Information Highways are widely considered as the next generation of high speed communication systems. These highways will be based on emerging Broadband Integrated Services Digital Networks (B-ISDN), which - at least in principle - are envisioned to support not only all the kinds of networking applications known today but also future applications which are not as yet understood fully or even anticipated. Thus, B-ISDNs release networking processes from the limitations which the communications medium has imposed historically. The operational generality stems from the versatility of Asynchronous Transfer Mode (ATM) which is the transfer mode adopted by ITU-T for broadband public ISDN as well as wide area private ISDN. A transfer mode which provides the transmission, multiplexing and switching core that lies at the foundations of a communication network.

ATM is designed to integrate existing and future voice, audio, image and data services. Moreover, ATM aims to minimise the complexity of switching and buffer management, to optimise intermediate node processing and buffering and to bound transmission delays. These design objectives are met at high transmission speeds by keeping the basic unit of ATM transmission - the ATM cell - short and of fixed length. However, to support such diverse range of services on one integrated communication platform, it is necessary to provide a most careful network engineering in order to achieve a fruitful balance amongst the conflicting requirements of different quality of service constraints, ensuring one service does not have adverse implications on another. Thus, performance evaluation and quantitative analysis of ATM networks are of extreme importance to both users and operators.

Experimental ATM networks have now been established worldwide, based on commercially available ATM products and switch architectures. Although the suitability and cost effectiveness of ATM to provide the B-ISDN core has been the subject of public debate, the authoritative endorsement of ATM by ITU-T and the subsequent investments in commercial ATM technology ensure that ATM will - in all likelihood - hold a place of prominence in the world of communications well into the new millennium!

Performance modelling, evaluation and prediction of ATM networks are very important in view of their ever expanding usage and the multiplicity of their component parts together with the complexity of their functioning. Over the recent years a considerable amount of effort has been devoted, both in industry and academia, towards the performance analysis of ATM networks. However, there is still a set of many interesting and important performance related research problems to be addressed and resolved before a global integrated broadband network infrastructure can be established. This includes traffic modelling and characterisation, flow and congestion control, routing and optimisation, ATM switch architectures and internetworking, IP/ATM networks integration, resource allocation and the provision of specified quality of service. Thus, it seems most essential both to comprehend recent advances made in the field and also to search for new evaluation techniques and tools for the performance optimisation of these future high speed networks.

The principal objective of the tutorial book 'Performance Evaluation and Applications of ATM Networks' is to present an overview of recent results, applications, future directions and comprehensive bibliographies relating to the fundamental performance evaluation and application issues of ATM networks. The book maintains an effective balance between descriptive and quantitative approaches towards the presentation of important ATM mechanisms and associated performance modelling techniques and applications. Moreover, it offers a fundamental source of reference on ATM networks' performance within both academic and industrial environments.

The book includes 17 tutorial papers by eminent researchers and practitioners in the field from industry and academia worldwide. All papers are invited works which were evaluated and selected, subject to rigorous international peer review. The tutorial papers can be used as essential introductory state-of-the-art material for both education and further research in the performance modelling and analysis field of ATM networks. In particular the tutorial book aims to unify ATM performance modelling material already known but dispersed in the literature, introduce readers to unfamiliar and unexplored ATM performance research areas and, generally, illustrate the diversity of research found in the ATM field of high growth.

The tutorial papers are broadly classified into six parts covering the following topics:

Part One ATM Traffic Modelling and Characterisation

Part Two ATM Traffic Management and Control

Part Three ATM Routing and Network Resilience

Part Four IP/ATM Networks Integration

Part Five ATM Special Topics: Optical, Wireless and Satellite Networks

Part Six Analytical Techniques for ATM Networks

An overview of the proposed tutorial papers of the book is presented below:

Part One on "ATM Traffic Modelling and Characterisation" includes three tutorial papers and is concerned with modelling, characterisation and performance implications of multiplexed streams of bursty and correlated ATM traffic in ATM networks. The first paper by John Cosmas (Brunel University, UK) on ‘Stochastic Source Models and Applications to ATM’ describes the theory of the relationships between the main statistical and model parameters of voice, data and video sources and how they relate to Usage Parameter Control (UPC) mechanisms in ATM networks. The second paper by Mark Bromirski and Wieslaw Lobejko (Military Communication Institute, Poland) on ’Fractals and Chaos for Modelling Multimedia ATM Traffic’ explores the fractal and chaotic properties of multimedia ATM traffic and their performance impact. The third paper by Timothy X. Brown (University of Colorado, USA) on ’Adaptive Statistical Multiplexing for Broadband Communication’ focuses on the ststistical multiplexing of traffic sources and reviews adaptive multiplexing in terms of statistical-classification-based decision functions and their applications.

Part Two on "ATM Traffic Management and Control" brings together five tutorial papers addressing fundamental objectives such as guaranteed network performance, traffic control and congestion schemes, traffic management and contracted quality-of-service (QoS).

The first paper by Chris Blondia (University of Antwerp, Belgium) and Olga Casals (Polytechnic University of Catalunia, Spain) on 'Traffic Management in ATM Networks: An Overview' provides a comprehensive overview of traffic service categories and transfer capabilities for ATM traffic managements together with some essential control and congestion schemes in ATM networks. The second paper by Khaled M. Fuad Elsayed (Cairo University Egypt) and Harry G. Perros (North Carolina State University, USA) on 'A Comparative Performance Analysis of Call Admission Control Schemes in ATM Networks' carries out a comparative study of the performance analysis of Call Admission Control (CAC) mechanisms devised to meet certain QoS requirements expressed in terms of cell loss probability and maximum delay. The third paper by Nikolas Mitrou (National Technical University of Athens, Greece) on 'Traffic Control in ATM: A Review, an Engineer's Critical View and a Novel Approach' reviews the main ATM control functions and describes an alternative approach to the traffic control problem, based on burst-level modelling. The latter explores the buffering gain and proposes the use of the M/D/1 model as a unified tool for engineering all necessary control mechanisms. The fourth paper by Gunnar Karlsson (Swidish Institute of Computer Science, Sweden) on 'Video over ATM Networks' is concerned with quality requirements posed on network transfers of video information and presents a review of video communication over ATM networks which includes source coding, bit rate regulation and quality constraints. The fifth paper by Michael Logothetis (University of Patras, Greece) on 'Optimal Resource Management in ATM Networks based on Virtual Path Bandwidth Control' discusses the impact of the optimal call-level virtual path bandwith (VPB) control towards the analytic minimisation of the worst call blocking probability of all virtual paths (VPs) of an ATM network.

Part Three on "ATM Routing" consists of two tutorial papers addressing inherent routing problems frequently encountered during the design and management of complex multiservice ATM networks involving information transfer from one to one or one to many recipients for multimedia applications. The first paper by John Crawford and Gill Waters (University of Kent at Canterbury, UK) on 'ATM Multicast Routing' reviews heuristics for multicast routing which support multimedia services in high speed networks such as B-ISDNs based on ATM, by minimising the multicast tree cost whilst maintaining a bound on delay. Relative performance comparisons involving different multicast heuristics are carried out and recommendations are made towards efficient solutions for a wide range of flat and hierarchical networks. The second paper by Paul Veitch (BT Labs., UK) on ‘Embedding Resilience in Core ATM Networks’ deals with the embedding of resilience mechanisms in core ATM network elements in order to provide restoration mechanisms and, thus, mitigate the impact of outages caused by cable breaks and node failures.

Part Four on "IP/ATM Networks Integration" includes a single tutorial paper by Andreas Skliros on ’IP Switching over ATM Networks’. The paper addresses performance and reliability problems associated with the unprecedented growth of IP traffic and reviews various approaches for integrating the flexibility of IP software with the high transmission speed and QoS guarantees of ATM networks. Particular emphasis is given on the new cost-effective IP switching architecture, its functionality and the management of QoS issues.

Part Five on "ATM Special Topics: Optical, Wireles and Satellite Networks" presents three tutorial papers dealing with some contemporary topics in the ATM field. The first paper on Maurice Gagnaire and Saso Stojanovski (ENST, France) on ’An Approach for Traffic Management over G.983 ATM-based Passive Optical Networks’ focuses on a new generation of access networks aiming to provide end-to-end broadband services. The state of the art in this field is presented by addressing both feeder networks and access networks with particular reference to ATM traffic management over passive optical networks. The second paper by Renato Lo Cigno (Politecnico di Torino, Italy) on ’Wireless ATM: An Introduction and Performance Issues’ reports an overview of the main characteristics of wireless ATM networks with radio access, network architecture and management. Moreover, performance issues and application areas are identified together with MAC protocols, handover implementation procedures and experimental projects. The fourth paper by Zhili Sun (Surrey University, UK) on ‘Satelite ATM Networks’ presents an overview of the major issues and recent developments of satellite systems for ATM networks (and broadband communication) including ATM satellite system structure and architecture, management and control over satellite, performance aspects of ATM over satellite, satellite bandwidth resource management, multimedia applications including current projects and future research issues on satellite constellations and convergence of ATM and Internet.

Part Six on "Analytical Techniques for ATM Networks" presents three tutorial papers reviewing exact and approximate analytic methodologies for the performance modelling, evaluation and prediction of ATM switching nodes and networks involving multistreams of bursty and /or correlated traffic under different buffer management policies. The first paper by Gilberto Mayor and John Silvester (University of Southern California, USA) on Performance Modelling and Network Management for Self-Similar Traffic' highlights the long-range dependence phenomenon exhibited by real network traffic and provides an overview of self-similar traffic models, based on a fractional Brownian motion envelope process. Moreover, analytical tools capable of computing bandwidth and buffer requirements in ATM are included, driven by aggregate, heterogeneous and self-similar processes. The second paper by Sabine Wittevrongel and Herwig Bruneel (University of Ghent, Belgium) on 'Discrete-Time ATM Queues with Independent and Correlated Arrival Streams' presents analytical techniques for the solution of discrete-time queueing models of ATM multiplexers and switching elements with either independent or correlated arrival streams and dedicated-buffer output queueing schemes. The Third paper by Demetres Kouvatsos (Bradford University, UK) on ' An Information Theoretic Methodology for Queueing Network Models (QNMs) of ATM Switch Architectures' reviews an information theoretic methodology for the credible and cost-effective approximate analysis of queueing models of some ATM switches and networks with short range dependence (SRD) correlated traffic streams and either cell-blocking or cell-loss, as appropriate. The methodology has its roots on the information theoretic principle of maximum entropy (ME) and implies a decomposition of the queueing network into individual finite capacity queues each of which can be solved in isolation.

Some of these papers are based on tutorial themes presented during the recent series of the International Federation of Information Processing (IFIP) Workshops on the 'Performance Modelling and Evaluation of ATM Networks' which were organised by Bradford University at Ilkley, West Yorkshire, England, UK and generated enormous international support from both industry and academia. I, therefore, wish to end this foreword by expressing my thanks to the IFIP TC6 on Communication Systems and all other supporting organisations, such as the Performance Engineering Groups of the British Computer Society (BCS) and British Telecom (BT). My thanks are also extended to all international referees for their invaluable and timely reviews of the tutorial papers and to Melissa Fearon, Kluwer Academic Publishers, U.S.A. for her technical advice and kind collaboration towards the preparation of the entire book.

Demetres Kouvatsos

Participants in the Review Process

Irfan Awan
Riaz Ahmad
Marco Ajmone-Marsan
Åke Arvidsson
Frank Ball
Monique Becker
Alexandre Brandwajn
Mark Bromirski
Chris Blondia
Timothy X Brown
Herwig Bruneel
Olga Casals
Tadeusz Czachorski
Marco Conti
Laune G. Cuthbert
Tien V. Do
Serge Fdida
Rod J. Fretwell
Maurice Gagnaire
Pawel Gburzynski
Erol Gelenbe
Nicolas Georganas
John M. Griffiths
Peter Harrison
Boudewijn Haverkort
Gérard Hébuterne
Christoph Herrmann
Frank Hübner-Szabo de Bucs
Ilias Iliadis
László Jereb
Mourad Kara
Gunnar Karlsson
Johan Karlsson
Ernest Koenigsberg
Demetres Kouvatsos
Hayri Korezlioglu
Koenraad Laevens
Renato Lo Cigno
Michael Logothetis
Xiaowen Mang
Brian G. Marchent
Phil Mars
Saverio Mascolo
John Mellor
Isi Mitrani
Nikos M. Mitrou
Sandor Molnár
Jogesh K. Muppala
Arne Nilsson
Raif Onvural
Rubem Pereira
Harry G. Perros
Guido Henri M. Petit
Michal Pióro
Guy Pujolle
Martyn J. Riley
Charalambos Skianis
Maria Simon
Geoff Smith
Andreas Skliros
Maciez Stasiak
Ioannis Stavrakakis
Yutaka Takahashi
Don Towsley
Paul A. Veitch
Sabine Wittevrongel
Michael E. Woodward
Kristiaan Wuyts
Hideaki Yamashita
Sufian Yousef
Yury Zlotnikov

Chapter 1

STOCHASTIC SOURCE MODELS AND APPLICATIONS TO ATM

John.P. Cosmas
Department ofElectronic and Computer Engineering, Brunel University, Uxbridge, England
john.cosmas@brunel.ac.uk

Abstract: The subject of this paper is the theory of the relationships between the main statistical parameters of voice, data and video sources. Examples are given throughout to illustrate how the source models can be parameterised and used. The mathematics is kept as simple and self-explanatory as possible.

Keywords: ATM Source Models

1. INTRODUCTION

Variable bit rate (VBR) voice, data and video coding schemes compress information more efficiently than constant bit rate (CBR) coding schemes and may be connected to either Synchronous Transfer Mode (STM) or ATM networks via a transmission link.

In STM networks the communications resources, in the form of circuits, are shared among users. Each user has sole access to a circuit during the use of the network which in the telephone network has a fixed capacity of 64 kbit/s. If more capacity is required then more circuits are allocated to the user. Since the capacity is solely allocated to a user whether or not it is required, constant bit rate (CBR) encoding techniques are used for information compression. Coding to compress information by removing redundancies is required so that a minimum amount of network resources (circuits) are used. However variable bit rate (VBR) coding schemes often compress information more efficiently. For VBR coders, if the output capacity of the source exceeds that allocated to it then some information is not able to be transmitted and is lost. However if the output capacity of the source is less than that allocated to it, then there is under utilisation of network resources. Therefore a switching technique is required which can efficiently carry information from highly coded variable bit rate sources without any significant loss of information. This technique is ATM.

In ATM networks, information from a source is broken up into short units called cells, which are transmitted individually through the network. The main benefit of cell switching is 'statistical multiplexing' which is the simultaneous use of the same communications circuits by a large number of sources on a demand basis. If there is a simultaneous requirement of a communications link by two cells each from different sources then there is queuing at the network nodes where a cell from one source waits until a cell from another source has been transmitted. This queuing at the network nodes is also designed to absorb cells from VBR coders thus making cell switching more suitable to efficiently coded VBR coders. The main dilemma is that the variable bit rate of voice, data and video codecs is dependent on the incidental nature of voice, data and video sequences that have yet to occur. For example, VBR Digital Speech Interpolation (DSI) voice codecs for normal conversation are known to have exponentially distributed mean talk duration of 3 seconds and silent periods of 7.5 seconds and can be negotiated as such with the ATM network. However if for some unpredictable reason there ensues a "heated" conversation with mean talk duration of 5 seconds and silent periods of 4 seconds then more cells will be generated than was originally negotiated and cells may require to be removed.

Therefore variable bit rate audio-visual terminals, which are interfaced to ATM networks will be required to decide on their traffic characteristics prior to call set up. This will require a terminal to measure its cell arrival statistics, decide on a traffic model and parameterise the traffic model from its cell arrival statistics. The traffic model and its parameters will be used to obtain connection acceptance from the ATM. Once a connection has been accepted the ATM network using a leaky bucket will police the cell arrivals. This paper gives a review of the basic principles of the most important source models for ATM. A review of source models and their statistical parameters for ATM is presented in [Cosm94]. Descriptions and mathematics of the source models can be found in [Klei75] [Krey70] [Cox87] [Mag88].

2. LEVELS OF RESOLUTIONS IN TIME

The traffic destined for transport using cells in an ATM network may show behaviour, which can be characterised by up to five resolutions in time: 1) Calendar level; 2) Connection level; 3) Dialogue level (voice and data sources); 4) Burst level; 5) Cell level.

The calendar level describes the daily, weekly and seasonal traffic variations of a traffic source. The connection level describes the behaviour of a traffic source on a virtual connection basis. The connection set-up and clear events, which delimit the connection duration, are the most macroscopic behaviour of a stationary traffic source. The duration of a connection is typically in the range of 100 ... 1000s, depending on the service. The dialogue level describes the interaction between voice or data agents at both ends of the connection. In principle four situations are possible: silence, A-subscriber transmitting, B-subscriber transmitting, both subscribers transmitting, so that the interaction can be modelled by a four state Finite State Machine. Typical duration of a transmission in the case of telephony is in the range of 10 seconds. In the case of unidirectional services e.g. file transfers, there is no dialogue level. The burst level describes the statistical behaviour of an active (transmitting) partner. For a telephone service the on-off characteristics of the cell generation process is modelled in this level. Duration of the on-time and the off-time are in the range of 0.1.. few seconds (voice transmission). During a burst interval the cell arrivals are approximated to the mean rate rather than a probability distribution. For a distributive video service the interscene change statistics is modelled in this level. Scene changes are defined to have occurred if there has been a physical operation on the camera (a positive or negative zoom or pan) or on the film (a cut of the film from one scene to another). Typical duration of a scene is in the range of 10..20 secs [Cosm94]. Distributive video exhibits burst level behaviour because it consists of a sequence of scenes each with their own inter-scene activity that can be considerably different from each other. During burst intervals the cell arrivals are approximated to a mean rate rather than a probability distribution. This is not the case for interactive video because there is only one class of scene (head and shoulders) with no physical operation on the camera or film. The cell level describes the behaviour of cell generation at the lowest level. From the (maximum) bitrate of the service and the length of a cell the (minimum) distance between cells can be derived. For a 622 Mbit/s link speed and a 48+5 bytes cell size, the corresponding time scale in the cell level is 0.6817 μs i.e. it corresponds to the minimum interval between two consecutive cells.

3. STATISTICAL PARAMETERS OF A DISCRETE RANDOM VARIABLE

3.1 DISCRETE RANDOM VARIABLE, PROBABILITY DISTRIBUTION AND PROBABILITY DISTRIBUTION FUNCTION

A random variable X is called a discrete random variable if X can assume only a countable number of values $\{x_1, x_2, x_3,....\}$. The complete set of probabilities $P\ [x_i]$ associated with the possible values of xi's of X is called the probability distribution of a discrete random variable X. The probability distribution and probability distribution function are shown in figure 1.1 and are related as:

$$F_X(x) = \sum_{x_i \le x} P[x_i]$$

where

$$F_X(\infty) = \sum_{x_i \le \infty} P[x_i] = 1$$

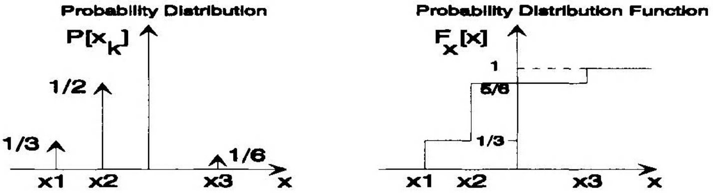


Figure 1.1. Probability Distribution and Probability Distribution Function of a Discrete Random Variable

3.2 EXPECTATION AND MEAN OF A DISCRETE RANDOM VARIABLE

The sample mean of a discrete random variable X where X_k (k=1..N) denotes the outcome of the kth sample is:

$$\overline{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$$

Let $P[x_i]$ denote the probability that the result is the outcome x_i. Then $E[X]$ the expectation or mean.

$$E[X] = \sum_{i=1}^{\infty} x_i P[x_i] = \mu_x$$

3.3 MOMENTS OF A DISCRETE RANDOM VARIABLE

If X is a discrete random variable, so is its nth power X^n. The sample n^{th} moment of X:

$$m_n = \frac{1}{N}\sum_{k=1}^{N} X_k^n$$

The n^{th} moment of X :

$$E[X^n] = \sum_{i=1}^{\infty} x_i^n P[x_i]$$

The sample n^{th} central moment of X :

$$s^n = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_k - \overline{X}\right)^n$$

The n^{th} central moment of X:

$$E\left[\left(X-\mu_x\right)^n\right] = \sum_{i=1}^{\infty}\left(x_i - \mu_X\right)^n P\left[x_i\right]$$

The second central moment is given the special name variance σ_X^2. It is the mean of the squared deviations (dispersion) of a random variable about its mean. The sample variance is given as:

$$s^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_k - \overline{X}\right)^2$$

The variance is given as:

$$E\left[\left(X-\mu_x\right)^2\right] = \sum_{i=1}^{\infty}\left(x_i - \mu_X\right)^2 P\left[x_i\right] = \sigma_x^2$$

The variance can also be given as :

$$E\left[\left(X-\mu_X\right)^2\right] = E\left[X^2\right] - E\left[X\right]^2$$

Since $E[X+Y] = E[X] + E[Y]$ and $E[kX] = kE[X]$

In order to have a measure of dispersion, which has the same dimensions as the random variable the square root of the variance is computed and is known as the standard deviation σ_X.

The third central moment is a measure of asymmetry of the random variable about its mean. If $E\left[(X-\mu_x)^3\right] = 0$ then the distribution of the random variable X about the mean is symmetric. If $E\left[(X-\mu_x)^3\right] > 0$ then the distribution of the random variable X about the mean has a longer tail on the positive side of the mean. If $E\left[(X-\mu_x)^3\right] < 0$ then the distribution of the random variable X about the mean has a longer tail on the negative side of the mean. The fourth central moment is sometimes used as a measure of peakedness of a distribution.

3.4 AUTOCORRELATION OF A DISCRETE RANDOM VARIABLE

To compute the correlation between two random variables is to measure the degree to which those two random variables are similar. Crosscorrelation is a measure of similarity between two different random variables X and Y whereas autocorrelation is when the measure of similarity of one random variable X with itself after a period of time τ, $X^{(\tau)}$.

The sample autocorrelation is given as:

$$m(X, X^{(\tau)}) = \frac{1}{N-\tau}\sum_{k=1}^{N-\tau} X_k X_{k+\tau}$$

The autocorrelation is given as:

$$R\left[X,X^{(\tau)}\right]=E\left[XX^{(\tau)}\right]=\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}x_i x_j P[x_i]P\left[x_j|x_i\right]^{(\tau)}$$

Where $P[x_j|x_i]^{(\tau)}$ is the probability of obtaining the outcome x_j given the outcome x_i, τ intervals after the outcome x_i.

The autocovariance $C[X,X^{(\tau)}]$ of a random variable X is the joint second central moment of the random variable X and $X^{(\tau)}$. The sample autocovariance is given as:

$$s\left(X,X^{(\tau)}\right)=\frac{1}{N-\tau}\sum_{k=1}^{N-\tau}\left(X_k-\overline{X}\right)\left(X_{k+\tau}-\overline{X}\right)$$

The autocovariance is given as:

$$C\left[X,X^{\tau}\right]=E\left[(X-\mu_x)\left(X^{(\tau)}-\mu_x\right)\right]=\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}(x_i-\mu_x)(x_j-\mu_x)P[x_i]P\left[x_j|x_i\right]^{(\tau)}$$

The autocovariance can also be given as:

$$E\left[(X-\mu_x)\left(X^{(\tau)}-\mu_x\right)\right]=E\left[X,X^{(\tau)}\right]-E[X]^2$$

The normalised autocovariance is given as

$$C\left[X,X^{\tau}\right]/C\left[X,X^{0}\right]$$

4. TYPES OF PROCESSES

4.1 DETERMINISTIC PROCESS

A deterministic process is one in which there is a constant outcome. A continuous bit rate source (CBR) is an example of a deterministic process.

4.2 BERNOULLI PROCESS

A Bernoulli process is the random counting process, which results from a Bernoulli experiment. There are one of two possible outcomes in a Bernoulli experiment: success or failure (corresponding to packet or no packet, cell or no cell, frame or no frame). A sequence of Bernoulli trials occurs when a Bernoulli experiment is performed several independent times so that the probability of success p remains the same from trial to trial. If the probability of success is p then the probability of failure is $(1\text{-}p) = q$. Let X be a random variable associated with a Bernoulli trial. Then

$$X_{\text{(success)}}=1 \text{ and } X_{\text{(failure)}}=0$$

The probability distribution function is written as

$$P[X] = p^x(1-p)^{1-x} \ldots\ldots\ldots x = 0,1$$
$$E[X] = \sum_{x=0}^{1} xp^x(1-p)^{1-x} = p$$
$$\sigma_X^2 = \sum_{x=0}^{1} (x-\mu_x)^2 p^x(1-p)^{1-x}$$
$$\Rightarrow \sigma_X^2 = p^2(1-p)+(1-p)^2 p = p(1-p)$$

The Bernoulli process has an autocovariance $C[X, X^\tau] = \sigma_X^2.\delta(\tau)$ (where $\delta(\tau)$ is the Kronecker delta function) because any sequence of Bernoulli trials is independent of each other. The Poisson process is the continuous time version of the discrete time Bernoulli Process.

The Bernoulli process can also be viewed as a time series, which has a geometrically distributed interarrival time α:

$$P[\alpha = \mathrm{n}] = p(1-p)^{n-1}$$

The mean interarrival time $E[\alpha]$ is thus:

$$E[\alpha] = \sum_{n=1}^{\infty} np(1-p)^{n-1} = (1-q)\sum_{n=1}^{\infty} nq^{n-1} = (1-q)(1+2q+3q^2+4q^3+\ldots..)$$
$$\Rightarrow E[\alpha] = (1+q+q^2+q^3+\ldots..) = \sum_{n=0}^{\infty} q^n = \frac{1}{(1-q)} = \frac{1}{p}$$

Since if $0 \leq q < 1$, then:

$$\sum_{n=0}^{\infty} q^n = \frac{1}{(1-q)}$$

The interarrival time variance σ_α^2 is thus:

$$E[\alpha^2] = \sum_{n=1}^{\infty} n^2 p(1-p)^{n-1} = (1-q)\sum_{n=1}^{\infty} n^2 q^{n-1} = (1-q)(1+4q+9q^2+16q^3+25q^4+\ldots)$$
$$\Rightarrow E[\alpha^2] = (1+3q+5q^2+7q^3+9q^4+\ldots) = (1+q)(1+2q+3q^2+4q^3+5q^4+.$$
$$\Rightarrow E[\alpha^2] = (1+q)(1+q+q^2+q^3+q^4+\ldots)(1+q+q^2+q^3+q^4+\ldots) = \frac{(1+q)}{(1-q)^2}$$
$$\Rightarrow \sigma_\alpha^2 = E[\alpha^2] - E[\alpha]^2 = \frac{(1+q)}{(1-q)^2} - \frac{1}{(1-q)^2} = \frac{1-p}{p^2}$$

4.3 PROBABILITY GENERATING FUNCTION

The probability generating function of a discrete random variable X is given as:

$$G(z) = \sum_{k=0}^{\infty} P[x_k] z^{x_k}$$

By differentiating and setting z=l, the first moment is obtained:

$$\frac{dG(z)}{dz}=\sum_{k=0}^{\infty}x_k P[x_k]z^{x_k-1}$$

$$\Rightarrow \left.\frac{dG(z)}{dz}\right|_{z=1}=\sum_{k=0}^{\infty}x_k P[x_k]=E[X]$$

By differentiating again and setting z=1, the first and second moment are obtained:

$$\frac{d^2G(z)}{dz^2}=\sum_{k=0}^{\infty}x_k(x_k-1)P[x_k]z^{x_k-2}$$

$$\Rightarrow \left.\frac{d^2G(z)}{dz^2}\right|_{z=1}=\sum_{k=0}^{\infty}x_k(x_k-1)P[x_k]=E[X^2]-E[X]$$

4.4 MARKOV PROCESS

A random sequence is said to be a Markov process if for every time t and all possible states the probability of any future state given the entire past and present states is independent of the past states and only dependent on the present state of the process.

4.4.1 Discrete Time Discrete 2-State Markov Process

This is a Markov model that alternates between two states 0 and 1 as shown in figure 1.2. The packet arrival process in states 0 and 1 can be deterministic, Bernoulli or any other type of stochastic process. It can be a time series (denoted by the packet interarrival time) or a counting process (denoted by the number of packet arrivals in an interval T). Assume that the arrival process is a deterministic, counting process with parameters A_0 or A_1.

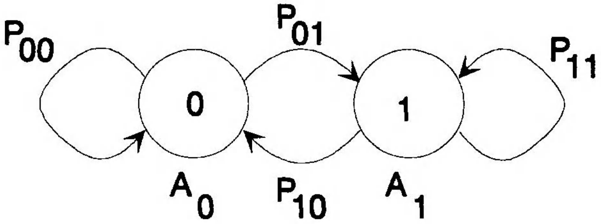


Figure 1.2. A 2-state Discrete Time Discrete State Markov Process

The transition probability matrix $[P]$ is given as

$$[P]=\begin{bmatrix}P_{00} & P_{01}\\ P_{10} & P_{11}\end{bmatrix}$$

Where P_{ij} denotes the probability of transition from state i to state j.
Let $s_i^{(n)}$ denote the probability of finding the system in state i at time n.
Then:

$$s_0^{(n)} = s_0^{(n-1)}p_{00} + s_1^{(n-1)}p_{10}$$
$$s_1^{(n)} = s_0^{(n-1)}p_{01} + s_1^{(n-1)}p_{11}$$

In matrix form:

$$[s^{(n)}] = [s^{(n-1)}][P]$$

On iteration:

$$[s^{(n)}] = [s^{(n-1)}][P] = [s^{(n-2)}][P]^2 = \ldots = [s^{(0)}][P]^n$$

Thus given the initial conditions $[s^{(0)}]$ and the matrix of transition probabilities $[P]$ we can find the state occupation probabilities at any time n.. After a sufficiently large number of iterations the system settles down to a condition of statistical equilibrium in which the state occupation probabilities are independent of initial conditions. Thus as $n \to \infty$ then $[s^{(n)}]=[\pi]$ where $[\pi]=[\pi_0 \;\; \pi_1]$ is the equilibrium probability distribution.

$$[\pi] = [\pi][P]$$

Given that $\pi_0 + \pi_1 = 1$, we can solve for π_0 and π_1.

$$\pi_0 = P_{10}/(P_{01} + P_{10})$$
$$\pi_1 = P_{01}/(P_{10} + P_{01})$$

Therefore:

$$E[X] = \sum_{i=0}^{1} A_i \pi_i$$

$$E[X^2] = \sum_{i=0}^{1} A_i^2 \pi_i$$

$$R[XX^{(\tau)}] = \sum_{i=0}^{1} \sum_{j=0}^{1} A_i A_j \pi_i P[x_j | x_i]^{(\tau)}$$

Where $P[x_j | x_i]^{(\tau)}$ is the probability of being in state j given state i after a lag of τ samples.

4.4.2 General Modulated Deterministic Process

The General Modulated Deterministic Process (GMDP) [Cosm94] is based on a finite state machine having N states (an example for $N=3$ is shown in figure 1.3).

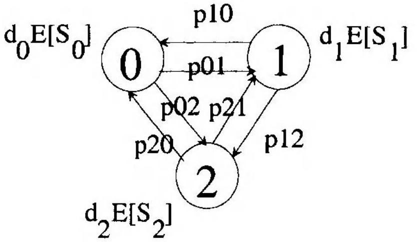


Figure 1.3. 3-state General Modulated Deterministic Process

In each state, cells are generated with constant interarrival time d_i, where the index i identifies the state. The number of cells (X_i) which are emitted in state i consecutively may have a general discrete distribution $f_i(k) = P[X_i = k]$. In general, the GMDP includes also silence states where no cells are generated, and the duration of these states may also have a general discrete distribution. The state changes of the underlying state machine are governed by a *NxN* transition matrix $P = (p_{ij})$, where p_{ij} is the probability that at the end of its sojourn time in state i the source moves to state j, i <> j.. For case studies a special case of the GMDP has been used, where X_i has a geometric distribution with a minimum of 1 (cell). In this case the process is called Markov Modulated Deterministic Process (MMDP), since the underlying state machine can now be described as a discrete-time discrete-state Markov Chain.

The transition matrix $[p]$ of the MMDP is closely related to the transition matrix $[P]$ of the Discrete Time Discrete State Markov Model. The main difference between both of the descriptions is the mechanism to generate arrivals: counts or intervals. For a three state model $[p]$ and $[P]$ are given as:

$$[p] = \begin{bmatrix} 0 & p_{01} & p_{02} \\ p_{10} & 0 & p_{12} \\ p_{20} & p_{21} & 0 \end{bmatrix} \qquad [P] = \begin{bmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{bmatrix}$$

Where:

$$p_{01} = \frac{P_{01}}{P_{01} + P_{02}} \qquad p_{02} = \frac{P_{02}}{P_{01} + P_{02}} \qquad p_{10} = \frac{P_{10}}{P_{10} + P_{12}}$$

$$p_{12} = \frac{P_{12}}{P_{10} + P_{12}} \qquad p_{20} = \frac{P_{20}}{P_{20} + P_{21}} \qquad p_{21} = \frac{P_{21}}{P_{20} + P_{21}}$$

P_{ii} is the probability of remaining in state i whereas 1- P_{ii} is the probability of moving to another state.

$$E[S_i] = \sum_{n=1}^{\infty} nP_{ii}^{n-1}(1-P_{ii}) = (1-P_{ii})(1+2P_{ii}+3P_{ii}^{2}+\ldots..) = \sum_{n=1}^{\infty} P_{ii}^{n} = \frac{1}{1-P_{ii}}$$

$E[S_i]$ is the mean number of time intervals spent in state i. By multiplying $E[S_i]$ by the time for one interval d_i, then the mean sojourn time $d_i E[S_i]$ is obtained. In ATM d_i (for all i) are the same and so we denote $d = d_i$. The MMDP is often preferred to the Discrete Time Discrete State Markov Model because it more closely relates to the human's understanding of the operation of a source.

If the Discrete Time Discrete State Markov Model models are simulated then a uniform random number generator between 0.0 to 1.0 is run at every time interval. This procedure is a costly on computation time. Since the Bernoulli Process is the discrete version of the continuous time Poisson Process (see section 5.4), an exponentially distributed interarrival time can be generated and discretised to form a Geometrically distributed interarrival time. The mean interarrival time for a Poisson Process is given as:

$$E[X] = \frac{1}{\lambda} \qquad (1.1)$$

The mean interarrival time for a Geometric Distribution is given as

$$E[dS_i] = \frac{d}{1 - P_{ii}} \tag{1.2}$$

Equating equations (1.1) and (1.2) we obtain:

$$\lambda = \frac{1 - P_{ii}}{d}$$

5. STATISTICAL PARAMETERS OF A CONTINUOUS RANDOM VARIABLE

5.1 CONTINUOUS RANDOM VARIABLE, PROBABILITY DENSITY AND PROBABILITY DISTRIBUTION FUNCTION

A random variable X is called a continuous random variable if its probability distribution function $F_X(x)$ is everywhere continuous as shown in figure 1.4.

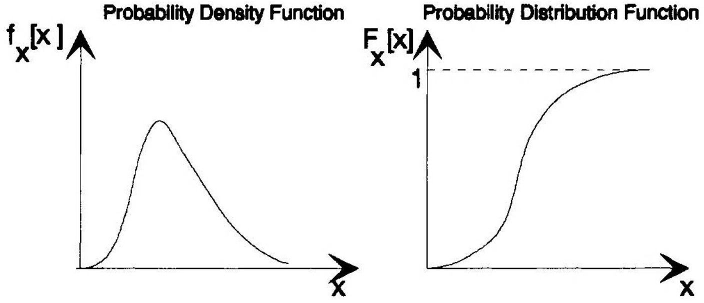


Figure 1.4. Probability Density Function and Probability Distribution Function of a Continuous Random Variable

The derivative of $F_X(x)$ is called the probability density function of X and is denoted by $f_X(x)$. Therefore

$$f_X(x) = \frac{dF_X(x)}{dx}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

5.2 EXPECTATION AND MEAN OF A CONTINUOUS RANDOM VARIABLE

$E[X]$ the expectation or mean of a continuous random variable,

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \mu_x$$

Provided the integral exists, then

$$\mu_X = \int_0^{\infty} x f_X(x)dx + \int_{-\infty}^{0} x f_X(x)dx$$

By applying integration by parts and by assuming that $E[|X|] < \infty$ then

$$\mu_X = \int_0^{\infty}[1 - F_X(x)]dx - \int_{-\infty}^{0} F_X(x)dx$$

Where $[1 - F_X(x)]$ is the complementary distribution function or the survivor function of the random variable *X*. If *X* is a non-negative random variable the following formula is obtained.

$$\mu_X = \int_0^{\infty}[1 - F_X(x)]dx$$

5.3 MOMENTS OF A CONTINUOUS RANDOM VARIABLE

The nth moment of *X*:

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)dx$$

The nth central moment of *X*:

$$E\left[\left(X - \mu_x\right)^n\right] = \int_{-\infty}^{\infty}\left(x - \mu_x\right)^n f_X(x)dx$$

The second central moment is given the special name variance σ_X^2. It is the mean of the squared deviations (dispersion) of a random variable about its mean. The variance is given as:

$$E\left[\left(X - \mu_x\right)^2\right] = \int_{-\infty}^{\infty}\left(x - \mu_x\right)^2 f_X(x)dx = E[X^2] - E[X]^2$$

5.4 THE POISSON PROCESS

Consider a finite time interval (0, *T*). Divide the period *T* into *m* subintervals each of length *h=T/m.* Letλ denote the average arrival rate of events as shown in figure 1.5.

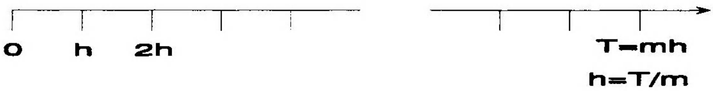


Figure 1.5. A finite time interval divided into m subintervals

For any subinterval the probability that one event arrives is $h\lambda$ + o(h), where $o(h)$ represents any quantity that approaches zero faster than h when $h \to 0$. The probability that no customers arrive is $1 - h\lambda - o(h)$. If we consider an arrival at an interval as a success of a Bernoulli trial, then the probability that exactly i customers arrive in the m subintervals is a binomial distribution. The arrivals are independent and identically distributed (iid).

$$p(i) = \binom{m}{i}[\lambda h + o(h)]^i[1 - \lambda h - o(h)]^{m-i}$$

Taking limits as $h \to 0$

$$p(i) = \binom{m}{i}[\lambda h]^i[1 - \lambda h]^{m-i} = \frac{m!}{i!(m-i)!}\left[\frac{\lambda T}{m}\right]^i\left[1 - \frac{\lambda T}{m}\right]^{m-i}$$

Taking limits as $m \to \infty$

$$p(i) = \frac{(\lambda T)^i}{i!} Lim_{m\to\infty} \frac{m!}{m^i(m-i)!}\left[1 - \frac{\lambda T}{m}\right]^{m-i}$$

but

$$Lim_{m\to\infty}\frac{m!}{m^i(m-i)!} = Lim_{m\to\infty}\frac{\overbrace{m(m-1)(m-2)....(m-i+1)}}{m^i} = 1$$

and

$$Lim_{m\to\infty}\left[1 - \frac{\lambda T}{m}\right]^{m-i} = Lim_{m\to\infty}\left[1 - (m-i)\frac{\lambda T}{m} + \frac{(m-i)(m-i-1)}{2!}\frac{(\lambda T)^2}{m^2}\frac{(m-i)(m-i-1)(m-i-2)}{3!}\frac{(\lambda T)^3}{m^3}.......\right]$$

$$\Rightarrow Lim_{m\to\infty}\left[1 - \frac{\lambda T}{m}\right]^{m-i} = 1 - \lambda T + \frac{(\lambda T)^2}{2!}\frac{(\lambda T)^3}{3!} + ... = e^{-\lambda T}$$

$$\Rightarrow p(i) = \frac{(\lambda T)^i}{i!} Lim_{m\to\infty}\frac{m!}{m^i(m-i)!}\left[1 - \frac{\lambda T}{m}\right]^{m-i} = \frac{(\lambda T)^i}{i!}e^{-\lambda T}$$

The number of arrivals i in the period T has the distribution above known as the Poisson Distribution. Let x be the interval from the time origin to the first arrival. No arrivals occur between 0 and x.

$$p_x(i = 0) = \frac{(\lambda x)^0}{i^0}e^{-\lambda x} = e^{-\lambda x}$$

$p_x(i \neq 0) = 1 - e^{-\lambda x} = F_x(x)$ the probability distribution function

$f_x(x) = \lambda e^{-\lambda x}$ the probability density function

5.5 STATISTICAL PARAMETERS OF THE POISSON PROCESS

The Poisson arrival process has exponentially distributed interarrival times as shown in figure 1.6.

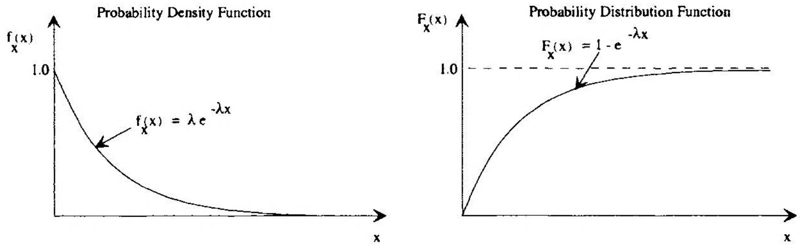


Figure 1.6. Probability Density Function and Probability Distribution Function of a Poisson Process

The probability distribution function is given as:

$$F_X(x) = 1 - e^{-\lambda x}$$

The probability density function is given as:

$$f_X(x) = \lambda e^{-\lambda x}$$

The characteristic of the exponential distribution is that it has a memoryless property. The past history of a random variable that is exponentially distributed plays no role in defining its future that is also an exponentially distributed random variable. It is therefore Markovian. A proof that the exponential distribution is memoryless can be found in section 5.6.

The expectation or mean:

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x}dx$$

$$E[X] = \lambda\left[\left(\frac{xe^{-\lambda x}}{-\lambda}\right) - \int\frac{e^{-\lambda x}}{-\lambda}dx\right]_0^\infty$$

$$E[X] = \lambda\left[\left(\frac{xe^{-\lambda x}}{-\lambda}\right) - \frac{e^{-\lambda x}}{\lambda^2}\right]_0^\infty$$

$$E[X] = \lambda\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda}$$

The second moment:

$$E\left[X^2\right] = \int_0^\infty x^2\lambda e^{-\lambda x}dx$$

$$E[X^2] = \lambda\left[\left(\frac{x^2e^{-\lambda x}}{-\lambda}\right) - \int\frac{2xe^{-\lambda x}}{-\lambda}dx\right]_0^\infty$$

$$E[X^2] = \lambda\left[\left(\frac{x^2e^{-\lambda x}}{-\lambda}\right) - \left(\frac{2xe^{-\lambda x}}{\lambda^2}\right) - \frac{2e^{-\lambda x}}{\lambda^3}\right]_0^\infty$$

$$E[X^2] = \lambda\left[\frac{2}{\lambda^3}\right] = \frac{2}{\lambda^2}$$

The variance:

$$E\left[X^2\right]-E\left[X\right]^2=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}$$

5.6 MEMORYLESS PROPERTY OF THE POISSON PROCESS

Consider a Poisson Process where arrivals had occurred at τ_0 and τ_1 as shown in figure 1.7.

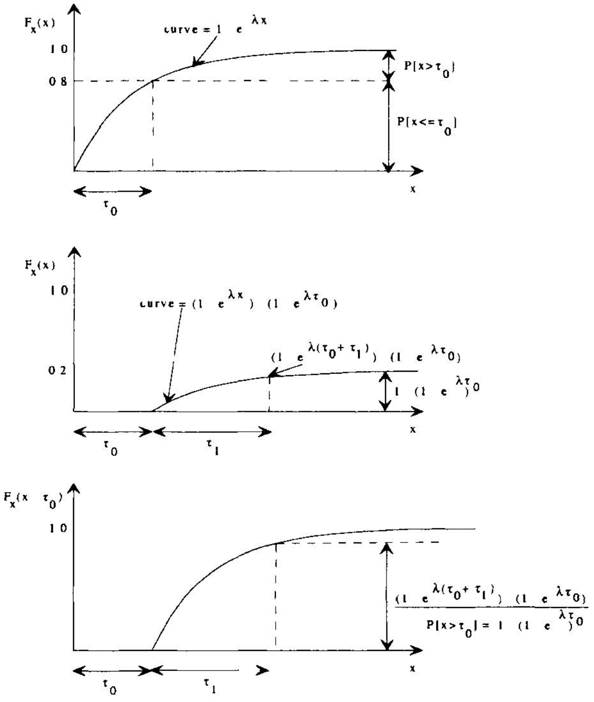


Figure 1.7. Memoryless property of Poisson Process

More formally:

$$P\left[A\ \ and\ \ B\right]=P[B]P[A|B]$$

$$\Rightarrow P[A\setminus B]=\frac{P\left[A\ \ and\ \ B\right]}{P[B]}$$

Applying this to the Poisson process:

$$\Rightarrow P\left[x \le \tau_0 + \tau_1 \middle| x > \tau_0\right] = \frac{P\left[\left(x \le \tau_0 + \tau_1\right) and \left(x > \tau_0\right)\right]}{P\left[x > \tau_0\right]} = \frac{P\left[\tau_0 < x \le \tau_0 + \tau_1\right]}{P\left[x > \tau_0\right]}$$

$$\Rightarrow P\left[x \le \tau_0 + \tau_1 \middle| x > \tau_0\right] = \frac{\left(1 - e^{-\lambda\left(\tau_0 + \tau_1\right)}\right) - \left(1 - e^{-\lambda\tau_0}\right)}{1 - \left(1 - e^{-\lambda\tau_0}\right)} = 1 - e^{-\lambda\tau_1}$$

Thus the Poisson process is said to have the memoryless property in that in calculating the probability of the remaining time before the next arrival, the time of the last arrival need not be considered.

5.6.1 Calling user model

Figure 1.8 shows a sequence of calls arriving at a Customer Premises Equipment (CPE) from its users. Each call is characterised by its duration, i.e. the time between its being set up and cleared. The interarrival time is the time between successive calls. Note - the expression 'interarrival time' derives from the fact that the process is traditionally considered from the point of view of the network. From the point of view of the population of users it is the time between successive call initiations. The setup and clearing times are not considered separately but are assumed to be included in the call duration for successful calls and zero for unsuccessful ones.

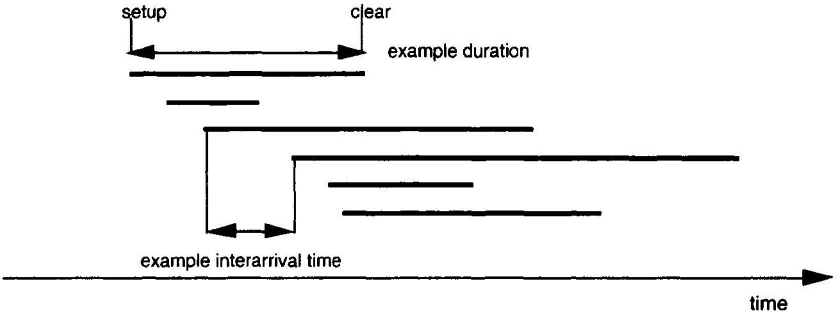


Figure 1.8. Call arrivals and duration

5.6.2 Called (answering) user model

In a real telephone network, if the called telephone is not busy, an incoming call causes the telephone to ring. The (human) user then takes a variable amount of time to answer the telephone or, indeed, may not answer it at all. If the telephone is in use (busy), the call is rejected by the called user's exchange.

A simplified model in Figure 1.9 shows the sequences of events for busy and not busy called users. If the telephone is busy then the call is immediately rejected by the called user's CPE. If the telephone is not busy then the call is answered immediately and is answered by the called user's CPE.

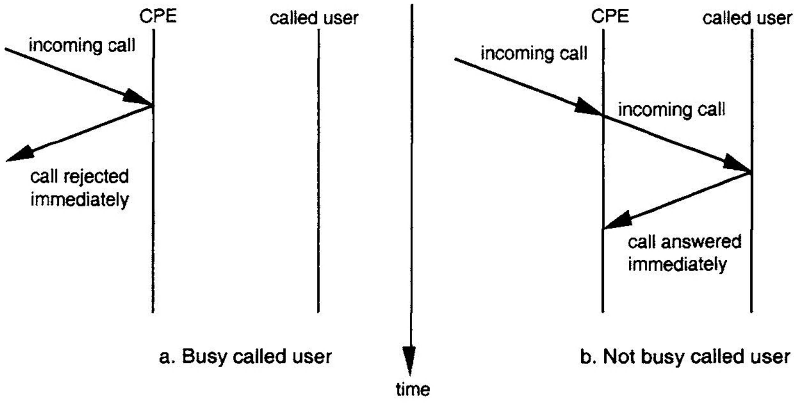


Figure 1.9. Called (answering) user model

5.6.3 Calling User Behaviour over time

A hierarchical model can represent the behaviour over a longer period of time. The values of mean inter call arrival time λ^{-1} and mean call holding time μ^{-1} can be changed at regular intervals of time and can have a trend within each time interval. These trends are shown in Table 1.1. These series of values can be generated, either in advance or during a simulation run, or can be based on measurements made on a real network.

Table 1.1. Traffic Trends

Model	Expression
Linear	$y_t = a + bt$
Parabolic	$y_t = a + bt + ct^2$
Exponential	$y_t = ae^{bt}$
Logistic	$y_t = \frac{M}{1 + ae^{bt}}$
Gompertz	$y_t = M(a)^{bt}$

6. CONTINUOUS TIME DISCRETE STATE MARKOV MODEL

6.1 MINISOURCE MODEL

This is the continuous time version of the Discrete Time Discrete State Markov Model and is shown in figure 1.10. The transitions between the two levels occur with exponential transition rates. The resultant rate $\lambda(t)$ is a continuous time process with discrete jumps at exponential transition rates.

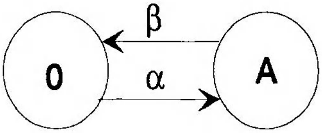


Figure 1.10. Minisource Model

- the information flow rate is A (cell rate) in the active state A and there is no information flow in the inactive state 0
- the mean burst duration 1/ß
  ß = rate of arrival to state 0 or rate of departure from state A
- the mean silence period 1/α
  α = rate of arrival to state A or rate of departure from state 0
- $P_i(t)$ probability of being in state i at continuous time t

The probability of remaining in state A after time Δt is the probability of being in state A times the rate of remaining in state A in time Δt plus the probability of being in state 0 times the rate of arrival from state 0 in time Δt.

$$P_A(t+\Delta t) = (1-\beta\Delta t)P_A(t) + \alpha\Delta t P_0(t)$$

$$\Rightarrow \frac{P_A(t+\Delta t) - P_A(t)}{\Delta t} = -\beta P_A(t) + \alpha P_0(t)$$

Taking the limit as Δt approaches 0 the left hand side of the equation represents the formal derivative of $P_A(t)$.

$$\Rightarrow \frac{dP_A(t)}{dt} = -\beta P_A(t) + \alpha P_0(t)$$

Also

$$P_0(t+\Delta t) = (1-\alpha\Delta t)P_0(t) + \Delta t \beta P_A(t)$$

$$\Rightarrow \frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\alpha P_0(t) + \beta P_A(t)$$

Taking the limit as Δt approaches 0 the left hand side of the equation represents the formal derivative of $P_0(t)$

$$\Rightarrow \frac{dP_0(t)}{dt} = -\alpha P_0(t) + \beta P_A(t)$$

Thus the forward equations that govern the evolution of the system are:

$$\frac{dP_A(t)}{dt} = -\beta P_A(t) + \alpha P_0(t)$$

$$\frac{dP_0(t)}{dt} = -\alpha P_0(t) + \beta P_A(t)$$

$$\Rightarrow \frac{dP_0(t)}{dt} = -\alpha P_0(t) + \beta - \beta P_0(t)$$

Since $P_0(t) + P_A(t) = 1$.

Taking the Laplace Transform to solve the differential equation:

$$sP_0(s) - P_0(0) = -\alpha P_0(s) + \frac{\beta}{s} - \beta P_0(s)$$

$$\Rightarrow P_0(s) = \frac{P_0(0)}{(s+\alpha+\beta)} + \frac{\beta}{s(s+\alpha+\beta)}$$

$$\Rightarrow P_0(s) = \frac{P_0(0)}{(s+\alpha+\beta)} + \frac{\beta}{s(\alpha+\beta)} - \frac{\beta}{(s+\alpha+\beta)(\alpha+\beta)}$$

Taking the inverse Laplace transform to obtain a solution for $P_0(t)$:

$$P_0(t) = \frac{\beta}{(\alpha+\beta)} + \left[P_0(0) - \frac{\beta}{(\alpha+\beta)}\right]e^{-(\alpha+\beta)t}$$

Similarly:

$$P_A(t) = \frac{\alpha}{(\alpha+\beta)} + \left[P_A(0) - \frac{\alpha}{(\alpha+\beta)}\right]e^{-(\alpha+\beta)t}$$

When $t \to \infty$ then:

$$P_0(t) = \frac{\beta}{(\alpha+\beta)}$$

$$P_A(t) = \frac{\alpha}{(\alpha+\beta)} = p$$

The mean arrival rate $E[\lambda(t)]$ as $t \to \infty$ is thus:

$$E[\lambda(t)] = \sum_{i=0,A} iP_i(t) = \frac{A\alpha}{(\alpha+\beta)} = Ap$$

The second moment of the arrival rate $E[\lambda^2(t)]$ as $t \to \infty$ is thus:

$$E[\lambda^2(t)] = \sum_{i=0,A} i^2P_i(t) = \frac{A^2\alpha}{(\alpha+\beta)} = A^2p$$

The variance of the arrival rate $Var[\lambda(t)]$ as $t \to \infty$ is thus:

$$Var[\lambda(t)] = E[\lambda^2(t)] - E[\lambda(t)]^2 = A^2p - A^2p^2 = A^2p(1-p)$$

The autocorrelation of the arrival rate $E[\lambda(t)\lambda(t+\tau)]$ as $t \to \infty$ is thus:

$$E[\lambda(t)\lambda(t+\tau)] = \sum_{j=0}^{1}\sum_{i=0}^{1} X_i X_j P[\lambda(t)=X_i]P[\lambda(t+\tau)=X_j|\lambda(t)=X_i]$$

$$E[\lambda(t)\lambda(t+\tau)] = A^2 P_A(t)P(\lambda(t+\tau)=A|\lambda(t)=A)$$

$$\Rightarrow E[\lambda(t)\lambda(t+\tau)] = A^2.\frac{\alpha}{(\alpha+\beta)}.P(\lambda(t+\tau)=A|\lambda(t)=A)$$

But

$$P(\lambda(t+\tau)=A|\lambda(t)=A) = \frac{\alpha}{(\alpha+\beta)} + \left[P_A(t) - \frac{\alpha}{(\alpha+\beta)}\right]e^{-(\alpha+\beta)\tau}$$

P_A(t) = 1 because the system starts in state A, so:

$$P(\lambda(t+\tau)=A|\lambda(t)=A) = \frac{\alpha}{(\alpha+\beta)} + \left[1 - \frac{\alpha}{(\alpha+\beta)}\right]e^{-(\alpha+\beta)\tau}$$

$$\Rightarrow P(\lambda(t+\tau)=A|\lambda(t)=A) = p + [1-p]e^{-(\alpha+\beta)\tau}$$

$$\Rightarrow E[\lambda(t)\lambda(t+\tau)] = A^2p^2 + A^2p[1-p]e^{-(\alpha+\beta)\tau}$$

The autocovariance of the arrival rate $C\,[\lambda(t)\lambda(t+\tau)]$ as t→ ∞is thus:

$$C[\lambda(t)\lambda(t+\tau)] = A^2p^2 + A^2p[1-p]e^{-(\alpha+\beta)\tau} - A^2p^2 = A^2p[1-p]e^{-(\alpha+\beta)\tau}$$

6.2 CONTINUOUS TIME DISCRETE STATE BIRTH-DEATH MARKOV MODEL (MULTI-MINISOURCE MODEL)

The multi-minisource model represents the superposition of M identical and independent minisource models and is shown in figure 1.11. It has also been proposed as a model for a video source. The model is based on a (M+l)-state continuous-time Markov Chain describing the number of sources that are currently active. The Markov Chain is a simple one-dimensional birth death process as shown in figure below, where the arrival rate in state i is given by iA. A is the information flow rate. The transitions between the M+1 levels occur with exponential transition rates. The parameter M is chosen such that any queue analysis [Anic82] generates queue length probabilities that are sufficiently similar to the queue length probabilities measured for the multiplexed video. The transition rates changes depending on the level of the bit rate. The resultant rate $\lambda(t)$ is a continuous time process with discrete jumps at exponential transition rates.

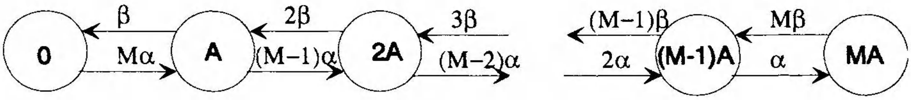


Figure 1.11. Multi-Minisource Model

For M identically distributed and independent minisources the mean arrival rate $E[\lambda_M(t)]$ as $t \to \infty$ is thus:

$$E[\lambda_M(t)] = MAp \tag{1.3}$$

For M identically distributed and independent minisources the variance of the arrival rate

$Var[\lambda_M(t)]$ as $t \to \infty$ is thus:

$$Var[\lambda_M(t)] = MA^2 p(1-p) \tag{1.4}$$

For M identically distributed and independent minisources the autocovariance of the arrival rate

$C[\lambda_M(t)\ \lambda_M(t+\tau)]$ as $t \to \infty$ is thus:

$$C[\lambda_M(t)\lambda_M(t+\tau)] = MA^2 p[1-p]e^{-(\alpha+\beta)\tau} \tag{1.5}$$

since $E[\ ]$ is a linear operator.

7. PARAMETERISATION OF MODELS

There are two main ways of parameterising models: 1) direct parameterisation, 2) parameterisation using unbiased estimators.

In direct parameterisation of the model parameters, if there is a direct relationship between the operation of a source and the states of a model then taking the appropriate measurements from the source can directly parameterise the parameters of the model.

In parameterisation using unbiased estimators, the mean and autocovariance can be expressed as equations in terms of the model parameters. These can then be equated to the measured mean and autocovariance of the source and solved for the model parameters.

For the discrete time discrete 2-state Markov process the expression of the autocorrelation contains a matrix operation which makes it unsuited for parameterisation using unbiased estimators since an equation of the autocorrelation at a given lag is complicated. For the continuous time discrete state birth-death Markov process the expression of the autocorrelation can be expressed as an equation and is thus suited for parameterisation using unbiased estimators.

Parameters for constant bit rate (cbr) and variable bit rate (vbr) traffic models for voice, video and data services, are proposed in [Cosm94]. Variable rate data services are modelled as 2-state (on/off) and 3-state discrete time Markov models. Variable rate video services are modelled as 5-state (birth-death) discrete time Markov models. Justifications for the models are given in [Cosm94]. The following coupling of sources with source models is proposed in Table 1.2.

Table 1.2. Coupling of Sources with Source Models

Source	Model	Comment
CBR Voice, Data and Video	Deterministic	
VBR Voice	Discrete Time Discrete State Markov	TalkSpurt/Silence Period codec
VBR Data	Discrete Time Discrete State Markov	
VBR Video Phone	Continuous Time Discrete State Markov	
VBR Video Distributive	Continuous Time Discrete State Markov + Discrete Time Discrete State Markov	For image sequence For camera zoom/pan and scene changes
Connection Level Voice, Data and Video	Poisson Process	Models duration of a call

8. FURTHER SOURCE MODELS

8.1 MARKOV MODULATED POISSON PROCESS

The Markov Modulated Poisson Process (MMPP) can be used to represent the superposition of on-off sources [Bai91], see figure 1.12. An MMPP is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain. For a two-state Markov chain, the mean transition rates out of states 1 and 2 are r_1^{-1} and r_2^{-1}, and the arrival process is Poisson with arrival rates λ_1 and λ_2. State 1 is referred to as the underload state because the sum of the arrival rates is always less than the maximum queue capacity. State 2 is referred to as the overload state because the sum of all the arrival rates is always greater than the maximum queue capacity.

The mean arrival rate in the underload state λ_1 is the probability of being in state i times the arrival rate in state I, iA where $0 \leq I \leq N$.

$$\lambda_1 = \sum_{i=0}^{M} iAF_i$$

The mean arrival rate in the overload state λ_2 is the probability of being in state i times the arrival rate in state I, iA where $N < I \leq M$.

$$\lambda_2 = \sum_{i=N+1}^{M} iAF_i$$

where:

$$F_i = \frac{1}{\left(1+\frac{\alpha}{\beta}\right)^M}\binom{M}{i}\left(\frac{\alpha}{\beta}\right)^i$$

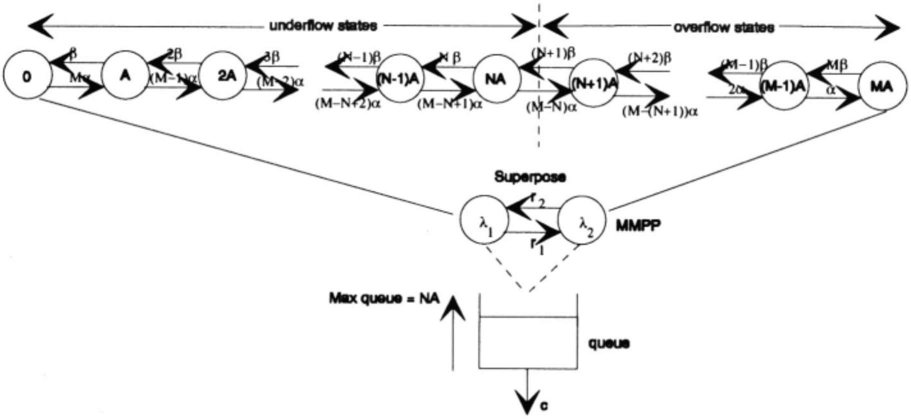


Figure 1.12. MMPP as a superposition of on-off sources

Figure 1.12: MMPP as a superposition of on-off sources

The system of differential equations, which describe the source behaviour, is given as:

$$\frac{d}{dt}\begin{bmatrix} P_0(t) \\ P_1(t) \\ P_{20}(t) \\ \cdots \\ P_{M-2}(t) \\ P_{M-1}(t) \\ P_M(t) \end{bmatrix} = \begin{bmatrix} -M\alpha & \beta & 0 & \cdots & 0 & 0 & 0 \\ M\alpha & -((M-1)\alpha+\beta) & 2\beta & \cdots & 0 & 0 & 0 \\ 0 & (M-1)\alpha & -((M-2)\alpha+2\beta) & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -(2\alpha+(M-2)\beta) & (M-1)\beta & 0 \\ 0 & 0 & 0 & \cdots & 2\alpha & -(\alpha+(M-1)\beta) & M\beta \\ 0 & 0 & 0 & \cdots & 0 & \alpha & -M\beta \end{bmatrix}\begin{bmatrix} P_0(t) \\ P_1(t) \\ P_{20}(t) \\ \cdots \\ P_{M-2}(t) \\ P_{M-1}(t) \\ P_M(t) \end{bmatrix}$$

A phase process is constructed from the system of differential equations, which consists of the overload states and an additional absorbing state M+1.

$$\frac{d}{dt}\begin{bmatrix} P_{N+1} \\ P_{N+2} \\ \cdots \\ P_{M-2} \\ P_{M-1} \\ P_M \\ P_{M+1} \end{bmatrix} = \begin{bmatrix} -((M-(N+1))\alpha+(N+1)\beta) & (N+2)\beta & \cdots & 0 & 0 & 0 & (M-N)\alpha \\ (M-(N+1))\alpha & -((M-(N+2))\alpha+(N+2)\beta) & \cdots & 0 & 0 & 0 & 0 \\ 0 & (M-(N+2))\alpha & \cdots & (M-2)\beta & 0 & 0 & 0 \\ 0 & 0 & \cdots & -(2\alpha+(M-2)\beta) & (M-1)\beta & 0 & 0 \\ 0 & 0 & \cdots & 2\alpha & -(\alpha+(M-1)\beta) & M\beta & 0 \\ 0 & 0 & \cdots & 0 & \alpha & -M\beta & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} P_{N+1} \\ P_{N+2} \\ \cdots \\ P_{M-2} \\ P_{M-1} \\ P_M \\ P_{M+1} \end{bmatrix}$$

Where:

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

Transition into the absorbing state M+l is certain. The solution of the differential equations provides expressions of the probabilities P_{N+1}, P_{N+2},... , P_{M-1} , P_M or the probability distribution of remaining in the transient (overload) states.

$$\frac{d}{dt}[P]=[T][P]$$

The solution is of the form:

$$[P]=\alpha.e^{[T]t}$$

Where the initial conditions are α:

The probability distribution function of the time to absorption is:

$$F(t)=1-[P][e]$$

$$\Rightarrow F(t)=\alpha.e^{[T]t}[e]$$

Where $[e]^T = [11...1]$

The time until absorption in [*T*] has a phase probability distribution. The complementary probability distribution function of the time to absorption [Neut81] is:

$$1-F(t)=ke^{-\eta t}+o\left(e^{-\eta t}\right) \quad \text{as } t\to\infty$$

It can be shown that the maximum real part eigenvalue of [*T*], which is real and negative, is -η.

Approximating the phase distribution as a negative exponential distribution, then

$$r_{ol}=\eta$$

8.2 DISCRETE MARKOV ARRIVAL PROCESS (DMAP)

The DMAP consists of two, transition matrix [*C*] and [*D*]. The [*D*] matrix consists of the probabilities of those transitions that go to an absorbing state *A* and then the process is immediately started in a transient state [Blon89]. Upon transition to the absorbing state an arrival is generated. The [*C*] matrix consists of the probabilities of those transitions that occur between the transient states where no arrivals are generated.

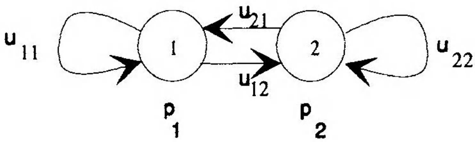


Figure 1.13. Discrete Time Markov Modulated Geometric Arrival Process

A discrete time Markov modulated geometric arrival process consists of a transition probability matrix $[U]$ and an arrival may occur at each state i with probability p_i. , as shown in figure 1.13. Figure 1.14 illustrates the two transition diagrams of the corresponding DMAP. The motivation of representing sources as a DMAP is that the performance analysis of the associated queuing model is tractable using matrix analytic techniques.

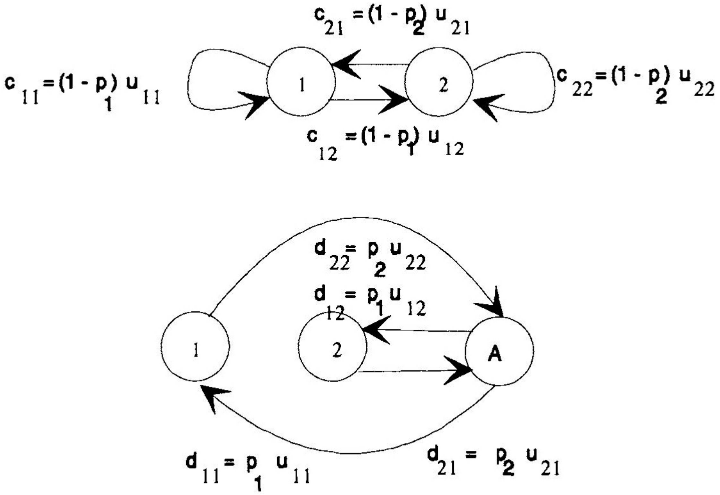

$c_{21} = (1 - p_2) u_{21}$
$c_{11} = (1 - p_1) u_{11}$
1
2
$c_{22} = (1 - p_2) u_{22}$
$c_{12} = (1 - p_1) u_{12}$
$d_{22} = p_2 u_{22}$
$d_{12} = p_1 u_{12}$
1
2
A
$d_{11} = p_1 u_{11}$
$d_{21} = p_2 u_{21}$

Figure 1.14. Discrete Markov Arrival Process

8.3 AUTOREGRESSIVE MOVING AVERAGE MODEL (ARMA)

Autoregressive model, shown in figure 1.15, moving average model, shown in figure 1.16, and autoregressive moving average model shown in figure 1.17 are mainly used to model call arrivals [Box70]. Differencing is used to factor out linear trends and spectral analysis/differencing is used to factor out periodic trends in call arrivals. The residual random variable is 'white noisy' and normally distributed. This can then be modelled as an autoregressive, moving average or autoregressive moving average model whose order need never be greater than two for almost all cases. Analysis of the autocorrelation function and the partial autocorrelation function is used to determine what type and order of autoregressive moving average model is required to model the residual random variable.

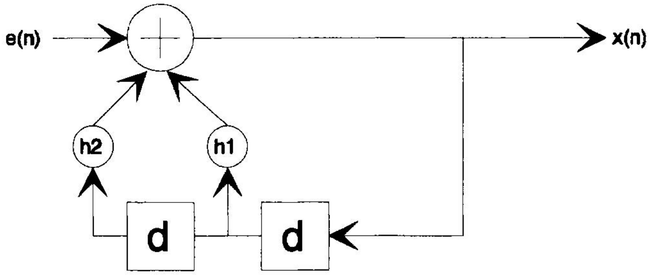


Figure 1.15. Autoregressive Process oforder 2

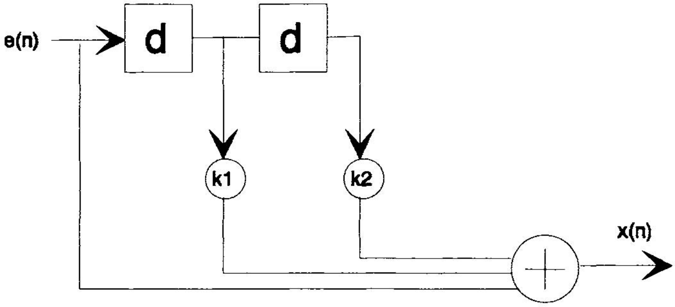


Figure 1.16. Moving Average Process oforder 2

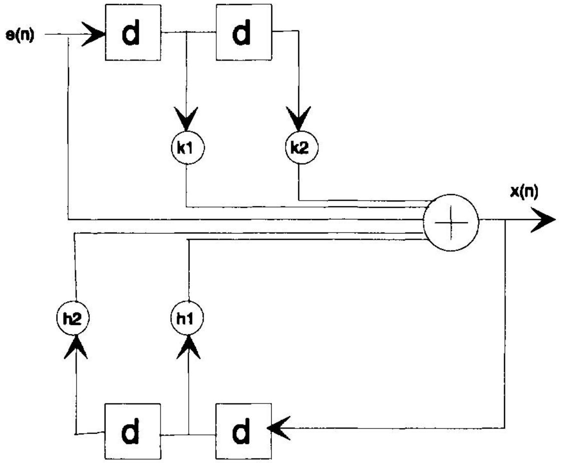


Figure 1.17. Autoregressive-Moving Average Process of order 2

References

[Anic82] D. Anick, D. Mitra, M. Sondhi: *'Stochastic Theory of a Data-Handling System with Multiple Sources'* The Bell System Technical Journal, Vol. 61, No. 8, pp. 1871-1894, October 1982.

[Baio91] A. Baiocchi, et al: *'Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources',* IEEE JSAC, Vol. SAC-9, No. 3, pp.388-393, 1991.

[Blon89] C. Blondia, T.Theimer *'A Discrete Time Modelfor ATM Traffic'* RACE 1022, PRLB_123_0018_CD_CC, October 1989.

[Box70] G.E.P. Box, G.M. Jenkins *'Time Series Analysis - Forecasting and Control'* Holden-Day, 1970.

[Cosm94] J. Cosmas et al *"A Review of Voice, Data and Video Source Modelsfor ATM "* European Transactions on Telecommunications, Vol. 5, No. 2, Mar-Apr 1994, ppll-26. ISSN 1120-3862

[Cox87] D.R. Cox, H.D. Miller *'The Theory ofStochastic Processes'* Chapman and Hall, 1987.

[Klei75] L. Kleinrock *'Queueing Systems Volume 1: Theory' J.* Wiley and Sons, 1975.

[Krey70] E. Kreyszig *'Introductory Mathematical Statistics'* J. Wiley & Sons, 1970.

[Mag88] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, J. Robbins *"Performance Models of Statistical Multiplexing in Packet Video Communications"* IEEE Trans. Commun., Vol. 36, NO. 7, July 1988.

[Neut79] M.F. Neuts "*A Versatile Markovian point Process*" J. Appl. Prob., Vol. 16, 1979, p764-779.

John Paul Cosmas (MIEE '90, MIEEE 90, CEng) obtained a BSc(Eng) honours degree in Electronic Engineering at Liverpool University in 1978 and a PhD in Image Processing and Pattern Recognition at Imperial College in 1987. Between 1978 and 1983, he worked as an electronics development engineer at Tube Investments and Fairchild Camera and Instruments. In 1983, he joined Imperial College as a Research Student and in 1986 Queen Mary and Westfield College as a Lecturer in Digital Systems Design, Computer Structures and Telecommunications. His research interests are focussed on video processing and Multimedia systems. In 1999 he joined Brunei University as a Reader of Multimedia systems. He has contributed towards the EEC RACE projects R1022 Technology for ATD' and R2072 'Mobile Audio-Visual Terminal' and ACTS projects A0098 'Mobile Multimedia System (MoMuSys)' and A0360 'CustomTV'. He is presently working on the EU Framework 5 project 'System for Advanced Multimedia Broadcast and IT Services (SAMBITS)'.

Chapter 2

FRACTALS AND CHAOS FOR MODELLING MULTIMEDIA ATM TRAFFIC

MAREK BROMIRSKI
WIESLAW LOBEJKO
Military Communications Institute, 05-130 Zegrze, Poland

Abstract The paper demonstrates that, beyond its statistical significance in traffic measurements, both fractal and chaotic properties have considerable impact on the performance of ATM system, and is a dominant characteristic for a number of packet traffic engineering problem. We discuss the underlying mathematical and statistical properties of dynamical chaotic systems, and indicate that self-similarity has serious implications for analysis and control of ATM traffic flow.

Keywords: Telecommunication traffic, Density fluctuation, Fractal, Chaos theory

1. INTRODUCTION

When modeling modern telecommunication network traffic, packet and connection arrivals are often assumed to be Poisson processes because such processes have attractive theoretical properties. A numer of studies have shown, however, that for both VBR and UBR traffic, the distribution of packet interarrival clearly differs from exponentials. In the last decade there has been much interest in the study of transitions to chaos and the onset of stochasticity in telecommunication systems [13]. It has been discovered that fractal structures often appear on the border between regular and chaotic dynamical behaviour [15]. In the best know examples like both LAN-LAN and multimedia services in ATM systems, the teletraffic exhibits robust scaling properties, and ideas from fractal geometry and chaos theory can be naturally applied [6].

This paper demonstrates that ideas from fractals and chaos are in fact applicable to many aspects of ATM systems and allow for efficient modelling of VBR and UBR telecommunication traffic. Although classical Markovian models can in principle always be used to accurately describe any finite set of teletraffic measurements, the resulting models needed to capture the fractal-like nature of measured traffic are bound to be very complex and highly parametrized [5]. We consider in this paper alternatives to finite time scale measurements that take account fractal properties.

The rest of the paper is organized as follows. In Section 2 some basic concepts and ideas from fractal geometry and chaos theory are discussed. Section 3 gives a comparison of both MPEG and LAN-LAN data traces from point of view of fractal description. Section 4 concludes with a summary of the paper.

2. FRACTALS AND CHAOS

One way to define ”fractal” is as a negation: a fractal is a set that does not look like a Euclidean object (point, line, plane, etc.) no matter how closely you look at it [24]. Euclidean geometry is a description of lines, circles, ellipses and so on [21]. Fractal geometry is described in algorithms - a set of instructions on how to create the fractal [1]. Computers translate the instructions into the magnificent patterns, such as the Mandelbrot set shown in Fig. 1.

The fractals are images of processes of a mathematical explorations of the space in which they are plotted. Let’s take the computer screen as representing a space. Each point on the screen is tested in some way. Usually an equation is iterated with this point as its starting value. That means a result is calculated using the equation, and this value is fed back into the equation leading to a further results being calculated. This process is repeated over and over and over. As a results of this calculation, the point on the screen at which we started is plotted in a particular colour. Then the computer repeats the process for the next point on the screen.

The fractal which is plotted in Figure 1 is far more than a flat picture. A fractal is infinitely complex. That is, if we zoom in on any part of the Mandelbrot set we will always find more detail. Each stage tends to have the same form as the original. So the fractal lacks scale. A small portion of the fractal is just as detailed as the original. The natural world has always had a fractal way about it. A piece of fern has a fractal form. As depicted on Figure 2, each segment of the fern is similar to the whole,

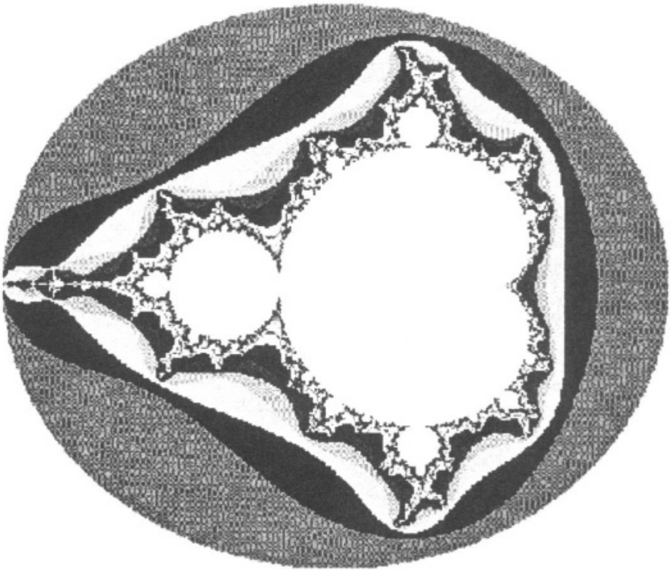


Figure 2.1. Mandelbrot set - an example of the fractal

and the segments then break down into smaller, similar segments. Ferns are limited by the real world boundaries. Mathematicians, computers and imaginations are not. Hence mathematically generated fractals go on forever, deep within themselves.

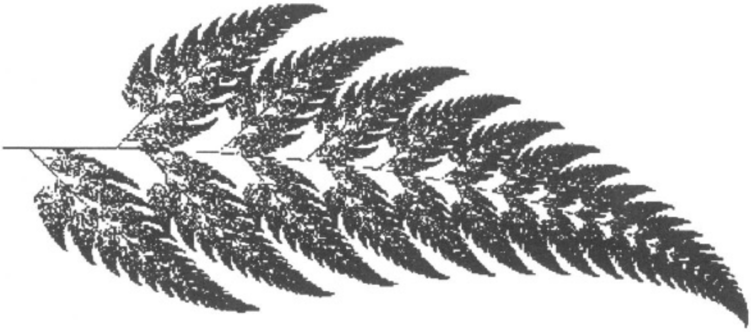


Figure 2.2. Self-similarity of fern

The amazing thing about fractals is that the formulae used to generate them are often extremely simple. A simple formula can lead to complex images. These images are very sensitive to the initial conditions. This is one of main properties of the all fractal-like or chaotic systems

The chaos (irregular or seemingly stochastic behaviour) exhibited by such system arises from a property known as Sensitive dependence on Initial Conditions (SIC) [25]. The Oxford Concise Dictionary defines chaos as ”Formless primordial matter”. The day has come when there is a need for an update. Population dynamics in one area which can be very sensitive to small changes in initial conditions. So can the weather. A butterfly flapping its wings in a South American jungle, it is said, can lead to a hurricane in China [17]. This is the signature of Chaos Theory.

As scientists studied chaotic systems, mathematics evolved which had already drawn interest from pure mathematicians. This mathematics involved iteration - taking the answer to the equation and feeding it back into the equation, over and over again. In watching the results of this process, some fascinating behaviours were observed. When the mathematicians and scientists got together, with the benefits of machines which could do their calculations within minutes, a new science was born. Chaotic systems are not random. They may appear to be. They have some simple defining features [2]:

- Chaotic systems are deterministic. This means they have some determining equation ruling their behaviour.
- Chaotic systems are very sensitive to the initial conditions. A very slight change in the starting point can lead to enormously different outcomes. This makes the system fairly unpredictable.
- Chaotic systems appear to be disorderly, even random, but they are not. Beneath the seemingly random behaviour is a sense of order and pattern. Truly random systems are not chaotic. The orderly systems predicted by classical physic are the exception. In this real world of our, chaos rules.

In general, chaos theory is riddled with strange patterns which underlie seemingly random and unpredictable behaviour. Although we don’t know for example which telecommunication traffic fluctuation will occur, there are traffic patterns which are possible, and likely, and others which are not. The likely patterns are called attractors [8]. These patterns attract the system into their state. In examining the traffic generation process by many particular traffic sources it is found certain patterns occur and often lead to some kind of bounded behaviour. These are attractors for the teletraffic processes which are presented more detailed in next section of this paper. There are a number of ways of modelling chaotic system with attractor. Most common strategy in this case is iterate a set of non-linear equations [14]. The simple functions that can take to the two dimensional attractor (named Henon's attractor) are given by following set of iteration formulas:

$$\begin{aligned} x_{n+1} &= (y_n + 1) - ax_n^2 \\ y_{n+1} &= bx_n \end{aligned} \tag{2.1}$$

Figure 3 shows the image of the Henon strange chaotic attractor for a=1.4 and b=0.3. In this instance the word strange refers to the geometry of the of the underlying attractor as exhibiting a fractal structure, while the word chaotic refers to the particular dynamics of orbits on the attractor.

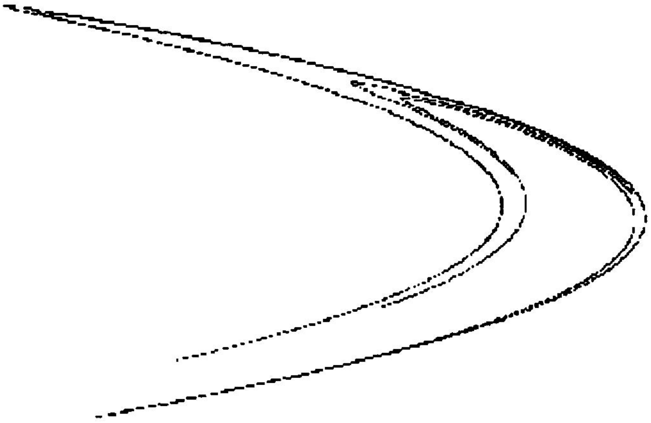


Figure 2.3. Trajectory of Hênon system

A chaotic attractor is an attractor for which nearby orbits diverge exponentially in time, displaying sensitive dependence on initial conditions (at least one positive Lyapunow exponent). A nonchaotic attractor, however, is an attractor for which nearby orbits typically do not diverge exponentially in time (no positive Lyapunow exponent). Consequently, a strange chaotic attractor is an attractor that is geometrically strange, for which nearby trajectories diverge exponentially. Figure 4 illustrates other strange chaotic attractor (named Lorenz attractor).

An analysis of strange attractors proceeds by determining the topological organization of unstable periodic orbits have been recognized as a

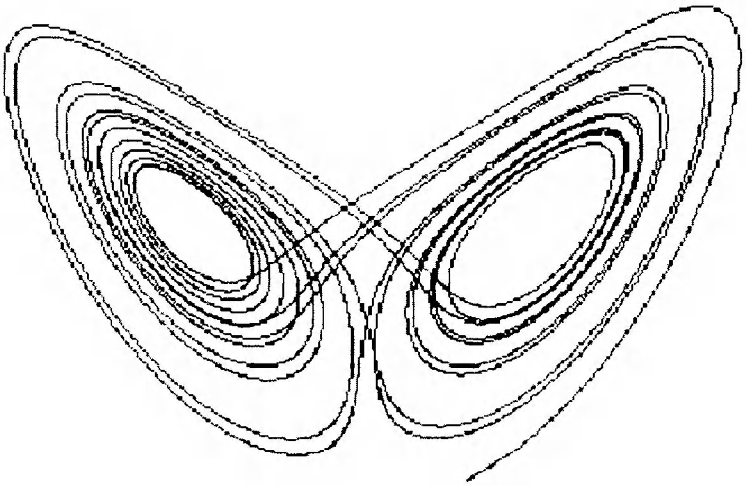


Figure 2.4. Lorenz attractor

major tool in characterising low dimensional chaotic behaviour, particularly because they can be extracted from experimental data series. Key to this approach is the fact that, for an attractor embedded in a three-dimensional phase space, topological invariants, such a linking numbers or knot polynomials, may be used to determine in which way its periodic orbits are knotted and linked with each other.

In order to identify attractors in experiments, we need to identify the potentially measurable signatures that characterize these attractors. In particular, we discuss below the Fourier amplitude spectra of time series, fractal dimensions, Hurst and Lyapunow exponents and bifurcation diagrams.

Fourier amplitude spectra - The discrete-time Fourier amplitude spectrum $\frac{1}{2}$ $S(f)$ $\frac{1}{2}$ as a function of frequency $\mathtt{f}$ for a time-varying quantity is the magnitude of the Fourier transform of the discrete-time series obtained by sampling the process under investigation [3]. In practice any $\mathtt{d}$ functions contained in $|S(f)|$ acquire a width and a finite peak value (e.g. due to the finite duration of the time series). A peak is defined as a local maximum in the discrete spectrum. Define the spectral distribution function $N(\sigma)$ as the number of peaks $|S(f)|$ with amplitude greater than σ. Distinct scaling relations for $N(\sigma)$ have been predicted for strange chaotic and nonchaotic attractors. Figure 5 illustrates an example of discrete-time Fourier amplitude spectrum of Bellcore LAN traffic.

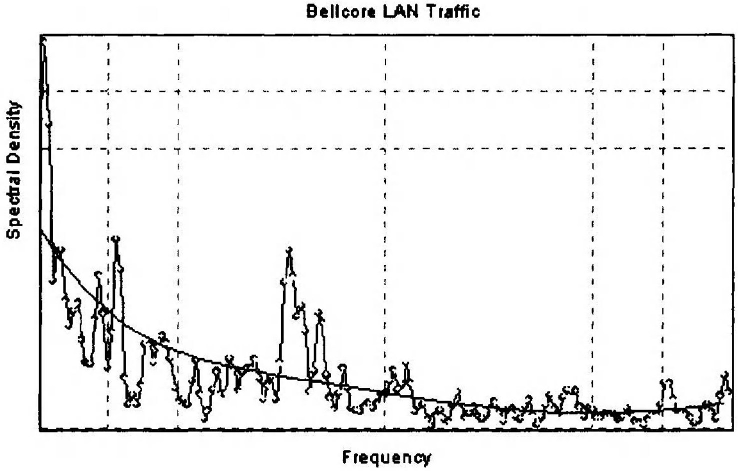


Figure 2.5. Discrete-time Fourier amplitude spectrum of Bellcore LAN traffic

Fractal dimension - The information dimension of an attractor is defined as [18]:

$$d_i = lim_{\epsilon \to 0} - \frac{I(\epsilon)}{ln(\epsilon)} \tag{2.2}$$

where the attractor has been covered by cubes from a Cartesian grid of spacing ϵ in the phase space. Here $I(\epsilon) = \sum_{i=1}^{N(\epsilon)} p_i ln(p_i)$ where p_i is the measure of the attractor in the i-th cube of the cover. In experiments p_i can be estimated as the fraction of time that a finite-duration orbit spends in cube i. Note that d_i=*2* does not rule out fractal structure. In particular, it has been claimed [11] than the capacity dimension d_c is strictly greater than the information dimension for strange attractors. To summarise, the information we will be using is as follows. A strange chaotic attractor occurs only if $d_i > 2$ and a strange nonchaotic attractor occurs if $d_i < 2$ with $N(\sigma) \sim \sigma^{-\alpha}$.

Hurst exponent - The Hurst exponent was developed to estimate the fluctuations which occurred in the time series of data. The Hurst exponent is named after the hydrologist who measured the daily water discharge levels [22]. He found a positive correlation between the "peaks" and "troughs" which fluctuate about an average value across the time-series of water level measurement. This statistic was used to quantify the persistence or antipersistence of feature details. We note that in one dimensional sequences the Hurst exponent falls in range of: $0 < H < 1$. A persistent trend is characterised by repetitive behaviour. For example, if a high value occurs at time t_x then at time t_{x+1} one would expect the probability of another high value to be greater than 0.5. Persistent trends fall in the range $0.5 < H \leq 1.0$. Note that a random walk process which exhibits no correlation between values has a Hurst exponent $H = 0.5$. This contrasts with an antipersistent trend where successive values are likely to alternate. For example, if a high value occurs at time t_x then at time t_{x+1} one would be more likely to see a low value. Similarly, antipersistent scaling has a Hurst exponent $0 \leq H < 0.5$. Figure 6 shows a typical example of the processes with different values of *H*. These processes were generated by *Random Midpoint Displacement* (RMD) method [22]. RMD is fast and allows the rapid generation of long traces, and it is known to be exact for all values of *H* parameter.

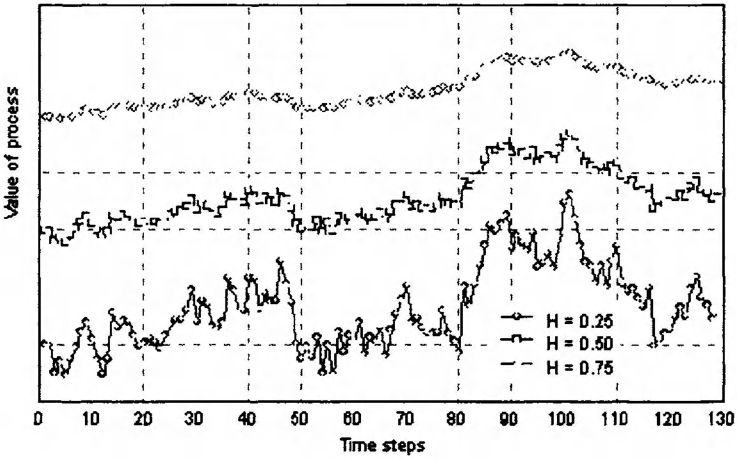


Figure 2.6. Example of the processes with different values of H

Lyapunow exponent - When one studies an attractor of a chaotic dynamical system quantitatively, one is often interested in estimating a "time average": the average of a given function of the state of the system over a typical trajectory on, or approaching, the attractor [12]. Of particular interest are Lyapunow exponents, which reflect average rates of linear expansion or contraction near the attractor [10]. In some very special cases, Lyapunow exponents can be determined exactly as follows [9]:

$$\lambda(x_0) = lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} log_2|f'(x_k)| \qquad (2.3)$$

where $\lambda(.)$ is Lyapunow exponent, x_0 is starting point of iteration and *f(.)* is characteristic function of the system under investigation. Unfortunately, in practical systems like flow of ATM cells on the output of the multiplexer, *f(.)* is unknown and the Lyapunow exponent must be computed with the aid of the simulation. The main approach to estimating a time average is of course to compute **N** trajectories near the point x_0. If $\lambda(.)<0$, the trajectories are attracted by x_0 (system is stable). In other cases $\lambda(.)$ ≥ 0) the system is chaotic.

Bifurcation diagrams - Most systems, however, are neither chaotic nor integrable but show a complicated mixture of regular and chaotic behaviour. Bifurcation is a phenomenon exhibited in systems with mixed phase space [4]. They are responsible for the rapid increase of the number of periodic orbits when an integrable system is transformed into a chaotic system, e. g. by changing an external parameter. If one changes this parameter by an arbitrarily small but finite amount, then in general an infinite number of bifurcations occur, since they take place any time that the stability angle of a stable orbit is a rational multiple of **2**.

There are different kinds of generic bifurcations, but the number of different forms is limited. They are characterized by normal forms which describe the characteristic classical motion in the vicinity of a periodic orbit. They have the property that a central periodic orbit bifurcates and other periodic orbits split from the central orbit whose primitive period is **m** times the primitive period of the central orbit. (An exception is the case **m=1** for which there is no periodic orbit before the bifurcation.) The most simple function that can take to the bifurcation is given by Verhulst (logistic) equation [19]:

$$p_{n+1} = rp_n(1 - p_n) \tag{2.4}$$

When the constant parameter *r* is less than **1**, the process gradually decreases to zero, and when *r* is between **1** and **3**, the process settles down to some stable equilibrium value. However, when *r* is greater than 3, some rather more interesting things start to happen (Figure 7).

Firstly, the stable equilibrium doesn't appear above *r=3.00.* However many times you calculate the number in the next iteration, the numbers will not stabilise. Up to *r=3.00,* the process oscillates between two values. To visualise more easily what happens, we can draw a bifurcation diagram (see Figure 8). This is drawn by taking each value for *r* in turn, and calculating the iterations which values are plotted in the vertical direction on the diagram.

So we can see in the plot, that at *r=3.50,* the process cycles between four distinct values, but at *r=3.8,* there is no pattern linking the steps of iteration, i.e. the system becomes chaotic. However, as *r* is increased

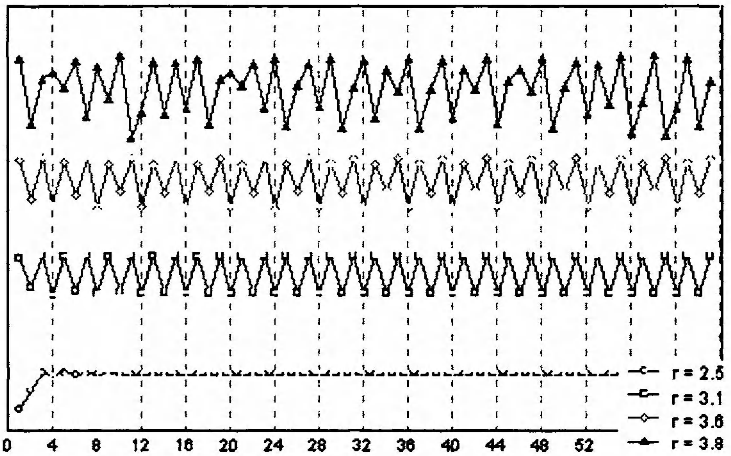


Figure 2.7. Logistic process for different values of r

more, we find small areas of stable behaviour; for example when *r=3.83*, the process cycles between three distinct values.

Combining spectral distribution scaling and dimension measurements one can systematically support or rule out the existence of the above-mentioned chaotic or nonchaotic attractors in an experimental time series. In practice the dimension measurements are performed on surface-of-section data, reducing all of the dimensions by one.

3. CHAOTIC PROPERTIES OF ATM TRAFFIC

Recent traffic measurements from a wide range of working packet networks have convincingly established the presence of significant statistical features that are characteristic of *fractal traffic processes* (FP), in the sense that these features span many time scales [7]. Of particular interest in packet traffic modelling is a property called long-range dependence (LRD) which is marked by the presence of correlations that can extend over many time scales. Leyland et al. [16] observed the Ethernet traffic seems to look the same in the large scales (min, h) as in the small (s, ms).

Leyland [16] re-visited the Bellcore Ethernet LAN traffic and extract from the aggregate traffic the traces generated by individual source-destination pairs. Statistical analysis of these traces reveals that:

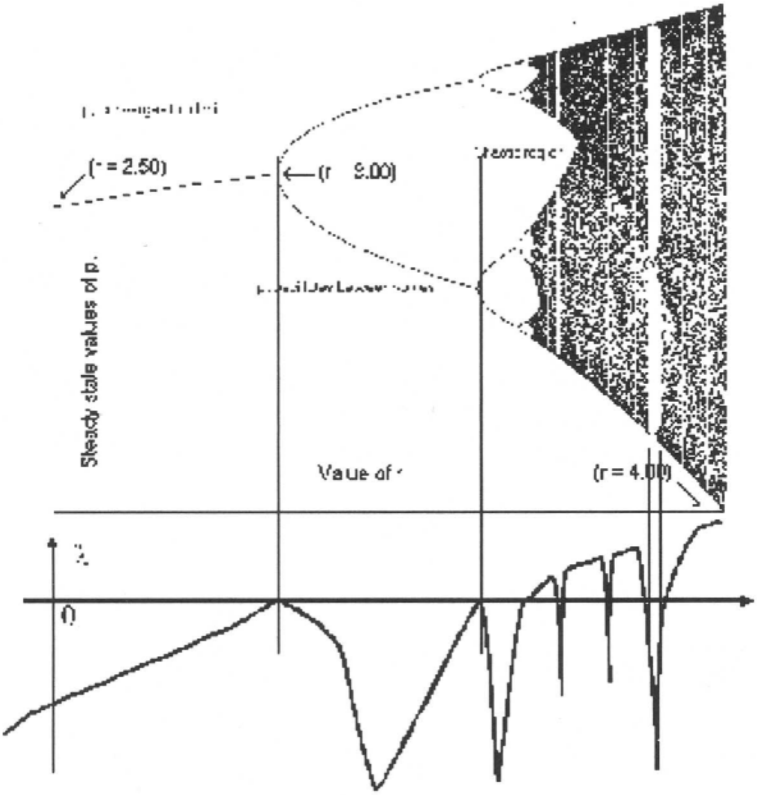


Figure 2.8. Bifurcation and Lyapunow exponent for the logistic process

- the traffic generated by each pair is consistent with an ON/OFF model;

- the distribution of the sojourn times in the ON/OFF states can be accurately described using Pareto-type distributions which exhibit infinite variance. Thus, the examined traffic data are not only consistent with self-similarity of aggregate packet traffic, but they are also in full agreement with given below explanation. It is reasonable to assume, that LAN traffic measured on Ethernet can be examined at three major levels of behaviour corresponding to certain resolution of time [5]:

- The connection level describes the human behaviour. The connection duration is determined by the file sending time and file length. In tactical LAN networks both parameters are additionally determined by specific requirements and limitations. The duration between calls on an Ethernet is typically in time range of 10 - 1000 s.

- The TCP/IP level describes the transport level. The traffic sent on the network depends of an uncontrollable number of parameters but the major influences on it is the network behaviour. The transmission duration of a TCP/IP packet varies typically from 0.01 - 10 s.
- The Ethernet network level where the sent traffic depends essentially on the local traffic flowing on the network. The time between sending and not sending a frame is typically in the range 1-50 ms.

In our considerations, we use an exactly self-similar model, based on *Fractional Brownian Motion* (FBM) which has been proposed by [20]. In this model the total amount of traffic arriving to a system until time *t* is given by

$$A(t) = mt + \sqrt{cm}z(t), t\epsilon(-\infty, \infty) \tag{2.5}$$

where *Z(t)* is normalized FBM characterized by the self-similarity Hurst parameter *H(0.5, 1).* Norros uses a scaling analysis to derive analytic expression with regards to the *Quality of Service* (QoS) criteria. In particular Norros shows that the complementary queue distribution is asymptotically bounded by a stretched exponential or Weibull form

$$P(L > x) \approx ae^{-\gamma x^{\beta}}, 0 \leq \beta \leq 1 \tag{2.6}$$

where $\gamma = f(c, m, H)$ and $\beta = 2 - 2H$. This form of the queue length distribution for $H > 0.5$, is much heavier than the exponential decay predicted by traditional model. The rest of this section presents the fractal properties of ABR as well as VBR data traces. The measured LAN-LAN traffic has been obtained by a working ATM network called VISTAnet which is a gigabit testbed sponsored by the National Science Foundation and was designed to implement a medical imaging application and LAN-LAN services over large distances in North Carolina. The real MPEG traffic under test consists of a number of MPEG2 files. The data was collected from the files available by Internet which consist in each case of not more than 2 MB coded videos. About 32 MB data were proceeded by software to obtain a form appropriate for use in the experiments. The bit-rate in MPEG as well as in LAN-LAN systems are plotted in Figure 9 and Figure 10 respectively.

As the first step toward extracting a template from the time series, a three dimensional embedding of a chaotic trajectory is created out of the scalar amplitude measurements made by computer analysis of the

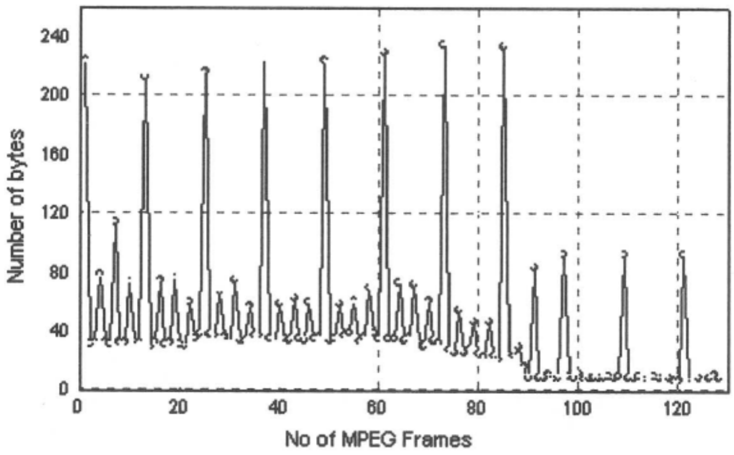


Figure 2.9. An example of bit-rate in MPEG system

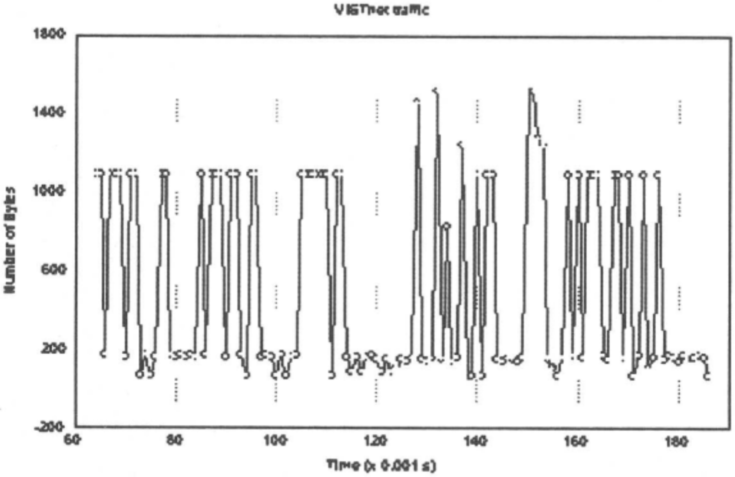


Figure 2.10. Aggregate bit-rate in VISTAnet system

traffic traces. This is accomplished via a time delayed embedding of the original data set. The offset for the delay was determined by the mutual information criterion [4]. Values for the offset range from 1 to 12 frames and from 1 to 100 ms for MPEG and LAN-LAN traces respectively. Figure 11 and Figure 12 show the trajectories that are realised after a torus-doubling route to chaos.

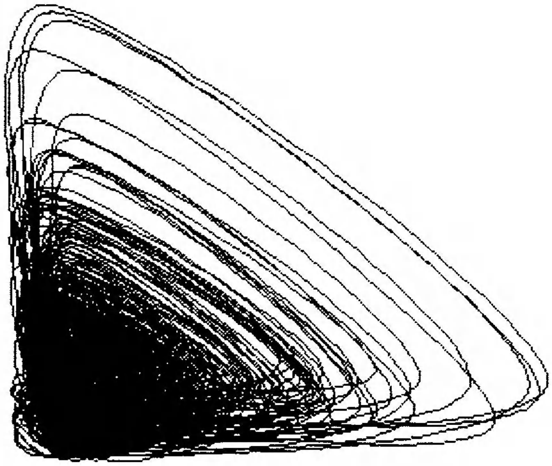


Figure 2.11. Attractor of MPEG data stream

Information-dimension calculations have been performed for the attractors presented in Figures 11 and 12. In each case a maximum of 1000 × 1000 equally sized gird boxes were used to cover the attractor. Fractal dimension calculations of these time series indicate a dimension 2.35 (MPEG) and 2.48 (LAN-LAN). The relevant point is that information dimension of both attractors is clearly well above 2. Thus we conclude that both attractors are chaotic. In conclusion, the plots as well as information dimension calculation provide compelling evidence that we have observed a strange chaotic attractor in both telecommunication systems under investigation. We have also calculated the Lyapunov exponents for both data traces. Calculating the Lyapunov exponents from an experimental data set is a notoriously difficult procedure. This is particularly true with regard to the negative exponents. In general the best that one can hope for is to calculate exponents that are consistent with itself and whatever additional facts are known about the dynamics. The method we use to calculate the Lyapunov exponents from the experimental data is a *Minor Variation Method* (MVM) [8]. MVM uses polynomials to map local neighbourhoods on the attractor into their time evolved images. In both cases (MPEG and LAN-LAN) the total number of vectors used to form the attractor is $N = 16\ 000$. Therefore, the size of the local neighbourhoods used for the polynomial fitting is essentially the same for all tests. To obtain numerical values of the Lyapunov exponents we first averaged the calculated values of λ over the different initial conditions and then averaged that value over the order of the fit for all fits greater than 2. Finally, we find $\lambda_{MPEG} =$ - 0.702 and $\lambda_{LAN} =$ - 0.655.

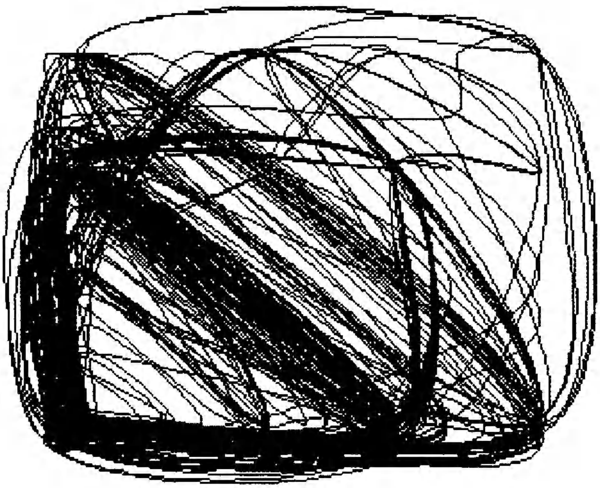


Figure 2.12. Attractor of teletraffic from VISTAnet system

In Figure 13 and Figure 14 we have plotted the *Fourier amplitude spectra* (FAS) of the two processes under investigation. The spectra shown were calculated using a Parzen window; however, we found the spectra density to be unaffected to the particular choice of window function. Clear power law behavior like $1/f^{\alpha}$ on the frequency f is seen in the FAS from Figure 14 (white line). We find that in this case α=1.8. In Figure 13 the power law may not be observed, because MPEG data stream has strongly periodic structure (I, P and B frames).

Figures 15 and 16 show the periodogram plots of both MPEG and LAN-LAN traffic traces. It can be seen for MPEG data stream that for a range of intermediate time scales, the plot shows very little changes before entering the asymptotic regime. This feature of periodogram plot suggests the presence of strong short-term correlation in the data, which makes MPEG traffic only "asymptotically" self-similar. In contrast, data traffic (Figure 16) shows essentially the same structure of periodogram plot over all time scales. This process can be modelled over time scales of engineering interest as exactly self-similar process.

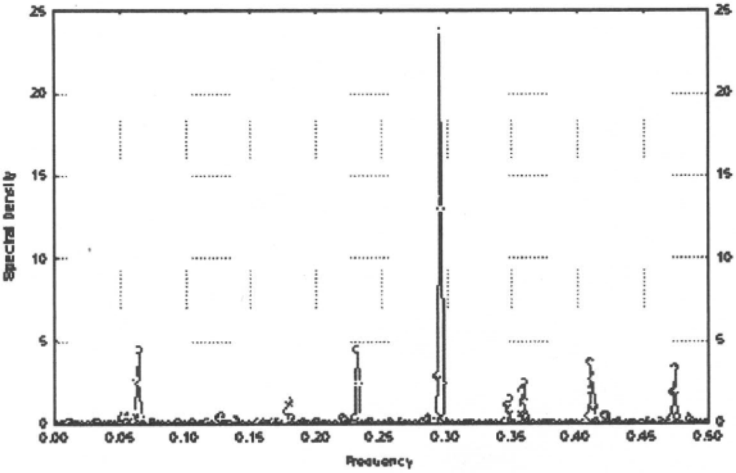


Figure 2.13. Spectral density of MPEG data stream

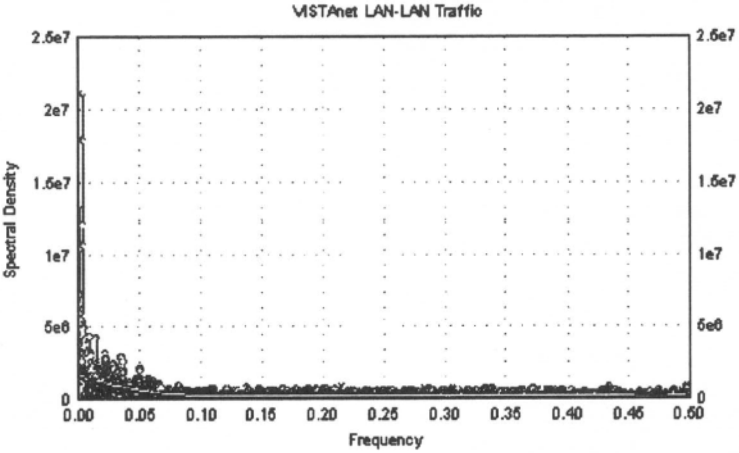


Figure 2.14. Spectral density of LAN-LAN traffic

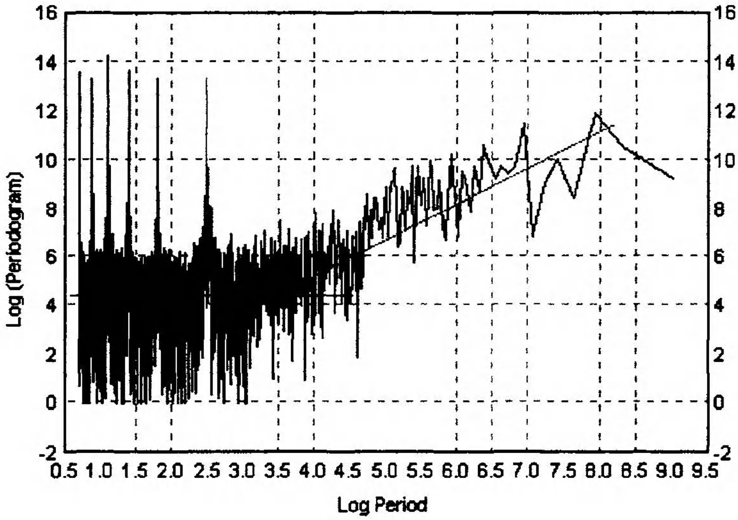


Figure 2.15. Periodogram of MPEG data stream

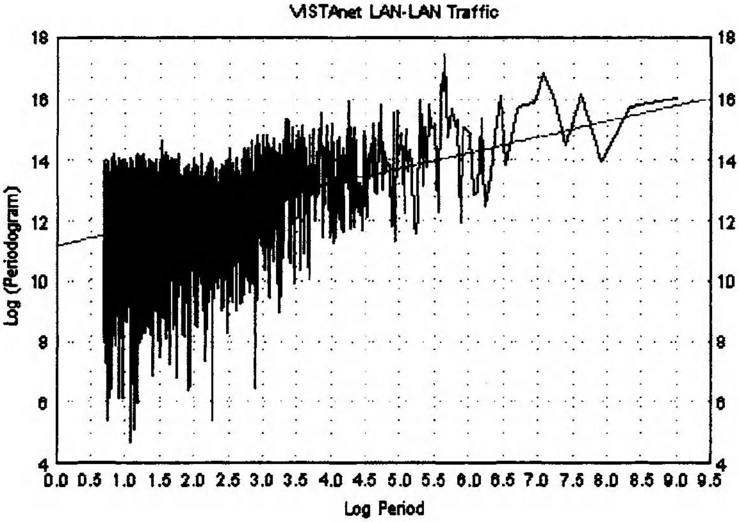


Figure 2.16. Periodogram of LAN-LAN traffic

4. CONCLUDING REMARKS

In the past few years, there has been considerable work in the rapidly growing area of fractal traffic description and modelling, driven by high resolution traffic measurement studies showing the existence of fractal features in packet traffic. Currently, the only measurements that are typically supported by packet switching and network operations systems are rate measurements over coarse time scales of the order 16 - 60 minutes. Such measurements capture the quantity or volume of traffic, but not the quantity of burstiness. It may be shown that a three parameter description of traffic (rate, the Hurst parameter and a peakedness parameter) is required to address many of the engineering problems of interest, such as buffer sizing, setting safe operating points etc. In principle, these parameters can be estimated from special study operational measurements that collect time series of packet count e.g. 1 second counts for a 15 minute period. The disadvantages of this approach and fractal analysis in general, are the difficulties in estimating fractal dimension and scaling region. This lead to the fact that this method does not yield a very robust model and could be relatively less accurate because of the mentioned difficulties. The greatest advantage of this approach is that, if it is possible to obtain the relevant fractal model, the estimation phase is greatly simplified when compared to other methods, and the sampled time scales become almost irrelevant. In particular, the self-similarity in packet traffic can be exploited to reduce measurement overhead. For example, if the traffic is known to be fractal, it is not necessary to have very fine time scale measurements; the relevant traffic parameters can be estimated from coarser measurements. Finally, the results of fractal traffic measurements are already being applied in the development of suitable traffic management methods than can be supported in practice. This is nevertheless a very young field, with considerable scope for innovative research addressing practical problems of relevance.

References

[1] L. Block et. al. *Global Theory of Dynamical Systems,* Lecture Notes in Mat. vol 819 Springer - Verlag, New York, 1980

[2] D. Campbell (Ed.). *Order in Chaos* North-Holland, Amsterdam, 1983

[3] R. Candy. *Signal Processing: The modern approach* McGraw-Hill, New York, 1988

[4] R. Devaney. *An Introduction to Chaotic Dynamical Systems* Addison-Wesley, Redwood City, 1989

[5] A. Erramilli et al. *Engineering for Realistic Traffic: A Fractal Analysis of Burstiness* in Proc. of ITC Special Congress, Bangalore, India, 1993

[6] A. Erramilli et al. *Experimental Queueing Analysis with Long-Range Dependent Packet Traffic* Trans. on Networking, vol.4, no.2, April 1996

[7] A. Erramilli et al. *Recent developments in Fractal Traffic Modeling* in Proc. St. Petersburg Inten. Teletraffic Semin., 1995

[8] F. Family (Ed.). *Dynamics of Fractal Surfaces* New Scientific, Singapore, 1991

[9] S. Feit. *Characteristic exponents and strange attractors* Commun. Math. Phys. vol.61, 1978

[10] M. Figenbaum. *Some Characterisations of strange sets* J. Stat. Phys. vol.46, 1987

[11] P. Grassberger. *Measuring the strangeness of strange attractors* Physica vol.9D, 1983

[12] C. Grebogi. *Critical exponents of chaotic transient in nonlinear dynamical systems* Phys. Rev. Lett. vol.37, 1986

[13] D. Heyman. *Statistical analysis and simulation study of video teleconference traffic in ATM networks* IEEE Trans. Circ. Syst. vol.2, 1992

[14] M. Hênon. *A two-dimensional mapping with a strange attractor* Commun. Math. Phys., vol.50, 1976

[15] F. Kishino. *Variable Bit-Rate Coding of Video Signals for ATM Networks* IEEE Selected Area in Commun., vol.7, 1989

[16] W. Leyland. *High time-resolution measurements and analysis of LAN traffic* in Proc. Infocom'91, Bal Harbour, 1991

[17] E. Lorenz. *The local structure of a chaotic attractor in four dimensions* Physica vol.13D, 1984

[18] B. Mandelbrot. *Self affine fractals and fractal dimension* Phys.Scr.vol.32, 1985

[19] D. Murray. *Mathematical Biology* Springer-Verlag, New York, 1989

[20] I. Norros. *A storage model with self-similar input* Queueing System Theory and Applications, vol.16, 1994

[21] T. Parker. *Practical Numerical Algorithms for Chaos Systems* Springer-Verlag, New York, 1989

[22] E. Peters. *Chaos and Order - A New View of Cycles* J. Wiley & Sons, New York, 1991

[23] H. Schuster. *Deterministic Chaos: An Introduction* VCH, Bonn, 1988

[24] N. Trufillaro et al. *An experimental approach to nonlinear dynamics and chaos* Addison-Wesley, Reading, 1992

[25] D. Veith. *Novel methods of description of broadband traffic* in Proc. 7th Australian Teletraffic Research Seminar, Murray River, Australia, 1992

Chapter 3

ADAPTIVE STATISTICAL MULTIPLEXING FOR BROADBAND COMMUNICATION

Timothy X Brown
Dept. of Electrical and Computer Engineering
University ofColorado, Boulder, CO 80309-0530
timxb@colorado.edu

Abstract Statistical multiplexing requires a decision function to classify which source combinations can be multiplexed through a given packet network node while meeting quality of service guarantees. This chapter shows there are no practical fixed statistical multiplexing decision functions that carry reasonable loads and rarely violate quality of service requirements under all distributions of source combinations. It reviews adaptive alternatives and presents statistical-classification-based decision functions that show promise across many distributions including difficult-to-analyze ethernet data, distributions with cross-source correlations, and traffic with mis-specified parameters.

Keywords: Asynchronous Transfer Mode, Quality of Service, Admission Control, Statistical Multiplexing, Adaptive Methods.

1. INTRODUCTION

Modern broadband services transport diverse sources—constant bit rate voice, variable-rate video, and bursty computer data—using packet-based protocols such as the asynchronous transfer mode (ATM). In Figure 3.1, packets arrive at a node from different sources and are multiplexed on an output link. Since the many traffic sources are uncoordinated and communication bandwidth is finite, links congest and the link introduces losses and delays. With enough congestion, delays grow, queues overflow, and service degrades. Unlike *best-effort* services such as the internet, we consider the case where traffic sources are given *quality of service* (QoS) guarantees. To be specific, this work focuses on packet-level QoS such as on the maximum delay, delay variation, or loss rate, rather than call-level QoS such as call blocking rates.

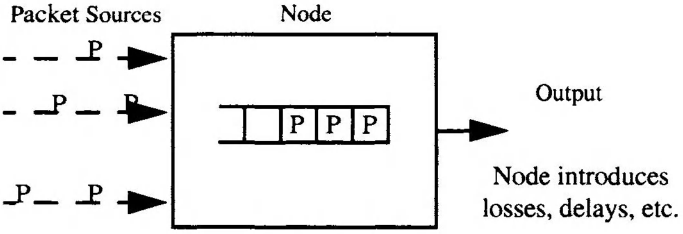


Figure 3.1 Network Node as a Black Box.

Providing packet-level QoS guarantees in broadband networks is a broad area of intense research (see [Gue99] and [Kni99] for an overview and extensive bibliography). While many aspects of QoS must be addressed, we would often like to answer a simple question: Under what conditions can a network meet a QoS guarantee. This chapter argues for new approaches to answering this question.

Standard multiplexing avoids congestion by rejecting a source combination if the total maximum source transmission rate would exceed the link bandwidth (so-called *peak-rate* multiplexing). This works well with constant bit rate sources. Variable rate and bursty sources generate packets at different rates over time. When many such sources are combined it is unlikely they all simultaneously communicate at their maximum rate. *Statistical multiplexing* exploits this fact to accept more sources and gain significantly higher utilizations. The key is an accurate *decision function* that classifies what combinations of sources can and can not be statistically multiplexed together on a given link while meeting QoS requirements. Successful statistical multiplexing is central to key tasks in broadband networks. For example, connection admission control avoids congestion by admitting new connections only when the new and existing connections will receive their requested QoS. A statistical multiplexing decision function could evaluate new connection requests for this purpose.

Statistical multiplexing can provide significant gains over peak rate allocation. If the utilization of a source type is low, for instance below 10%, then many such sources could be statistically multiplexed together providing up to 10 times greater network utilization. Deciding exactly how many such sources could be multiplexed together and still meet QoS requirements is part of the decision function design.

Two paths can be taken to developing the statistical multiplexing decision function as shown in Figure 3.2. The first path develops a model of the node function and traffic processes and then reduces this model to a decision function. This we denote the *fixed* method since the decision only applies to the modeled system. It is also fixed since the decision function is typically considered accurate for any combination of sources and therefore the same

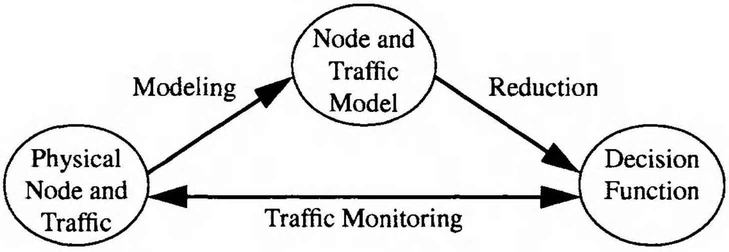


Figure 3.2 Two paths to developing statistical multiplexing decision functions.

without regard to the distribution of source combinations from the modeled traffic processes. The fixed method can fail if either the models are not accurate or if the model do not yield tractable decision functions and compromising simplifications are made.

The second path assumes little about the traffic or node. Many protocols provide monitoring data or network simulations can generate data with samples of traffic source combinations and the observed QoS. Using methods described in this chapter, such samples can be combined directly into a decision function that classifies what combinations do and do not meet QoS requirements without developing any explicit analytical node and traffic models. This we denote the *adaptive* method since the decision can be modified by the actual behavior of the node and traffic. The adaptive decision function is accurate only after observing the network performance and as a result may have an initial period of low accuracy. But, with enough observation, the adaptive method has the potential to learn an optimal decision function.

A simple example will make the distinction clear. Given Poisson arrivals into a finite FIFO buffer with exponential service time (i.e. an *M/M/*1 queue) and a QoS requirement on the maximum blocking probability, the fixed approach would derive the relationship between total load and blocking. A decision threshold would be derived and only loads up to the threshold would be accepted. The adaptive method simply would use examples of the observed loss rate at different loads to set the threshold. The adaptive method applies equally if the arrival process, service time distribution or queueing discipline changed, whereas the fixed approach would do well only on certain models and then only if the model was known.

No particular source model is assumed in Figure 3.1. The sources could be homogeneous or heterogeneous, independent or correlated. No particular node model is assumed in Figure 3.1 either. The queues could be simple FIFO, or implement a more complex scheme such as multiple priority queues or weighted fair queueing. The service rate could be constant or vary over time. Feedback mechanisms may be in place such as for ABR traffic. This chapter treats statistical multiplexing decision functions that apply quite gen-

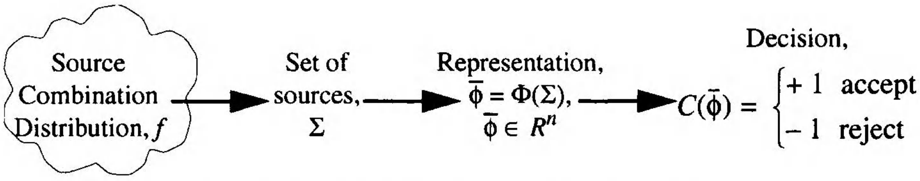

Source Combination Distribution, f
Set of sources, Σ
Representation, $\bar{\phi} = \Phi(\Sigma)$, $\bar{\phi} \in R^n$
Decision,
$C(\bar{\phi}) = \begin{cases} +1 & \text{accept} \\ -1 & \text{reject} \end{cases}$

Figure 3.3 **Relationship of formal elements to problem.**

erally to a wide range of scenarios.

The body of this chapter is divided into four sections. Section 2 is a formal introduction to the statistical multiplexing decision function; the minimum necessary components; and metrics for evaluating the decision function effectiveness. Section 3 argues that any reasonable fixed controller either carries arbitrarily low loads relative to what is possible, is not robust to differing traffic structure, or does not treat artifacts of real networks such as inter-source correlations and misspecified parameters. Section 4 introduces adaptive statistical multiplexing and develops a theoretical foundation. Section 5 presents several experiments with adaptive multiplexing that show it has promise to be both robust and efficient across a variety of node types and traffic distributions including those with inter-source correlations and misspecified traffic parameters.

2. STATISTICAL MULTIPLEXING DECISION FUNCTIONS

The role of the decision function is to answer the question of whether the node can carry a set of sources and meet QoS guarantees. The set of sources need a description called a representation that can be used as inputs to the decision function. The performance of a decision function depends on the environment where it is applied. The environment is defined by the distribution of source combinations that will be seen by the controller. The elements of the statistical multiplexing decision function are shown in Figure 3.3. The rest of this section elaborates on these concepts.

2.1 DECISION MODEL

A source combination, Σ, consists of a number of sources. Such as three MPEG-2 video sources and five 10 BaseT ethernet links. The space for Σ depends on the application and will not be explicitly defined here. It is only necessary that a probability distribution, $f(\Sigma)$, can be defined from the space of possible distributions. Each source can generate packets according to its own traffic process. Since the number of sources is unbounded, the different traffic types vary greatly, and decisions must be made in a reasonable time; source combinations are described by an intermediate feature vector, $\bar{\phi} = \Phi(\Sigma) \in R^n$ for some fixed dimension n. The function, Φ, is the *representation*, with, for example, statistics of Σ such as the total load or the number of sources within different traffic classes.

We define QoS at two levels. At the source level, $Q(\Sigma) = (Q_1(\Sigma), \ldots, Q_l(\Sigma))$ is a vector of l QoS metrics for Σ; for example, the cell loss rate and mean delay for this combination are two possible metrics. With non-homogeneous traffic classes this could be a vector of QoS values for each traffic class. The vector $Q(\Sigma)$ does not describe a particular instance of the node carrying Σ. It is the long term expected QoS metrics for the source combination Σ. In connection access control this is known as conservative control. An aggressive type controller might momentarily allow combinations that violate QoS if averaged over time the system meets QoS. This depends on the dynamics of the problem which we will not consider here, but is considered elsewhere [Mit98, Ton99].

For a given distribution of source combinations, $f(\Sigma)$, and representation, Φ, each feature vector, $\bar{\phi}$, has an associated QoS vector, $q(\bar{\phi}) = (q_1(\bar{\phi}), \ldots, q_l(\bar{\phi}))$ that is the average[1] over the source combinations having feature $\bar{\phi}$, e.g.:

$$q_i(\bar{\phi}) = \text{average of QoS metric } i \text{ at } \bar{\phi} = \frac{\int_{\{\Sigma \mid \Phi(\Sigma) = \bar{\phi}\}} Q_i(\Sigma) f(\Sigma) d\Sigma}{\int_{\{\Sigma \mid \Phi(\Sigma) = \bar{\phi}\}} f(\Sigma) d\Sigma}. \quad (3.1)$$

We formulate the QoS requirements in terms of the QoS metric vector and a threshold vector $\tau = (\tau_1, \ldots, \tau_l)$:

$$q_i(\bar{\phi}) < \tau_i \text{ for all } i. \quad (3.2)$$

These notions of QoS are quite general. Most QoS requirements can be put in this form (e.g. if $x(\bar{\phi})$ is the expected delay, a requirement on delay between 1ms and 2ms can be represented by $q_i(\bar{\phi}) = |x(\bar{\phi}) - 1.5|$ and $\tau_i = 0.5$ or by $q_i(\bar{\phi}) = x(\bar{\phi})$, $q_{i+1}(\bar{\phi}) = -x(\bar{\phi})$, $\tau_i = 2$, and $\tau_{i+1} = -1$).

Having defined the representation and QoS, we turn to the decision function. The decision function can be treated as a classifier, $C(\bar{\phi}) \in \{+1, -1\}$, that classifies which $\bar{\phi}$ meet and don't meet QoS requirements. If $C(\bar{\phi}) = -1$ we say the classifier *rejects* the source combination, otherwise it *accepts* it. The optimal classifier accepts a source combination if and only if $q(\bar{\phi})$ meets the QoS requirements (3.2). Noting $q(\bar{\phi})$ is implicitly a function of the source distribution, $f(\Sigma)$, the optimal decision function is defined as:

[1] Other criteria could be defined such as the infimum (i.e. worst case) QoS of all Σ that have $\bar{\phi}$ as a representation.

$$C_f(\bar{\phi}) = \begin{cases} +1 & \text{if } q(\bar{\phi}) \text{ meets all QoS requirements} \\ -1 & \text{o.w.} \end{cases} . \tag{3.3}$$

To be clear, the optimal classifier depends on the space of sources, their distribution, the representation, the QoS metrics, and the QoS requirements. It also depends on the source traffic processes, the interaction of the traffic processes, and the functionality of the network node. The traffic processes and node functionality are fixed but not necessarily known. Their effect is captured by the QoS, $q(\bar{\phi})$.

2.2 PERFORMANCE METRICS

Ideally the decision function is defined by (3.3). How does a given classifier, C, compare with Cf? A classifier can misclassify by rejecting combinations that would have been accepted by the optimal classifier, or by accepting combinations that do not meet QoS. The first reduces the utilization of the network. The second increases the fraction of the source combinations accepted that violate QoS guarantees. We define two performance measures of a given classifier C to capture these notions. Each of these measures is defined in terms of a specific source distribution, or independent of the source distribution.

Let $U(\Sigma)$ be the utilization of the node output link with source combination Σ. The utilization can be defined quite generally as carried load, generated revenue, etc. The distribution dependent efficiency is

$$E_f(C) = \frac{\text{avg. utilization with } C}{\text{avg. utilization with } C_f} = \frac{\int_{\{\Sigma | C(\Phi(\Sigma)) = +1\}} U(\Sigma) f(\Sigma) d\Sigma}{\int_{\{\Sigma | C_f(\Phi(\Sigma)) = +1\}} U(\Sigma) f(\Sigma) d\Sigma} . \tag{3.4}$$

$E_f(C) > 1$ is possible if the classifier accepts source combinations that do not meet QoS requirements. If numerator and denominator are both zero then by definition $E_f(C) = 1$. For a given distribution, a classifier can have a high efficiency if it rarely rejects source combinations that meet QoS or the utilization of rejected source combinations is low. The distribution free efficiency,

$$E(C) = \inf_f \{E_f(C)\}, \tag{3.5}$$

is the worst case efficiency over all source distributions.

The fraction of correct accepts for a given distribution is

$$R_f(C) = \frac{\text{pr. } C \text{ correctly accepts}}{\text{probability } C \text{ accepts}} = \frac{\int_{\{\Sigma | C(\Phi(\Sigma)) = C_f(\Phi(\Sigma)) = +1\}} f(\Sigma) d\Sigma}{\int_{\{\Sigma | C(\Phi(\Sigma)) = +1\}} f(\Sigma) d\Sigma} \tag{3.6}$$

One minus $R_f(C)$ is how often the classifier falsely accepts a source combination and violates QoS requirements. If numerator and denominator are both zero, then $R_f(C) = 1$. The robustness,

$$R(C) = \inf_f\{R_f(C)\}, \tag{3.7}$$

is the fraction of correct accept decisions in the worst case.

This section emphasizes statistical multiplexing decision functions classify a feature space which is defined via a representation function on the space of source combinations. The performance of this classification is defined by the types of errors relative to the source combination distribution. Since this distribution may not be known a priori, we have also considered the worst case performance over all distributions.

3. FIXED DECISION FUNCTIONS

There exist many proposed QoS decision functions either explicitly or implicitly in terms of admission control or equivalent bandwidth strategies. Typically these are based on assumed models of the traffic and node function. These we denote as fixed decision functions because they are designed to apply to any distribution of traffic from a given class of traffic models. Formally, we define a fixed decision function as a classification function that depends on the space of possible source combinations, the representation function, the source traffic processes, and node function; but is independent of the distribution of source combinations. This section focuses on the distribution of sources and source types and the representation function. For any decision function there exists source distributions for which the method is optimal in the sense of having maximal utilization relative to what is possible (efficiency) and correctly classifying the source combinations realized from this distribution (robustness). This section demonstrates that for any reasonable fixed decision function there exist source distributions for which the function either has zero efficiency or zero robustness. This argument is theoretical. We also show that in practice, fixed decision functions are fragile to realistic variations from typical assumed models.

3.1 THE REPRESENTATION

This section discusses the representation's role in the fixed classifier's efficiency and robustness. A representation, $\Phi(\Sigma)$, is *separable* if for all Σ_1 and Σ_2 where $\Phi(\Sigma_1) = \Phi(\Sigma_2)$: either both $Q(\Sigma_1)$ and $Q(\Sigma_2)$ meet or both do not meet QoS requirements. Appendix A shows if a representation is not separable, then there is always some distribution of source combinations that has either zero efficiency, or only accepts source combinations violating QoS requirements (zero robustness). Therefore, a good classifier over many source distributions requires a good representation.

One might ask if separable representations exist. At one extreme, if $\Phi(\Sigma_1) = \Phi(\Sigma_2)$ only if $\Sigma_1 = \Sigma_2$ then by definition Φ is separable. As an example that this is always possible, define every source by listing every packet's arrival time in order starting with the first packet. By mathematical artifices such as a diagonal counting of these arrival times, and interleaving of the decimal digits of these arrival times a single (albeit infinite precision) real number can uniquely represent any source combination.

At the other extreme $\Phi(\Sigma) = 1$ if Σ meets all its QoS requirements $\Phi(\Sigma) = 0$ otherwise is also separable: that is, the representation function is the decision function. In any real system the representation lies between these extremes and furthermore is fixed by existing protocol or hardware limitations. The next section will look at typical representations and show they are not separable. Further, several examples will make clear the resulting low efficiency or robustness is likely to be observed in practice.

3.2 CONVENTIONAL REPRESENTATIONS

The most robust decision function is based on simply the peak rates of the sources. Others such as the stationary (zero-buffer) approach in [Gue91] use both the peak and average rate. Although for CBR traffic they are optimal, the efficiency is zero or low with bursty video and data sources [DeP92]. For example, with the Ethernet data of Figure 3.7 the utilizations are much less than 1% for a 10Mbps link bandwidth. If the link bandwidth was any value less than 10Mbps all of these sources would be rejected under peak rate and efficiency would be zero. Thus, as is well known, peak rate allocation does not yield an efficient classifier. The stationary approach is asymptotically (in the number of sources) optimal. While asymptotically it is optimal, for small numbers of sources it is either equivalent to peak rate (for high utilization sources) with its low efficiency, or the approximation assumption on which it is based is violated and it accepts too many sources (for low utilization sources) resulting in low robustness.

The traffic descriptors specified by the ATM Forum include peak and average rate as well as a measure of burstiness. Unfortunately these equally represent a wide range of traffic types that have varying effect on network performance [Gal95]. Earlier work to include burstiness assumed traffic from the ON/OFF model of Figure 3.4 assuming exponential holding times (so the model is Markovian) [Gue91][Elw93][Cho94].

It has been well documented that the ON/OFF model with exponential

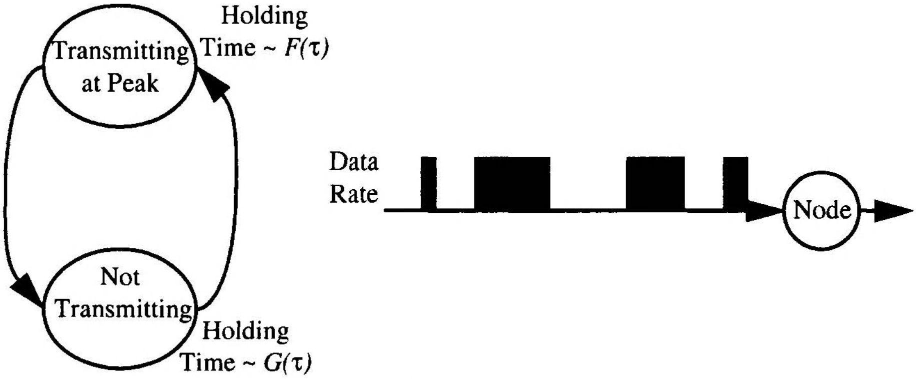
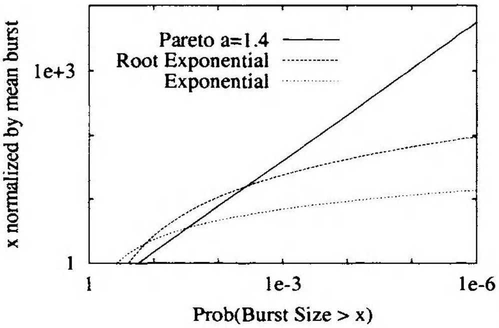


Figure 3.4 Two-State Source Traffic Model

holding times does not reflect the fundamental characteristics of traffic. Analysis of ethernet traffic [Lel93][Nor94] and variable rate video [Gar94] indicate such traffic types are decidedly not simple Poisson processes. Nor are they Poisson burst processes with geometrically distributed burst size. Typically the ON or OFF holding times are characterized by heavy tails with respect to the exponential distribution, i.e., outlier events occur with greater frequency than predicted by the exponential. TCP/IP traffic has also been analyzed and although the session arrivals (telnet, ftp, etc.) are well modeled as a Poisson Process, the traffic within the sessions also exhibits heavy tailed properties [Pax94]. The ethernet data, for instance, has been shown to have interarrival times with finite means and infinite higher moments. Even bounded packet sizes lead to heavy tailed distribution on the number of arrivals in a given period [Kri95]. A heavy tailed interarrival time implies a few very long periods offset by many short periods. So, even though the individual packets are bounded they tend to come in "trains" of packets one after the other followed by long "inter-train" periods.

The model of Figure 3.4 with exponential (or its discrete equivalent, geometric) holding times is completely represented by three components; the ON rate, the mean ON time, and the mean OFF time. Models relying on this representation fail since it is easy to construct distributions that have the same representation, but yet fail to meet the QoS requirements. For instance, as shown in Figure 3.5, an ON/OFF source with exponential holding times will rarely have a burst greater than 13 times the mean bursts (i.e. with probability ~1 in a million), but using Pareto holding times with parameters determined in [Wil95] to match ethernet data, bursts 100's of times longer than the mean occur often. More standard distributions such as the root exponential[2] also have much longer bursts in the tails.

The effect on traffic can be seen in Table 3.1 which uses the Markov model based technique in [Gue91] to calculate the highest load of four



Distribution	P{period length > x}	Variance
Exponential	$e^{-x/m}$	m^2
Root Exponential	$e^{-\sqrt{(2x)/m}}$	$5m^2$
Pareto	$\left(\frac{x\alpha}{m(\alpha-1)}\right)^{-\alpha}, x > \frac{m(\alpha-1)}{\alpha}$	∞ $(\alpha < 2)$

Figure 3.5 Comparing exponential, root exponential, and Pareto holding times (average burst size is m).

sources that can be carried on a given link.[3] Using this load in the model in Figure 3.4 and the node model in Appendix B, 2x10^{10} packet time periods are simulated with different holding time distributions. The statistical multiplexing gain (carried load over the greatest load carried with peak rate) and the net loss rate are shown. Since the ON and OFF periods have the same mean period, the utilization is 50% and the greatest multiplexing gain is 2. With short bursts compared to the buffer size, a large multiplexing gain (out of a possible gain of 2) is possible with the geometric source. With long bursts, only 3% more traffic than allowed by peak rate is accepted. Even with this small deviation

[2] The root exponential is just a special case of the Weibull distribution with **Prob{period length >** x**} = exp**$\left[-\left(\frac{(\Gamma(1+b))}{m}\right)^{1/b}\right]$ and variance $m^2\left(\frac{\Gamma(1+2b)}{\Gamma^2(1+b)}-1\right)$ for parameter $b > 0$. Figure 3.5 would plot $\frac{-(\log(z))^b}{\Gamma(1+b)}$ vs. z. The exponential and root exponential are $b = 1$ and $b = 2$. By choosing large enough b the variance and tail can be made arbitrarily large although unlike the Pareto, the tail plot of Figure 3.5 would always be sub-linear.

[3] This experiment could be repeated with many more than four sources. Four were used for simplicity.

from peak rate allocation, the Pareto distribution packet loss rate is still many orders of magnitude higher than the 10^{-6} target loss rate.

Section 3.1 argues a better representation is needed to perform better. For instance, the Hurst parameter is one candidate that would differentiate traffic models with infinite holding time variance (e.g. the Pareto) from finite variance distributions [Err96]. As seen in Table 3.1, the geometric and root exponential (both producing Hurst parameter 0.5) have dramatically different loss rates. Another approach taken in [Hey96] attempts to characterize video traffic via a general Weibull distribution and concludes a single model based on a few physically meaningful parameters that applies to all video sequences does not seem possible.

Simulation studies in [Kni99] on a range of fixed decision techniques concludes that simple representations will yield low utilizations for bursty traffic flows, while more detailed representations put undo burden on network clients and policing.

It should be clear from these results that given any simple traffic statistics, a wide range of traffic streams can be generated having these statistics. Conversely and more significantly, given a wide range of realistic traffic it is unlikely a representation consisting of a single set of simple statistics will be separable. Therefore, low efficiency or robustness can be expected when fixed decision functions based on such representations are used either across many traffic streams or for extended periods when new usage and new applications can alter the fundamental structure and distribution of traffic.

Table 3.1: $2\text{x}10^{10}$ time slot simulation using Markov-based equivalent bandwidth allocation

Mean Period On/ Off	Multiplexing Gain	Source Distribution	Loss Rate (10^{-6} target)
100/100	1.75	Geometric	$5\text{x}10^{-8}$
		Root Exponential	$4\text{x}10^{-3}$
		Pareto ($\alpha = 1.4$)	$2\text{x}10^{-2}$
10000/10000	1.03	Geometric	$4\text{x}10^{-9}$
		Root Exponential	$3\text{x}10^{-5}$
		Pareto ($\alpha = 1.4$)	$2\text{x}10^{-5}$

3.3 DIFFICULTIES IN REAL NETWORKS

None of the statistical multiplexing methods known to this author specifically addresses the problem of correlated sources. Instead all sources are assumed independent. Two high-rate sources could be synchronized identical outputs violating this assumption as in a three-way video conference where the traffic from one participant to the other two may be sent as two identical streams that have partially overlapping paths. As a trivial extreme, even peak rate allocation can produce losses if the buffer size is less than the number of sources and the sources generate their packets simultaneously (despite the “Asynchronous” in ATM, sources can potentially be highly correlated).

Most methods assume the traffic descriptors are accurate. Due to traffic shaping by the network, bursts from independent sources can become coupled [Lau93], and even CBR sources may enter a node in bursts [Lee96]. For sources with long-range dependencies, such as ethernet traffic, the measured average traffic rate varies widely from one averaging period to the next on averaging periods ranging up to 100’s of seconds [Lel93]. This indicates accurate traffic measurements are not possible. Similarly, many sources may not have any traditional measure of how to describe the traffic other than crude limits and broad classes despite having useful information e.g. “residential world-wide-web surfer from 28.8 baud modem.” Within the framework of Figure 3.3 these practical difficulties are reduced in (3.1) to asking whether the average over all source types with representation $\bar{\phi}$ will meet QoS.

These results suggest the solution to good statistical multiplexing decision functions is not strictly better modeling or better representations, but rather a method that is optimal for the given representation and robust to deviations from the model on which the representation is based.

4. ADAPTIVE METHODS

Adaptive schemes allow the decision function to depend on the results of carried traffic performance. They have the potential to make decisions that vary according to the source distribution, *f*. For example, if only one source combination is possible (as in the proof in Appendix A), a reasonable adaptive method would learn whether the combination should be rejected or accepted. This section describes the steps and elements to this adaptation. It is derived from the formalism of statistical function approximation [Dud73][Bis95] rather than adaptive control.

4.1 OVERVIEW AND EVALUATION CRITERIA

The adaptive method collects a performance data set, $X = \{(\bar{\phi}_i, \bar{\mu}_i)\}$, where $\bar{\phi}_i$

is the feature vector, $\bar{\mu}_i$ is a real vector of monitoring information, and $1 \leq i \leq |X|$ ($|X|$ is the number of elements in X). A sample in this data set represents the output of monitoring hardware with the measured performance, $\bar{\mu}_i$, from a source combination with feature vector, $\bar{\phi}_i$. As before, the source combinations are distributed according to f. The monitoring information, $\bar{\mu}_i$, might be as simple as 1 or –1 depending on whether or not the source combination met its QoS. Or, it might be more detailed, like the number of packets sent and the number of packets lost, delay statistics, etc. The classification function derived from the data set X is denoted as $C_X(\bar{\phi})$.

We specify two criteria for $C_X(\bar{\phi})$. The first, *consistency,* is an asymptotic property:

$$\lim_{|X| \to \infty} \text{Prob}\{C_X(\bar{\phi}) \neq C_f(\bar{\phi})\} = 0. \tag{3.8}$$

The probability is over source combinations chosen from f. This says that as we collect more data the probability of decision error goes to zero. The second is a finite sample property that requires the fraction of correct accepts to be greater than a confidence level Λ:

$$R_f(C_X) > \Lambda \tag{3.9}$$

An easy way to satisfy (3.9) is to simply not accept any source combination so by definition $R_f = 1$. When X has few samples, this may be a viable strategy. The requirement in (3.8) ensures that with more data, the classifier converges in probability to the true classifier and $R_f(C_X) = E_f(C_X) = 1$. Since this applies for any f, we conclude a classifier satisfying (3.8) will yield $R(C_X) = E(C_X) = 1$ as more data is collected. With limited data, (3.9) bounds QoS violations to probability less than $1 - \Lambda$.

Is any adaptive decision function consistent? Can any adaptive decision function give confident estimates with finite samples? In general, the answer to both questions is yes. Appendix C discusses these questions further.

4.2 PRIOR WORK

A variety of researchers have examined the adaptive approach. The methods in [Che92], [Nev93], and [Nor93] choose a particular traffic model and then refine parameters based on the controller's performance. These are based on Markov source models and thus suffer the same deficiencies as noted in the previous section when traffic is non-Markovian. The method in [Jam92], while adaptive, controls only for delay. As noted in [Lev97], delay is a much more stable parameter than packet loss rate. Also, as is the method in [Kaw95], it is based on short term traffic measurements which can be misleading for data with long-range dependencies [Pax94]. The methods in [Hir90],[Tra92], and [Est94] are closest to the method presented here. They do not assume a particular traffic model in developing the decision function and can choose long time-scales to adapt over (e.g. days or weeks). While promising, they are applied to very simple models (e.g. combinations of different numbers of only one or two source types). It is not clear how the methods would scale to many heterogeneous copies of source types. Further, as will be elaborated shortly, these approaches have a distinct bias that under real-world conditions leads to accepting source combinations that miss QoS targets by orders of magnitude. Incorporating preprocessing methods to eliminate this bias is critical and two methods from prior work by the author will be described. Unlike this previous work, the methods are applied to a range of source models, difficult-to-model ethernet traffic, sources with intra-source correlations, and sources with misspecified parameters.

4.3 STATISTICAL CLASSIFICATION

The adaptive methods in this chapter generate a decision function using statistical classification methods. The reader is directed to any of a number of books in the statistical classification area often under the label of "pattern recognition" or "neural networks". Two good examples are [Dud73] and [Bis95]. Using statistical classification, the decision function is derived from examples of previously carried sources and their received QoS. A statistical classifier is given a *training set*, $Y = \{(\overline{\phi}_i, d_i, w_i)\}$, consisting of feature vectors, $\overline{\phi}$, with corresponding desired output classification, $d_i \in \{+1, -1\}$ and positive real-valued sample weight, w_i. For many applications all samples are weighted equally and the weight is disregarded. A classification function, $C(\overline{\phi};\overline{v})$, parameterized by a real-valued vector $\overline{v}$, divides the feature space into positive and negative regions separated by a *decision boundary* and can be used to classify future feature vectors. Based on the training set, a classifier (i.e., $\overline{v}$) is selected that minimizes some criteria (so-called *training*). We describe in turn the data collection, decision function model, and objective criteria.

4.4 GENERAL SCENARIO AND COLLECTING DATA

This chapter focuses on the scenario in Figure 3.1 where a node is multiplexing multiple sources. The goal is to collect a data set about the node in the form $X = \{(\overline{\phi}_i, \overline{\mu}_i)\}$ consisting of monitoring information, $\overline{\mu}_i$, for source combinations represented by $\overline{\phi}_i$. The node architecture, feature vector representation, and source distributions are assumed fixed but not necessarily known. In a running network sources dynamically arrive/depart and at transitions network monitors record information about the traffic carried and QoS since the last transition. Alternatively, off-line network simulations of different traffic combinations could be used. These are not equivalent since, off-line, any combination can be simulated without regard to the QoS given, while in an operating network customers care about received QoS. Also in an on-line admission control scenario, the observed distribution of source combinations is a function of what connection requests are accepted or rejected. This issue is discussed in [Hir95] and work in [Bro99a] shows the interaction is stable and does not fundamentally change the problem. All of the results presented in this chapter are based on simulated source combinations.

For simplicity the rest of this chapter focuses on a single QoS criterion. Since, as noted earlier, packet loss is the most difficult parameter to control, we focus on a system where the QoS criterion is in terms of a maximum packet loss rate, p^*. It should be clear the general technique applies to any QoS metric. With these disclaimers we assume the samples contain the number of packet arrivals, T, the number of lost packets, s, and the feature vector, $\bar{\phi}$. The underlying QoS, $q(\bar{\phi})$, is the average loss rate of source combinations with representation $\bar{\phi}$. The data given is $X = \{(\bar{\phi}_i, s_i, T_i)\}$ and the training set is $Y = \{(\bar{\phi}_i, d_i, w_i)\}$ where $|X| = |Y|$. How is Y computed from X, what is the form of the objective function that will be minimized, and what decision function model will be used? We look at the latter question first.

4.5 DECISION FUNCTION MODEL

Given a training set, Y, the decision function, $C(\bar{\phi};\bar{v})$, can use many models [Bis95]. The basic consideration is the so-called bias-variance trade off. To illustrate this trade-off we consider two extremes. At one extreme the decision function is unconstrained; for instance, $C(\bar{\phi};\bar{v})$ is an arbitrarily large look-up table. This table stores a value for every unique $\bar{\phi}$ in Y and $C(\bar{\phi};\bar{v})$ is simply the value stored at $\bar{\phi}$. While this can always produce an unbiased estimate for each $\bar{\phi}$ in Y, it is not defined for other $\bar{\phi}$ and the output will have a high variability from one Y to the next. At the other extreme the decision function is highly constrained; for instance, $C(\bar{\phi};\bar{v})$ is a constant. If the output is always accept or always reject then this is a sufficient model. In more interesting cases, the output will not be a constant. On the other hand the output will vary little from data set to data set and have low variance.

Since the decision errors can be decomposed into bias and variance components selecting a decision function is a trade-off between models that can capture the correct decision function, and models with few parameters that can be trained quickly with small data sets. We highlight this issue to show that selecting a model requires some care and including prior knowledge about the type of decision function we expect is more likely to produce simple efficient and robust decision functions.

In order to focus on the central issues of this chapter, we use a simple model; the linear discriminant:

$$C(\bar{\phi};\bar{v}) = \operatorname{sign}\left(\sum_{i=1}^{n} v_i\phi_i + v_0\right). \tag{3.10}$$

The parameters, $\bar{v}$, are determined by minimizing an objective (next section) with respect to the weights. For the experiments in this chapter the features are loads. The optimal classifier is monotonic in that if $\bar{\phi}$ is rejected, then any feature with greater load is rejected. The linear decision function also has this property. If there is only one feature in the feature vector, then the linear decision function reduces to a threshold on the feature; which is optimal if the feature is a load. By Appendix C, the linear discriminant can form a consistent estimator in this case with the correct objective function. Therefore, although (3.10) is a simple model, it is sufficient for the experiments in this chapter.

4.6 OBJECTIVE FUNCTION

Given a training set, $Y = \{(\bar{\phi}_i, d_i, w_i)\}$, and a classifier, $C(\bar{\phi};\bar{v})$, a set of parameters is chosen that minimizes an objective function. This chapter minimizes a weighted sum squared error. More general criteria also work well [Bro99b]:

$$E(\bar{v}) = \sum_i [w_i(C(\bar{\phi}_i;\bar{v}) - d_i)^2]. \tag{3.11}$$

An unconstrained classifier, will set $C(\bar{\phi};\bar{v}) = d_i$if all the $\bar{\phi}_i$ are different. With multiple samples at the same $\bar{\phi}$, the error in (3.11) is minimized when

$$C(\bar{\phi};\bar{v}) = \operatorname{sign}\left(\sum_{\{i|\bar{\phi}_i = \bar{\phi}\}} w_i d_i\right). \tag{3.12}$$

If the classifier is more constrained (e.g. a low dimension linear classifier) or no data is precisely at $\bar{\phi}$, $C(\bar{\phi};\bar{v})$ will be the weighted average of the d_i in the neighborhood of $\bar{\phi}$, where the neighborhood is, in general, an unspecified function of the classifier. A more direct form of averaging would be to choose a specific neighborhood around $\bar{\phi}$ and average over samples in this neighborhood. This suffers from having to store all the samples in the decision mechanism, and incurs a significant computational burden to find the samples in the neighborhood. More significant is how to decide the size of the neighborhood. If it is fixed, in sparse regions no samples may be in the neighborhood. In dense regions near decision boundaries, it may average over too wide a range for accurate estimates. Dynamically setting the neighborhood so that it always contains the *k* nearest neighbors solves this problem, but does not account for the size of the samples. We will return to this in Section 4.6.2.

4.6.1 The Small Sample Problem

If sample sizes are large ($Tp^* >> 1$), then $d_i = \text{sign}(p^* - s_i/T_i)$ accurately estimates $\text{sign}(p^* - q(\bar{\phi}_i))$ and the problem reduces to fitting a function to $Y = \{(\bar{\phi}_i, d_i)\}$ using standard statistical classification techniques. For example, this approach has been used in [Hir90][Tra92][Est94] and is denoted the *normal* method in Table 3.2. When sample sizes are small, ($Tp^* << 1$), the number of losses, s, will be zero with high probability while even one loss yields a sample estimate, $s/T >> p^*$, so that individual samples are poor estimates of the underlying rate. As will be shown, the above procedure when applied to small samples can accept a $\bar{\phi}$ despite $q(\bar{\phi})$ being orders of magnitude larger than p^*.

An alternative approach in [Hir95] attempts to estimate $q(\bar{\phi})$ directly using s_i/T_i as the measured loss rate at $\bar{\phi}_i$, and then using regression techniques to make an estimate, $\hat{q}(\bar{\phi})$, anddefining $C_Y(\bar{\phi}) = \text{sign}(p^* - \hat{q}(\bar{\phi}))$. The probabilities can vary over orders of magnitude making accurate estimates difficult. Estimating the less variable $\log(q(\bar{\phi}))$ is inconsistent for small samples where most of the samples have no losses and $s = 0$, and the logarithm must be artificially defined. Preliminary work in [Ton98] indicates a proper modeling of the regression problem may lead to satisfactory results that are unbiased and are insensitive to intra-sample correlations (c.f. Section 4.6.3).

One obvious solution is to have large samples. In communication networks, such as packet data networks, sample sizes are limited by three effects. First, desired loss rates are often small; typically in the range 10^{-6}–10^{-12}. This implies large samples must be at least 10^7–10^{13} observed packets. For 10^{13}, even a Gbps packet network with short packets requires samples lasting several hours. At typical rates, samples of size 10^7 require samples lasting minutes. Second, in dynamic data networks, while individual connections may last for significant periods, the aggregate flow of connect and disconnect requests prevents traffic combination for lasting the requisite period. Third, in any queueing system, even with uncorrelated arrival traffic, the buffering introduces memory in the system. A typical sample with losses may contain 100 losses,. But, a loss trace would show the losses occurred in a single short overload event. Thus, the number of independent trials can be several orders smaller than the raw sample size indicating the loads must be stable for hours, days, or even years to get samples that lead to unbiased classification.

4.6.2 Consistent and Confident Training Sets

We present without proof two preprocessing methods derived and analyzed in

[Bro95, Bro99b]. The first chooses an appropriate d and w so (3.12) corresponds to a consistent maximum likelihood solution. This is the *weighting* method shown in Table 3.2.

The second preprocessing method assigns uniform weighting, but classifies $d_i = 1$ only if a certain confidence level, Λ, is met that the sample represents a combination where $q(\bar{\phi}) < p*$. Such a confidence was derived in [Bro99b]:

$$\mathrm{Prob}\{q(\bar{\phi}) < p^* | s, T\} = 1 - B(s, T, p^*) \tag{3.13}$$

where

$$B(s, T, p) = \sum_{k=0}^{k \le s} \binom{T}{k} p^k (1-p)^{T-k}. \tag{3.14}$$

For small T (e.g. $Tp^* < 1$ and $\Lambda > 1 - 1/e$), even if $s = 0$ (no losses), this level is not met. But, a neighborhood of samples with similar load combinations may all have no losses indicating this sample can be classified as having $q(\bar{\phi}) < p*$. Choosing a neighborhood requires a metric, m, between feature vectors, $\bar{\phi}$. In this chapter we simply use Euclidean distance. Using the above and solving for T when $s = 0$, the smallest meaningful neighborhood size is the smallest k such that the aggregate sample is greater than a critical size,

$$T^* = \ln(1 - \Lambda)/\ln(1 - p^*). \tag{3.15}$$

From (3.13), this guarantees that if no packets in the aggregate sample are lost we can classify it as having $p(\bar{\phi}) < p*$ within our confidence level. For larger samples, or where samples are more plentiful and k can afford to be large, (3.13) can be used directly. Table 3.3 summarizes this *aggregate* method.

4.6.3 Generating Samples of Independent Bernoulli Trials

The above preprocessing methods assume the training samples consist of independent samples of Bernoulli trials. Because of memory introduced by the buffer and possible correlations in the arrivals, this is decidedly not true. The methods can still be applied, if samples can be subsampled at every Ith trial where I is large enough so the samples are pseudo-independent, i.e. the dependency is not significant for our application. As indicated in [Gro96],

Table 3.2: Summary of Methods.

Method	d_i	w_i
Normal	$\mathrm{sign}(p^*T_i - s_i)$	1
Weighting	$\mathrm{sign}(p^*T_i - s_i)$	$p^*T_i - s_i$
Aggregate	Table 3.3	1

buffered systems have a finite time horizon beyond which the impact on loss of correlations in the arrival process become nil even for sources with long-range dependencies. We therefore can expect to find a suitable I for most traffic types. An explicit bound on this time horizon appears in [Gro96], which would be a suitable I. This bound has not yet been tried for this chapter. It is expected the bound will be larger than necessary for our purposes. Further, it depends on knowing explicit characteristics of the node and source arrival process that may not be known to the statistical multiplexer.

A simple graphical method for determining I is given in [Bro99b]. The weighting and desired output:

$$\begin{aligned} w_i &= 1 - B(T_i p^*, T_i, p^*) \text{ if } d_i = 1; \; w_i = B(T_i p^*, T_i, p^*) \text{ if } d_i = -1 \\ d_i &= \operatorname{sign}(T_i p^* - s_i) \end{aligned} \quad .(3.16)$$

has the property that with uncorrelated trials it produces a consistent estimator. Alternatively, if T overstates the true sample size by a large factor, $w_i = 0.5$ and (3.16) is the same as the normal scheme. This is the case with correlated samples. The sample size, T, overstates the number of independent trials. As will be shown, this implies the decision boundary is biased to orders of magnitude above the true boundary. As the subsample factor is increased, the subsample size becomes smaller, the trials become increasingly independent, the weighting becomes more appropriate, and the decision boundary moves closer to the true decision boundary. At some point, the samples are sufficiently independent so that sparser subsampling does not change the decision boundary. By plotting the decision boundary of the classifier as a function of I, the point where the boundary is independent of the subsample factor indicates a suitable choice for I.

If the subsample factor is known, the packets can be subsampled explicitly as the data is collected. As will be seen, the subsample factors are large, implying an easy-to-implement sparse monitoring is possible in an on-line system. If the raw $(\bar{\phi}, s, T)$ samples are given, then as a worst case they are

Table 3.3: Aggregate Classification Algorithm

1. Given sample (ϕ_i, s_i, T_i) from training set $\{(\phi_i, s_i, T_i)\}$, choose confidence level Λ and metric m on the feature space.
2. Calculate T^* from (3.15).
3. Find nearest neighbor sequence $n_0, n_1, \ldots$ where $n_0 = i$ and $m(\bar{\phi}_{n_j}, \bar{\phi}_i) \le m(\bar{\phi}_{n_{j+1}}, \bar{\phi}_i)$ for $j \ge 0$.
4. Choose smallest k s.t. $T' = \sum_{j=0}^{k} T_{n_j} \ge T^*$. Let $s' = \sum_{j=0}^{k} s_{n_j}$.
5. $d_i = \operatorname{sign}(1 - \Lambda - B(s', T', p^*))$.

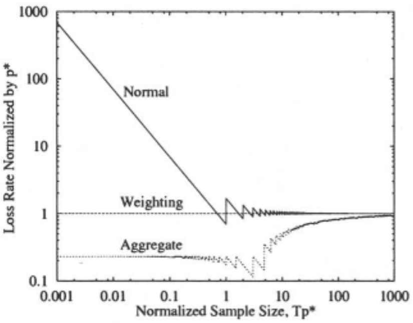


Figure 3.6 Expected Decision Normalized by $p*$. The nominal boundary is $p/p* = 1$. The aggregate method uses $\Lambda = 0.95$.

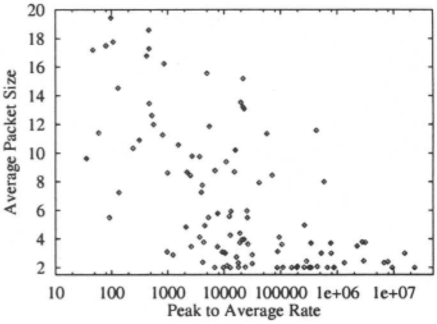


Figure 3.7 Average Packet Size (in 48-byte units) vs. Peak to Average Rate Ratio for Ethernet Data. The peak rate is 10Mbps for all connections.

subsampled by dividing s and T by I. The results are rounded up with probability proportional to the remainder.

In this way, despite the correlations in the data, we can produce independent trials for the statistical classification methods. Looking at Table 3.2, subsampling only scales the weighting by a factor of $1/I$ for all samples. Since this has no essential effect on the minimization of (3.11), the weighting method is independent of these correlations and thus does not need to estimate I.

4.7 SUMMARY AND PRIOR RESULTS

This section has presented a method for using samples of the QoS for different source combinations to be combined into a consistent decision function. The procedure consists of collecting traffic data at different combinations of traffic loads that do and do not meet QoS. These are then subsampled with a factor I determined as in Section 4.6.3. Then one of the methods for computing a training set, summarized in Table 3.2, are applied to the data. This training set is then used in any statistical classification scheme. Analysis in [Bro99b] derives the expected bias (shown in Figure 3.6) of the methods when used with an ideal classifier. The normal method can be arbitrarily biased, the weighting method is unbiased, and the aggregate method chooses a conservative boundary. Simulation experiments in [Bro99b] with a well characterized M/M/1 queueing system to determine acceptable loads showed the weighting method was able to produce unbiased threshold estimates over a range of values; and the aggregate method produced conservative estimates that were always below the desired threshold, although in terms of traffic load were only 5% smaller. Even in this simple system, where the input traffic is uncorrelated (but the losses become correlated due the memory in the queue), the subsample factor was 12, meaning good results required more than 90% of the data be thrown out.

5. EXPERIMENTS

A range of experiments are performed using the node model of Appendix B under different source models. The experiments are not necessarily designed to be realistic, but rather to demonstrate the method under a variety of conditions. Each experiment, used the three methods of Table 3.2 for creating a training set, Y, from the monitoring data, X, based on a QoS requirement of maximum packet lost rate, $p^* = 10^{-6}$, confidence of $\Lambda = 95\%$ for the aggregate method, and the decisionfunction, $C(\overline{\phi};\overline{v})$, is the linear discriminant decision function (3.10). A simple representation, total load, is used in each case.

5.1 SIMULATED SOURCES

In this section, the arrival process consisted of 4 identical ON/OFF sources from the model in Figure 3.4 similar to the experiments in Table 3.1 with equal, short mean ON/OFF periods of size 100 time slots. The training set for a given holding time distribution consisted of at least 10,000 simulations at randomly chosen loads for 10^7 timeslots. The load per source are uniformly distributed between 0.25 (accepted by peak rate) and 0.5 (a net load of 1, i.e. 4 sources times 50% duty cycle times a load of 0.5). The representation, ϕ, is simply the total load.

To create pseudo-independent trials necessary for the aggregate methods, we subsampled every Ith packet. Using the graphical method of Section 4.6.3, the resulting I are shown in column 4 of Table 3.4. The median subsample factor is ~200. The sample sizes ranged up to 10^7 plackets, But, after subsampling by a factor of 200, even for the largest samples, $p^*T < 0.05 \ll 1$.

These results for the geometric and Pareto holding time distributions appear in Table 3.4. A third distribution, labeled correlated, introduces a small cross-source correlation to the geometric distribution where in 10% of the samples, 2 of the 4 sources are 100% correlated. To test the results we use the threshold found by the aggregate method and simulate against the other distri-

Table 3.4: Experiments with Simulated Sources.

Mean Period On/Off	Distribution	Training Set Size	Subsample Factor	Threshold Found		
				Normal	Weighting	Aggregate
100/100	Geometric	15112	60	0.909	0.895	0.884
	Pareto ($\alpha = 1.4$)	10492	520	0.659	0.573	0.567
	Correlated	16791	180	0.907	0.874	0.855

butions for 10^{10} time slots as shown in Table 3.5. When the threshold is tested against the same distribution as was trained, the losses are less than the target $p^* = 10^{-6}$. Comparing the thresholds, the geometric-based threshold carries 50% greater loads than with the Pareto-based threshold. Conversely, if the geometric-based threshold is given Pareto traffic, then the QoS targets are missed by more than 4 orders of magnitude. The Correlated and geometric thresholds differ by only a few percent. Yet, the resulting loss rates vary by orders of magnitude. Conversely, since the loss rates are all within an order of magnitude of the target value when the same distribution is used for testing and training, these thresholds are within a few percent of the optimal. Thus given different distributions, the adaptive method can find a nearly optimal decision function for each. This is evidence the adaptive method gains significant efficiency and robustness.

5.2 ETHERNET DATA

We now turn from simulated traffic data to ethernet traffic traces. An earlier version of this section appeared in [Bro97]. The ethernet data is described in [Lel93] as the August 89 busy hour containing traffic ranging from busy file-servers/routers to users with just a handful of packets. The detailed data set records every packet's arrival time (to the nearest 100μsec), size, plus source and destination tags. From this, 108 "data traffic" sources are generated, one for each computer that generated traffic on the ethernet link. With ATM in mind[4], each ethernet packet (which ranged from 64 to 1518 bytes) is split into 2 to 32 48-byte packets (partial packets were padded to 48 bytes). Depending on the experiment, a link data rate was fixed at 10, 17.5, 30, or 100Mbps. Dividing this by 48x8 = 384 produced a link packet rate. Each ethernet packet arrival time is mapped into a particular time slot in the node model. Since the

Table 3.5: Aggregate Thresholds in Table 3.4 Tested for 10^{10} Timeslots.

Train Distribution	Adaptive On Rate[a]	Multiplexing Gain	Test Distribution[b]:		
			Geometric	Correlated	Pareto
Geometric	0.884	1.77	3×10^{-7}	2×10^{-6}	3×10^{-2}
Correlated	0.855	1.71	0	1×10^{-7}	1×10^{-2}
Pareto ($\alpha = 1.4$)	0.567	1.13	0	0	1×10^{-6}

a. Target Loss Rate = 10^{-6}
b. From 10 Billion Packet Timeslot Simulation

[4] Assuming, for simplicity, the cell payload is 48 bytes.

sources came from the same 10Mbps Ethernet link and only packets without collisions were recorded in the data trace, the traces contain at most one packet being sent at any given time. Decorrelating the sources via random starting offsets, produces independent data sources with the potential for multiple packet arrivals and overloads. Multiple copies at different offsets produces sufficient loads even for bandwidths greater than 10Mbps.

The peak data rate with the ethernet data is fixed, while the load (the average rate over the one hour trace normalized by the 10Mbps peak rate) ranges over five orders of magnitude (see Figure 3.7). Also troubling, analysis of this data [Lel93] shows the aggregate traffic exhibits chaotic self-similar properties and suggests that it may be due to the sources' packet inter-arrival time distribution following an extremely heavy tailed distribution with finite mean and infinite higher order moments such as the Pareto distribution in Figure 3.5.

We divided the data into two roughly similar groups of 54 sources each; one for training and one for testing. To create sample combinations we assign a distribution over the different training sources, choose a source combination from this distribution, and choose a random, uniform (over the period of the trace) starting time for each source. Simulations that reach the end of a trace wrap around to the beginning of the trace. The sources are described by a single feature corresponding to the average load of the source over the one hour data trace. A source combination, Σ, is described by a single variable, ϕ, the sum of the average loads. The source distribution, $f(\Sigma)$, was a uniformly chosen 0–*M* copies of each of the 54 training samples. *M* was dynamically chosen for each experiment so the link would be sufficiently loaded to cause losses. Each sample combination was processed for 3x10^7 time slots, recording the load combination, the number of packet arrivals, *T*, and the number blocked, *s*. The experiment was repeated for a range of bandwidths. The bandwidths and number of samples at each bandwidth are shown in Table 3.6

The thresholds found by each method and an estimate of loss rate at this threshold[5] are shown in Table 3.6. The normal scheme is clearly flawed with

Table 3.6: Ethernet Experiments at Different Link Bandwidth.

Band-width (Mbps)	Number of Samples		Sub-sample Factor	Threshold Found & Loss Rate at Threshold on (train,test) Set relative to $p* = 10^{-6}$		
	Train	Test		Normal	Weighting	Aggregate
10	1569	1080	230	0.232 (10, 4)	0.139 (0.8, 1)	0.105 (0.1, 0.08)
17.5	2447	3724	180	0.415 (20, 30)	0.268 (0.5, 0.9)	0.215 (0.003, 0.4)
30	6696	4219	230	0.508 (7, 40)	0.333 (4, 0.05)	0.286 (0.3, 0.02)
100	1862	N.A.	180	0.688 (10, N.A.)	0.566 (0.5, N.A.)	0.494 (0/N.A.)

losses 10 times higher than p^*, the weighting scheme's loss rate is apparently unbiased with results around p^*, while the aggregate scheme develops a conservative boundary below p^*. To test the boundaries, we repeated the experiment generating source combination samples using the 54 sources not used in the training. Table 3.6 also shows the losses on this test set and indicates the training set boundaries produce similar results on the test data.

Applying the Markov model based method in [Gue91] to this ethernet data (treating each packet arrival as an ON period and calculating necessary parameters from there[6]), all but the very highest loads in the training sets are classified as acceptable indicating the loss rate would be orders of magnitude higher than p^*. If the conservative peak rate is applied, since the original ethernet data rate is 10Mbps, only one source is accepted per 10Mbps of bandwidth. The average source load is 0.0014 and does not increase with increasing bandwidth. The adaptive method takes advantage of better trunking at increasing bandwidths, and carries two orders of magnitude more traffic.

5.3 MIS-SPECIFIED PARAMETERS

The adaptive method as described here does not assign any physical meaning to the representation. Thus any systematic monotonic transformation of the parameters will not change the resulting decisions. For instance, multiplying the feature vector by a and recomputing the classifier results in the parameter vector in (3.10) being divided by α, yielding the same decision boundary.

Random noise added to the parameters would affect the decisions. For instance, the average load could be a measurement over a short period and not be an accurate estimate. To test the method's ability to capture this disturbance[7], the Geometric data from Table 3.4, was randomly divided into two equal sized sets. To the first, a uniform random $[-x, +x]$ was added to the load

[5] To get loss rate estimates at these thresholds, the data set, $\{(\phi_i, s_i, T_i)\}$, were ordered by ϕ and the 20% of the data set's samples below each method's threshold were averaged via $\Sigma_i\, s_i/\Sigma_i\, T_i$. Since accepted loads would be below the threshold this is a typical loss rate. 20% was chosen since smaller percentages had a high sensitivity on their loss estimates (in particular they were not monotonic in the threshold).

[6] This is an unrealistic use of the method in [Gue91] since the data is known to be non-Markovian and, in ethernet, large packets are broken into blocks of less than 1518 bytes so the average ON period estimate was optimistically small. Several variations of ON and OFF period definitions were tried, and none did any better. The point is to show what a typical Markov based technique would predict.

[7] While in general the adaptive methods would be able to capture this effect, the preprocessing methods in Section 4 do not strictly apply since they assume a single underlying loss rate at $\bar{\phi}$ and not a distribution of loss rates. Despite this good results are obtained except for the largest noise values.

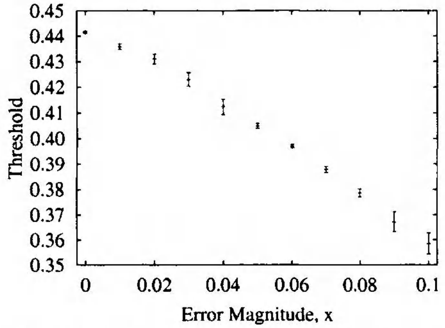


Figure 3.8 Decision threshold as a function of uniform error within [$-x$, $+x$] added to the input feature. Points are mean and one standard deviation of six different splits into test and train sets.

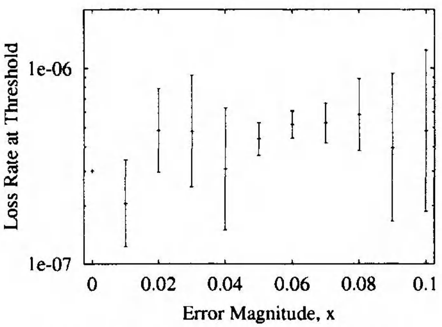


Figure 3.9 Average loss rate of test set samples within [$-x$, $+x$] of threshold ($p^* = 10^{-6}$). Points are geometric mean and one standard deviation of six different splits into test and train sets.

feature. The aggregate method was applied. Future samples at the resulting threshold would have true load uniformly distributed within $\pm x$, so the error rate was computed by averaging all samples in the second set within x of the threshold[8]. The results are plotted in Figures 3.8 and 3.9. As expected, the decision threshold decreased with increasing error magnitude to allow for the noise in the measured loads. Except for the largest noise values, the loss rate was kept below the target, $p^* = 10^{-6}$.

6. CONCLUSION

This chapter introduces a formal notion of statistical multiplexing decision functions for classifying which combinations of traffic will meet or not meet QoS. These functions form the core of admission control and routing algorithms. Theoretically and experimentally it was shown:

1. Fixed decision functions depend on source and node models that do not universally apply resulting in significantly lower efficiency or robustness than an optimal decision function.

2. Adaptive decision functions make few if any assumptions about the source and node models, and can achieve optimal decision boundaries.

For this reason it is argued adaptive schemes should be considered immediately for existing network applications.

[8] The number of underlying simulated timeslots represented by this average was large, ~ $x.1\text{x}10^{11}$. In the no noise case ($x = 0$), the loss rate from Table 3.5 was used.

The adaptive method is optimal *for a given representation and distribution of source combinations.* Thus, it benefits from additional work on source models and their statistical parameters that better represent traffic. It was shown to work well across varied traffic distributions, and successfully treat small samples, correlations between sources, and mis-specified parameters.

Future work needs to address several practical questions: How the adaptive scheme would be implemented in an operating network so as to avoid excessive QoS violations while adapting; on what time-scale should adaptation take place; and can it address more realistic scenarios, such as multiple QoS classes, and other QoS parameters than packet loss rate. This chapter gives confidence these questions can be successfully answered.

Appendix A: Separable Representations

This appendix proves separable representations are required for fixed decision functions to have non-zero efficiency and robustness.

Theorem 1: Given a representation, $\Phi(\Sigma)$, there exists a fixed $C(\bar{\phi})$ with $E(C) > 0$ and $R(C) > 0$ if and only if $\Phi(\Sigma)$ is separable.

Proof: Suppose there exist two source combinations, Σ_1 and Σ_2, such that $\Phi(\Sigma_1) = \Phi(\Sigma_2)$ and $Q(\Sigma_1)$ meets QoS while $Q(\Sigma_2)$ does not. Let $\bar{\phi} = \Phi(\Sigma_1) = \Phi(\Sigma_2)$. Choose any C. If $C(\bar{\phi}) = -1$ then choose a distribution consisting solely of Σ_1. The optimal classifier accepts this combination so $E(C) = 0$. If $C(\bar{\phi}) = +1$, choose a distribution consisting solely of Σ_2. In this case, the optimal classifier will reject the combination so $R(C) = 0$.

Suppose instead $\Phi(\Sigma)$ is separable. Then, $C(\Phi(\Sigma)) = 1$ if $Q(\Sigma)$ meets QoS requirements, $C(\Phi(\Sigma)) = -1$ otherwise is well defined and will always be optimal regardless of $f(\Sigma)$. Q.E.D.

Appendix B: Simulation Model and Ethernet Data

We describe simple node and traffic models used in this chapter's simulations. The node is modeled as a discrete-time single-server queueing model where in each time slot one packet can be processed and zero or more packets can arrive from different sources. All the packets arriving in a time slot are immediately added to a buffer, any buffer overflows would be discarded (and counted as lost), and if the buffer was non-empty at the start of the timeslot, one packet sent. The server's buffer is fixed at 1000 packets. All rates are normalized by the service rate.

Appendix C: Asymptotic Optimality of the Adaptive Method

Are there any consistent or robust adaptive multiplexing decision functions? The answer is yes. This appendix will not show this rigorously, but instead sketches the proof outline. For more details the reader is directed to [Bro95]

[Bro99b]. For a specific feature, $\bar{\phi}$, we can consider three cases. In the first case, according to the distribution, *f*, the probability of getting samples at or near $\bar{\phi}$ is zero, i.e. $\bar{\phi}$ is *unsupported* by *f*. In this case, we don't care, since the value of the classifier at this feature does not affect consistency or robustness. In the second case, $q_i(\bar{\phi}) = \tau_i$ for some *i* (see eq. (3.2)). In this case, the classifier may or may not agree with the optimal classifier defined by (3.3). But, for continuous valued QoS metrics the measure of $\bar{\phi}$ where $q_i(\bar{\phi}) = \tau_i$ is likely zero.

The third case, is where $q_i(\bar{\phi}) \neq \tau_i$ for all *i* and $\bar{\phi}$ is supported by *f*. This is the usual case we are concerned with. We assume that $q_i(\bar{\phi})$ is continuous and there exists some unbiased and consistent point estimator of the $q_i(\bar{\phi})$. For a given data set we can define a neighborhood around $\bar{\phi}$ and use the estimators of the QoS metrics in (3.2) to decide the classifier output. As the sample size grows we can simultaneously shrink the neighborhood size while increasing the number of samples in the neighborhood so the estimates are consistent estimates of the true QoS metric. Thus, a consistent estimator is possible.

While this approach guarantees consistency (an asymptotic result), it says nothing about the confidence of the estimates, (3.9), for finite sample sizes. Unfortunately there are many pitfalls possible depending on *f*, and q_i. For this reason, Section 4, uses the approach of first deciding where confident estimates can be made and then fitting a function to these estimates, implicitly encapsulating assumptions about the smoothness of *f*, and q_i.

Acknowledgments

We gratefully acknowledge helpful discussions with Krishnan Komandar, Mark Garrett, Walter Willinger. The Bellcore traces were gathered by D.V. Wilson. This work was supported by NSF CAREER Award NCR-9624791.

References

[Bis95] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford U. Press, Oxford, 1992. 482p.

[Bro95] Brown, T.X, "Classifying loss rates with small samples," in *Proc. of IWANNT*, Erlbaum, Hillsdale, NJ, 1995. pp. 153-161,

[Bro97] Brown, T.X, "Adaptive access control applied to ethernet data," *Advances in Neural Information Processing Systems,* 9, MIT Press, 1997. pp. 932-8.

[Bro99a]Brown, T. X, Tong, H., Singh, S., "Optimizing admission control while ensuring quality of service in multimedia networks via reinforcement learning," in *Advances in Neural Information Processing Systems*, 11, MIT Press, 1999, pp. 982-8.

[Bro99b]Brown, T. X, “Classifying loss rates in broadband networks,” in *1NFOCOMM ‘99*, New York, April v. 1, pp. 361-70, 1999.

[Che92]Chen, X., Leslie, I.M., “Neural adaptive congestion control for broadband ATM,” *IEE Proc.-I,* v. 139, n. 3, pp. 233–40, 1992.

[Cho94]Choudhury, G.L., Lucantoni, D.M., Whitt, W., “On the effectiveness of admission control in ATM networks,” in the 1*4th International Teletraffic Congress* in France, June 6-10, 1994. pp. 411–20.

[Dud73]Duda, R.O., Hart, P.E., *Pattern Classification and Scene Analysis*, Wiley & Sons, New York, 1973.

[Elw93]Elwalid, A. I., Mitra, D. “Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks,” *IEEE/ACM Trans. on Networking,* v. 1, n. 3, June 1993.

[Est94] Estrella, A.D., Jurado, A., Sandoval, F., “New training pattern selection method for ATM call admission neural control,” *Elec. Let.*, v. 30, n. 7, pp. 577-9, Mar. 1994.

[Err96] Erramilli, A., Narayan, O., Willinger, W., “Experimental queueing analysis with long-range dependent packet traffic,” *IEEE/ACM T. on Networking*, v. 4, n. 2, pp. 209-3, April 1996.

[Gal95] Galmes, S., et al., “Effectiveness of the ATM forum source traffic description,” in *Local and Metropolitan Communication Systems, v.* 3. ed. Hasegawa, T., et al. Chapman and Hall, 1995. pp. 93-107.

[Gar94] Garrett, M.W., Willinger, W., “Analysis, modeling and generation of self-similar VBR video traffic,” in *Proc. of ACM SIGCOMM,* 1996. pp. 269-80.

[Gro96] Grossglauser, M., Bolot, J-C., “On the relevance of long-range dependence in network traffic,” in *Proc. of ACM SIGCOMM,* 1994. pp. 15-24.

[Gue91] Guerin, R., Ahmadi, H., Naghshineh, M., “Equivalent capacity and its application to bandwidth allocation in high-speed networks,” *IEEE JSAC*, v. 9, n. 7, pp. 968-81, 1991.

[Gue99]Guerin, R., Peris, V., “Quality of service in packet networks: basic mechanisms and directions,” *Computer Networks and ISDN Systems*, v. 31, n. 3, 1999. pp. 169-89

[Hey96]Heyman, D.P, Lakshman, T.V., “Source models for VBR broadcast-video traffic,” *IEEE/ACM T. on Networking*, v. 4, n. 6, pp. 40–8, 1996.

[Hir90] Hiramatsu, A., “ATM communications network control by neural networks,” *IEEE T. on Neural Networks*, v. 1, n. 1, pp. 122–30, 1990.

[Hir95] Hiramatsu, A., “Training techniques for neural network applications in ATM,” *IEEE Comm. Mag.*, October, pp. 58–67, 1995.

[Jam92]Jamin, S., et al., "An admission control algorithm for predictive real-time service," *Third Int. Workshop Proc. of Network and Operating Systems Support for Digital Audio and Video*, 1992. pp. 349-56.

[Kaw95]Kawamura, Y., Saito, H., "VP bandwidth management with dynamic connection admission control in ATM networks," in *Local and Metropolitan Communication Systems*, vol. 3. ed. Hasegawa, T., et al. Chapman and Hall, London, 1995. pp. 233–52.

[Kni99] Knightly, E.W., Shroff, N.B., "Admission Control for Statistical QoS: Theory and Practice," *IEEE Network,* March/April 1999, pp. 20–9.

[Kri95] Krishnan, K.R., "The Hurst parameter of non-Markovian on-off traffic sources," *Bellcore Technical Memorandum*, Feb., 1995.

[Lau93]Lau, W.C., Li, S.Q., "Traffic analysis in large-scale high-speed integrated networks: validation of nodal decomposition approach" *Proc. of lNFOCOMM*, v. 3, 1993. pp. 1320–29.

[Lee96]Lee, D.C., "Worst-case fraction of CBR teletraffic unpunctual due to statistical multiplexing," *IEEE/ACM Tran. on Networking*, v. 4, n. 1, Feb. 1996. pp. 98-105.

[Lel93] Leland, W.E., et al., "On the self-similar nature of ethernet traffic," in *Proc. ofACM S1GCOMM* 1993. pp. 183–3, also in *IEEE/ACM T. on Networking,* v. 2, n. 1, pp. 1-15, 1994.

[Lev97]Levin, B., Ericsson Project Report, to appear.

[Mit88] Mitra, D., "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Prob.,* v.20, pp.646–76, 1988.

[Mit98] Mitra, D., Reiman, M.I., Wang, J., "Robust dynamic admission control for unified cell and call QoS in statistical multiplexers," *IEEE JSAC*, v. 16, n. 5, pp. 692-707, 1998.

[Nev93]Neves, J.E., et al., "ATM call control by neural networks," in *Proc. Inter. Workshop on Applications of Neural Networks to Telecommunication*," Erlbaum, Hillsdale, NJ, pp. 210–7, 1993.

[Nor93]Nordstrom, E., "A hybrid admission control scheme for broadband ATM traffic," in *Proc. IWANNT,* Erlbaum, pp. 77-84, 1993.

[Nor94]Norros, I., "A storage model with self-similar input," *Queueing Systems,* v. 16, pp. 387-96, 1994

[Pax94] Paxson, V., Floyd, S., "Wide-area traffic: The failure of Poisson modeling," in *Proc. ofACM SIGCOMM,* 1994. pp. 257–68.

[Ton98]Tong, H., Brown, T. X, "Estimating Loss Rates in an Integrated Services Network by Neural Networks," in *Proc. of Global Telecommunications Conference (GLOBECOM 98),* v. 1, pp. 19-24, 1998.

[Ton99]Tong, H., Brown, T.X, "Adaptive call admission control under quality of service constraints: a reinforcement learning solution," to appear in *IEEE JSAC,* Feb. 2000.

[Tra92] Tran-Gia, P., Gropp, O., "Performance of a neural net used as admission controller in ATM systems," *Proc. GLOBECOM 92*, Orlando, FL, pp. 1303-9.

[Wil95] Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V., "Self-similarity through high-variability: statistical analysis of ethernet LAN traffic at the source level," *Bellcore Internal Memo,* Feb. 7, 1995. Also in *IEEE/ACM T. on Networking,* v. 5, n. 1, pp. 71-86, 1997.

Timothy X Brown received his B.S. in physics from Pennsylvania State University in 1986 and his Ph.D. in electrical engineering from California Institute of Technology in 1991. He has worked at the Jet Propulsion Laboratory and Bell Communications Research. Since 1995 he is an Assistant Professor at the University of Colorado, Boulder. His teaching and research areas include: Telecommunication Systems, Wireless, Switching, ATM, Networking, and Machine Learning. He received the NSF CAREER Award in 1996.

Chapter 4

TRAFFIC MANAGEMENT IN ATM NETWORKS: AN OVERVIEW[1]

C. Blondia

University of Antwerp, Department of Computer Sciences and Mathematics, Universiteitsplein 1, B-2610 Antwerpen, Belgium. blondia@uia.ua.ac.be

O. Casals

Polytechnic University of Catalonia, Computer Architecture Department, Jordi Girona 1-3, Módulo D6, E-08034 Barcelona, Spain. olga@ac.upc.es

Abstract: The main objectives of traffic management in ATM networks are to protect the user and the network in order to achieve network performance objectives and to use the available resources in an efficient way. In order to achieve these objectives the profile of the cell stream of each connection needs to be described adequately by means of a set of traffic parameters, together with an indication of the required level of QoS. The relationship between network performance and traffic characteristics and QoS is structured by means of ATM layer Service Categories and Transfer Capabilities. Each Category/Capability is provided with a number of traffic congestion and traffic control mechanisms needed to guarantee the required QoS of the category while achieving a high level of efficiency. This paper presents a state-of-the-art of traffic management in ATM networks. An overview is given of the Service Categories, together with the most important control and congestion schemes: CAC, UPC, traffic shaping, priority control, resource management, flow control, packet discarding schemes.

Keywords: ATM, Traffic Management, CAC, UPC, traffic shaping, flow control

[1] This work was supported by the European Union under project AC094 (EXPERT). The first author was also supported by Vlaams Actieprogramma Informatietechnologie under project ITA/950214/INTEC (Design and control of broadband networks for multimedia applications) and by the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT), under the BATMAN project. The second author was supported by the Spanish Ministry of Education under projects TIC96-2042-CE and TIC98-1115-C02-01.

1. INTRODUCTION

The Asynchronous Transfer Mode (ATM) has been chosen as the transfer mode for B-ISDN because of its flexibility to support various types of services, each having their own traffic characteristics and performance requirements, and because of its efficiency with respect to resource utilisation, due to the potential gain by statistically multiplexing bursty traffic. Since ATM has to provide differentiated Quality of Service (QoS) to the various applications, there is a need for efficient, effective and simple functions which control the traffic streams and their resource utilization. These ATM layer traffic and congestion control functions are referred to as *Traffic Management* mechanisms. They are defined and standardised by ITU-T in Recommendation I.371 (Traffic Control and Congestion Control in B-ISDN, see [1371]) and by the ATM Forum in Traffic Management Specification 4.0 (see [ATM95]). The objective of traffic management is twofold.

- To achieve well-defined *performance objectives* by protecting both the user and the network against congestion. These performance objectives can be expressed in terms of cell loss probabilities, cell transfer delay, cell delay variations, etc.
- To achieve *efficiency* and *optimisation* of the usage of network resources needed to ensure the above mentioned performance requirements.

Traffic management mechanisms should be able to take the appropriate actions under all possible traffic conditions, such as

- temporarily *overload conditions* due to the statistical fluctuation of variable bit rate traffic
- *malicious users*, who deliberately offer more traffic to the network to obtain operational and/or economical advantage with respect to the other users
- *malfunctioning* of terminal equipment, leading to unexpected traffic volumes entering the network.

In order to structure the relationship between traffic characteristics and QoS requirements on one hand and network behaviour on the other hand, ATM Service Categories (ATM Forum terminology) or ATM Transfer Capabilities (ITU-T terminology) have been introduced. These service categories are intended to support a number of ATM Service Classes and associated QoS by means of a set of appropriate traffic management mechanisms.

The aim of this paper is to give an overview of these mechanisms. It is structured as follows. In Section 2 the parameters needed to define the notion of QoS and to characterize the traffic are introduced. Section 3 gives an overview of the ATM Service Categories and ATM Transfer Capabilities currently defined or under definition. In Section 4, we discuss the most important traffic control mechanisms: CAC, UPC/NPC, traffic shaping, priority control and resource management mechanism. Section 5 deals with congestion control mechanisms for Best Effort type of service. Here we discuss the ABR flow control scheme, several intelligent packet discarding schemes for the UBR Service Category and the mechanisms related to the Guaranteed Frame Rate Service Category. Finally conclusions are drawn in Section 6.

2. PARAMETERS OF QUALITY OF SERVICE AND TRAFFIC CHARACTERIZATION

2.1 ATM LAYER QUALITY OF SERVICE

The ATM layer Quality of Service (QoS) is defined by means of a set of parameters that characterise the end-to-end performance of a connection at the ATM layer. These parameters can be divided into two classes, namely parameters that may be negotiated between the end-systems and the network, two of which are related to cell delay and one to cell loss, and parameters that are given by the network.

The *Maximum Cell Transfer Delay* (maxCTD) is defined to be the $(1-\alpha)$ quantile of the Cell Transfer Delay (CTD). The *Peak-to-peak Cell Delay Variation* (Peak-to-peak CDV) is defined to be the $(1-\alpha)$ quantile of the CTD minus the fixed CTD (which represents the component of the delay due to propagation and switch processing). This measure quantifies the difference between the best and the worst case of CTD. The *Cell Loss Ratio* (CLR) is defined to be the number of lost cells divided by the total number of transmitted cells, including those that are delivered late w.r.t. the $(1-\alpha)$ quantile of the CTD. There are three non-negotiated QoS parameters: the *Cell Error Ratio* (CER), the *Severely Errored Cell Block Ratio* (SEBR) and the *Cell Misinsertion Rate* (CMR).

2.2 TRAFFIC PARAMETERS AND GENERIC CELL RATE ALGORITHM

Traffic parameters are used to describe traffic characteristics of an ATM connection. A major requirement of an ATM traffic parameter is its suitability to test whether a connection behaves conform the values of this parameter. Therefore, these parameters are given an operational definition, rather than a statistical definition, allowing conformance testing in a direct way, opposite to e.g. the mean bit rate. The algorithm used to define the traffic parameters in an operational way is the *Generic Cell Rate Algorithm* (GCRA). There are two equivalent definitions of the GCRA, namely the

Virtual Scheduling Algorithm and the *Continuous Leaky Bucket Algorithm.* We give both definitions and leave it to the reader to check the equivalence. GCRA is defined by means of two parameters T and τ, T being the *increment* and τ being the *limit* and is denoted by GCRA(T, τ).

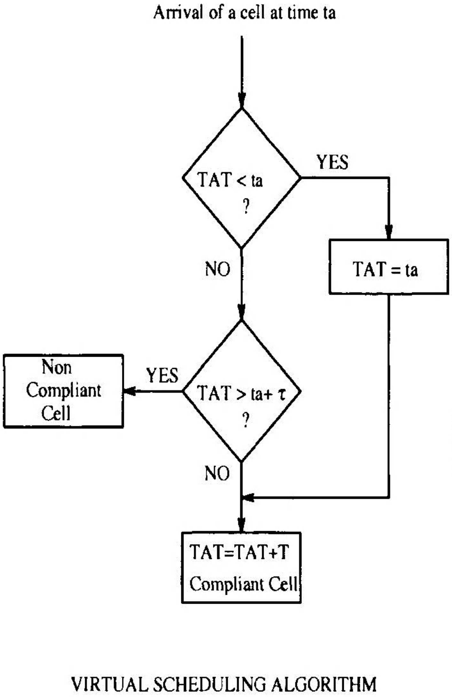


Figure 4.1. Virtual Scheduling Algorithm

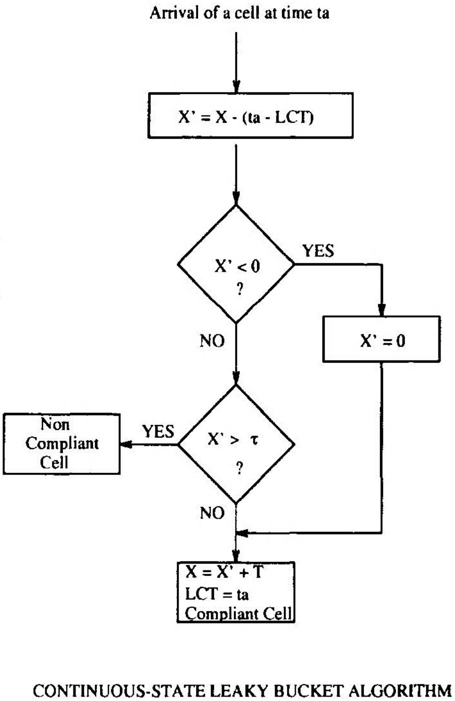


Figure 4.2. Continuous-State Leaky Bucket

Figure 4.1 shows the Virtual Scheduling Algorithm, t_a denotes the actual arrival time of a cell, TAT the Theoretical Arrival Time, based on the assumption that cells arrive equally spaced (the interarrival time being T) and τ represents a certain tolerance. The continuous-state LB is a finite capacity queue (equal to T+τ) with a continuous leak of 1, of which the content increases by T every time a cell arrives. Its operation is depicted in Figure 4.2. X denotes the contents of the LB, while LCT denotes the Last Conformance Time.

2.3 THE CONNECTION TRAFFIC DESCRIPTOR

The connection traffic descriptor consists of two parts: the source traffic descriptor, being the peak cell rate (PCR), the sustainable cell rate (SCR), the burst tolerance (BT) and the cell delay variation tolerance (CDVT).

2.3.1 The Peak Cell Rate

The *Peak Cell Rate* (PCR) R_p of a connection is defined at the Physical Layer Service Access Point (SAP), as the inverse of T, the minimum time between the emission of two cells from this connection.

2.3.2 The Cell Delay Variation (CDV) Tolerance

The cell stream of a connection may experience variable delay before entering the network (i.e. before the T_B interface), and hence before being submitted to the policing function. This Cell Delay Variation (CDV) is due to ATM Layer functions (multiplexing of connections introduces variable delay), Physical Layer functions, the insertion of OAM cells and customer equipment.Therefore, the UPC function can not operate purely on basis of the PCR. Some tolerance to cope with the CDV has to be built in. This tolerance is defined using the GCRA.The CDV tolerance τ, is defined as the second parameter GCRA(T,τ), where T denotes the inverse of the PCR.

2.3.3 The Sustainable Cell Rate

The PCR and the CDVT describe the cell rate of a CBR connection in an adequate way. However, an important part of the traffic carried by an ATM network consists of VBR traffic (e.g. video). Restricting the traffic descriptor to PCR, would lead to resource allocation on basis of the PCR, and no statistical gain could be achieved. Hence a parameter is needed which reflects a kind of average bandwidth utilization of a connection. Since the mean peak rate is not suited for policing purposes (see [RAG]), we define the *Sustainable Cell Rate* (SCR) R_s as the inverse of T_s which takes a value between the minimal cell interarrival time T and the mean cell interarrival time).

2.3.4 The Burst Tolerance

The Burst Tolerance τ_s is defined as the second parameter in the GCRA(T_s , τ_s), where T_s denotes the inverse of the SCR defined above. It gives an upper bound on the length of a burst transmitted at peak cell rate. It is easy to show that the maximal burst size B, given T, T_s and τ_s, satisfies

$B = 1 + \lfloor \tau_s / T_s - T \rfloor$ where $\lfloor r \rfloor$ denotes the largest integer value less than or equal to r. Remark that when a connection has generated a burst at PCR with length B, it has to be idle for a while before generating another burst. Hence, while the PCR and the CDV tolerance control the peak cell rate of a connection, the SCR and the burst tolerance control the burstiness of a connection.

2.4 TRAFFIC CONTRACT

During the connection set-up, a traffic contract between the user and the network is negotiated. This contract contains

- the requested QoS class : these classes are defined using the delay and cell loss parameters defined in 2.1.
- the traffic descriptor : the source traffic descriptor (PCR, SCR, BT) and the CDVT as defined in 2.3.
- the definition of a compliant connection : conformity is defined by means of one or more GCRAs.

3. ATM SERVICE CATEGORIES

In order to support efficiently the various services and applications with their specific QoS requirements, a number of ATM Service Categories (ATM Forum terminology) or ATM Transfer Capabilities (ITU-T terminology) have been defined. For each Service Category a set of appropriate traffic control and congestion control functions has to be identified, in order to achieve the required QoS of each class.

The ATM Forum has identified the following classes : Continuous Bit Rate (CBR), real-time Variable Bit Rate (rt-VBR), non-real-time Variable Bit Rate (nrt-VBR), Available Bit Rate (ABR), Unspecified Bit Rate (UBR) and Guaranteed Frame Rate (GFR). The ITU-T defines a similar structure, with the exception that no difference is made between real-time and non-real-time VBR, CBR is called Deterministic Bit Rate (DBR), VBR is called Statistical Bit Rate (SBR), UBR and GFR are not defined, but on the other hand the ATM Block Transfer (ABT) Capability is defined.

3.1 CONTINUOUS BIT RATE (CBR)

The CBR Service Category is intended for connections with stringent time relationship and bounded CTD and CDV requirements, which need a fixed amount of bandwidth for the whole duration of the connection. This bandwidth is characterised by the Peak Cell Rate. Typical applications are telephony, CBR video and circuit emulation services. The traffic parameters used for this class are PCR and CDVT. The QoS parameters are CLR, peak-to-peak CDV and maxCTD.

3.2 REAL-TIME VARIABLE BIT RATE (RT-VBR)

This Service Category is used for traffic streams with stringent time constraints (as CBR) but which transmit their information at a variable rate. As such, they exhibit a bursty character and hence are suited for statistical multiplexing gain. Typical applications are voice with silence detection and VBR video. The traffic parameters used for this class are PCR, SCR and BT. The QoS guarantees given are CLR, peak-to-peak CDV and maxCTD.

3.3 NON-REAL-TIME VARIABLE BIT RATE (NRT-VBR)

This Service Category is meant for non-real-time applications which exhibit a bursty character. As there are less stringent timing constraints, they are very well suited to achieve a high statistical multiplexing gain. Typical applications using this Service Category are response time critical transaction processing such as airline reservations and banking transactions. The traffic parameters that are used are PCR, SCR and BT. The only QoS guarantee is the CLR.

3.4 AVAILABLE BIT RATE (ABR)

The Available Bit Rate Service Category (ABR) has been introduced to support connections originating from users which are willing to accept unreserved bandwidth and which are able to adapt their cell rate to changing network conditions and available resources. Information about the state of the network (e.g. with respect to congestion) and the availability of resources is sent to the source as feedback information through special control cells, called Resource Management cells (RM cells). Services which are compliant to this feedback control information experience a low cell loss ratio and obtain a fair share of the available bandwidth. There is no guarantee with respect to the delay or delay variation. As this control scheme operates at the time scale of a complete round trip delay, the ABR Service Category requires large buffers to be present in the network. The traffic parameters used for this category are the Peak Cell Rate and a minimal usable bandwidth, called the Minimum Cell Rate (MCR). The only QoS guarantee is the CLR. The available bandwidth may vary in time, but shall never be lower than the MCR. Typical applications using this category are Remote Procedure Calls, Distributed File Transfer, Computer Process Swapping, etc.

3.5 UNSPECIFIED BIT RATE (UBR)

The Unspecified Bit Rate Service Category (UBR) is meant for traditional computer communication applications (such as e-mail, file transfer, etc.), where no specific QoS guarantees are required. It is the Best Effort ATM Service Category. No guarantees are offered with respect to CLR or CTD. The source will specify a PCR.

3.6 ATM BLOCK TRANSFER (ABT)

The ATM Block Transfer Capability (ABT), defined by ITU-T but not considered by the ATM Forum, provides a service with transfer characteristics negotiated on an ATM block basis. As such, it can be considered as a “non-permanent” CBR service. When a block is accepted by the network, sufficient network resources are allocated such that the QoS guarantees are equivalent to those offered to a CBR connection with the PCR negotiated for the transmission of a block. There are two variants of the ABT Transfer Capability : with Delayed Transmission (ABT/DT) and with Immediate Transmission (ABT/IT). In the first case an ATM block is transmitted only after the block cell rate has been confirmed by the network (i.e. after the network has reserved the required resources to transmit the block according to the agreed QoS). In the second case the block is transmitted immediately without waiting for the acknowledgement. This may result in a loss of the whole block if one ore more network elements on the path are short of resources. The traffic parameters specified by the source are PCR, SCR and BT. The QoS guarantees are the CLR, CTD, CDV and the blocking probability.

3.7 GUARANTEED FRAME RATE (GFR)

The Guaranteed Frame Rate Service Category (GFR) was first proposed in [GUE96] with a different name (UBR+). The objective of this service is to incentive users to migrate to ATM technology. Many existing users are not able to specify the traffic parameters required by the previous ATM services. For these users the only possibility to access ATM networks is through UBR connections which do not give any of the ATM QoS guarantees. GFR keeps the simplicity of UBR while providing the user with a minimum cell rate (MCR) guarantee as long as the user sends AAL5 frames of size less than the specified value. The service also allows the user a fair share of the spare bandwidth, i.e. the excess traffic of each user will get a fair access to the available resources. The traffic parameters used by GFR are the PCR, CDVT, MCR and the maximum AAL5-PDU size.

4. TRAFFIC CONTROL MECHANISMS

The basic ATM control functions we discuss in this section are: Connection Admission Control (CAC), Usage/Network Parameter Control (UPC/NPC), Priority Control and Selective Cell Discarding, Traffic Shaping and Resource Management.

4.1 CONNECTION ADMISSION CONTROL (CAC)

According to ITU-T Recommendation I.371, *Connection Admission Control* (CAC) is the set of actions taken by the network at the call set-up phase (or during the call re-negotiation phase) in order to establish whether a VC/VP connection can be accepted or rejected. A connection is to be accepted at its required Quality of Service (QoS) while maintaining the agreed QoS of already existing connections. The decision depends on the network resources that are available (and hence on the load of the network) and on the characteristics of the connection to be established.

4.1.1 Connection Admission Control for CBR Traffic

When the traffic that is offered to a multiplexer has a constant bit rate, then a straightforward approach could be simply admit connections as long as the sum of the PCRs does not exceed the capacity of the link. The buffer behavior can then be evaluated using the $N \times D/D/1$ or the $\sum N_i \times D_i / D/1$ model (see [VR89] and [RV91]). However, the presence of CDV makes this simple rule not necessary valid, unless the CDV that is allowed is negligible (see [COST242], Section 5.1.1 for a discussion on negligible CDV). When the CVD is not negligible, one may keep the constraint that $\sum PCR_i \le C$, for a link with capacity C, in addition to the condition that $\sum_i b_i \le B$, where b_i is the burst size of source i and B is the buffer capacity of the multiplexer. Based on those conditions, when C, B and PCR are given, the bucket depth b_i for source i has to be limited by $b_i = r_i \frac{B}{C}$. In this model worst case assumptions are supposed.

4.1.2 Connection Admission Control for VBR Traffic

Assume that a number of VBR sources are to be multiplexed on a link. When the buffer of the multiplexer is intended to absorb cell scale congestion we refer to this type of multiplexer as *Rate Envelope Multiplexing* (REM). If the buffer capacity is large enough to cope with burst scale congestion, we refer to *Rate Sharing Multiplexing* (RSM). In what follows we discuss these two multiplexing schemes and related CAC algorithms in more detail.

Rate Envelope Multiplexing (REM)

When dealing with services which have to meet strict delay requirements, such as interactive voice and video, small buffers (of the order of 100) able to absorb cell scale congestion (i.e. congestion due to a concentration of cell arrivals from different sources) are sufficient. The aim of CAC in this case is to limit the arrival rate such that the probability that the arrival rate exceeds the service rate is negligible. This type of multiplexing is also called *bufferless multiplexing.* With respect to the multiplexing efficiency, studies have shown that REM is efficient for bursty sources with peak cell rates which are low with respect to the link rate. The key idea is to define a notion of *Effective Bandwidth* (EB) which is used by the CAC algorithm. To determine the value of the Effective Bandwidth of a source, one may use statistical knowledge about the source (e.g. mean, variance) (in particular when measurement-based CAC is performed, as explained later in this Section) or one may assume worst case assumptions based on the traffic parameters defined by one or more GCRAs. Examples of EB definition based on statistical characteristics may be found in [KEL91], where a Chernoff bound is used to compute the probability of resource saturation. In [ROB92], an empirical expression is used to determine the EB based on the mean and the variance of the source rate. A worst case Effective Bandwidth definitions based PCR, SCR and MBS (maximum burst size) is given in [EMW95]. Traffic with the given parameters is considered to be of ON/OFF type, which transmits at PCR during on periods of duration MBS and at rate 0 during off times, such that the mean rate is SCR.

Rate Sharing Multiplexing (RSM)

In RSM, the probability that the input rate exceeds the link rate is non-negligible. Large buffers are needed to absorb this momentary input rate excess. Such situations occur in particular in data networks (with less strict timing constraints) where connections may have large peak bit rates compared to the link rate. Rate Sharing performance heavily depends on the traffic characteristics of the input traffic. For example in the case of simple ON/OFF sources, the notion of Effective Bandwidth in REM only depends on the peak and mean rate, while for RSM also the distribution of the duration of the ON and OFF periods and the correlation between successive bursts have a significant impact.

In order to simplify CAC, also for RSM a notion of Effective Bandwidth is introduced. It can be determined on basis of the asymptotic slope of the complementary queue length distribution (see e.g. [GAN91], [EM93], etc.). When the complementary buffer occupation distribution in a multiplexer with bandwidth C is given by $P(B > b) \approx e^{-\eta_i(c)b}$, then the EB needed to obtain an overflow probability with a buffer of size B less than 6 is given by

$c_i = \frac{1}{\eta_i(c)}(-\log\frac{\varepsilon}{B})$. The function $\eta_i(c)$ is determined by the statistical properties of the traffic source. Remark that the above asymptotic behaviour is valid for Markovian input but fails to be true for traffic with for example Long Range Dependence characteristics (see e.g. [DB98]).

4.1.3 Connection Admission Control of ABR Traffic

The ABR flow control scheme (see Section 5) aims at exploiting the available bandwidth, while achieving fair sharing of this bandwidth between contending connections. Assuming that new connections arrive according to a Poisson process, a link can be modelled as an M/G/1 processor sharing queue (see [ROB98]). It is well known that the mean transfer delay of a file of size x is linear in x. A generalisation of this queue can be obtained by letting the connection have a minimum and maximum throughput. Results of this system can be found in [COH79]. The blocking probability of a new connection in this setting depends on the file size distribution only through its mean value.

4.1.4 Measurement-Based Connection Admission Control

The parameters used to describe CBR and VBR traffic constitute in general a limited representation of the traffic variability, and as such may lead to an inefficient resource usage. An alternative approach consists of taking CAC decisions based on traffic measurements. Two different methods can be distinguished. First, a global measurement of the mean and/or variance of the bit rate of the aggregate traffic on a link is performed. An example of such a CAC algorithm, based on the Hoeffding bound can be found in [BS97] and [BS98]. Secondly, the traffic on a link is divided into classes of traffic with similar statistical properties (e.g. peak bit rate, burstiness) and the measurements are made per class. When the peak bit rate is a declared parameter and the mean is measured, one may define an equivalent bandwidth per class which is used in a CAC algorithm. An example of per class measured-based CAC can be found in [GK97]. In case all the connections have small peak bit rate with respect to the link rate, a global measurement is sufficient. The second approach, which is more complex, is recommended in case the connections have significantly different peak bit rate values.

4.2 USAGE/NETWORK PARAMETER CONTROL (UPC/NPC)

Once the contract between the user and the network is established and the connection is accepted, the network needs mechanisms (i) to check that the traffic is generated according to the specification and (ii) to enforce the compliance in case of violation. These actions can be performed at the User-Network Interface (UNI) and in this case it is called Usage Parameter Control (UPC) or at the Network-Node Interface (NNI), where it is referred to as Network Parameter Control (NPC). The mechanisms involved are called policing mechanisms. Once a connection is accepted, the CAC informs the UPC about the traffic contract.

4.2.1 UPC/NPC Requirements

The UPC/NPC is defined as the set of actions taken by the network to monitor and control traffic in terms of traffic offered and validity of the ATM connection, at the user access and the network access respectively [I371]. The main purpose is to protect network resources from malicious as well as unintentional misbehaviour which can affect the QoS of other already established connections by detecting violations of negotiated parameters and taking actions. In general, any UPC/NPC mechanism has to comply with the following requirements:

- the ability to detect any illegal traffic situation
- the ability to determine if the traffic is compliant
- fast reaction to parameter violations
- transparency to compliant traffic
- easy to implement.

A UPC/NPC mechanism has to decide whether a random cell flow is conforming or not. Such a mechanism can not be perfect and, even if the user respects its traffic contract, a certain number of cells will be erroneously detected as non-conforming or violating cells will be declared conforming. These errors should be kept very low (typically lower than the CLR).

4.2.2 UPC Location and Actions

The policing function is part of the public network, but it should be located as close as possible to the user. Therefore, the UPC function is located where the Virtual Channel Connections (VCC) or Virtual Path Connections (VPC) are terminated within the network. This implies that UPC is performed before the first switching activity takes place.

A UPC mechanisms may perform the following actions at the cell level: cell passing, cell re-scheduling, cell tagging and cell discarding. Cell passing and cell re-scheduling are performed on cells which are identified by a UPC/NPC as compliant. Cell re-scheduling is performed when traffic shaping and UPC are combined. Cell tagging and cell discarding are performed on cells which are identified by a UPC/NPC as non-compliant. Cell tagging operates on CLP=0 cells only by overwriting the CLP bit to 1.

4.2.3 Policing Mechanisms

Some authors (see e.g. [BOY92a], [GUI92]) distinguish between two classes of control mechanisms: the so-called "pick-up" mechanisms and those which shape the traffic.
A pick-up mechanism observes a cell flow and detects the exceeding cells. Therefore, the cells pass transparently through the policing device, or else they are detected as violating the contract and they are dropped or tagged. A shaper, in general, modifies the traffic even if it is non-conforming. Shaping will be discussed in the next section. The most well known pickup mechanisms are the Virtual Scheduling Algorithm and the Continuous-State Leaky Bucket Algorithm that have been proposed for conformance definition.

4.3 TRAFFIC SHAPING

Traffic shaping is a traffic control mechanism which alters the characteristics of a cell stream. It can perform the following actions: reduce the PCR, limit the burst length, reduce the CDV by suitably spacing cells in time.
An important class of traffic shaping mechanisms are the *Spacers*. By spacing the cells of a connection in time, the peak bit rate may be reduced, the Cell Delay Variation (CDV) may be controlled or the burst duration may be limited. Traffic shaping may be performed on different locations.
Traffic shaping in the Customer's Premises Network (*CPN*).
As mentioned before, the UPC function uses the traffic descriptor to check whether the cells stream offered to the network is conforming the contract. In order to enforce a source to be conform to the traffic contract, its cell stream may be shaped in the CPN before entering the network to obtain the required traffic characteristics.
Traffic shaping in the ATM network.
Passage through multiplexers and switches may alter the characteristics of a traffic stream considerably. In particular due to the queueing delays, the stream is jittered leading to cell clumping and dispersion of cells. These phenomena imply an important decrease of network utilisation to obtain a given QoS. Therefore, traffic shaping within the network is applied to change the traffic characteristics such that a higher utilisation is achieved.

4.3.1 Scheduling disciplines

Several queue service schemes have been proposed in order to be able to provide multiple QoS.

4.3.1.1 Generalised Round Robin (GRR)

The GRR [ROJ94] distinguishes an individual queue for each multiplexed connection. In the classical round robin discipline each queue is visited cyclically with at most one customer being served at each visit. The generalization consists of allowing the visit frequency to be different for each queue. The visit frequency would be determined by a bandwidth reservation parameter which, e.g., could be for a high speed data connection the sustainable cell rate, for a CBR connection the peak rate and for a low peak VBR connection some intermediate "equivalent rate". A "queueing engine" allowing GRR is described in [KMI92].

4.3.1.2 Fair Queuing (FQ)

FQ [DKS89] defines a separate FCFS queue for each connection and if k of these queues are currently not empty, then each non empty queue receives 1/k-th of the link bandwidth. Different bandwidth demands can be expressed using relative weights (Weighted Fair Queuing) [CSZ92].

4.3.1.3 Virtual Clock (VC)

In this scheme [ZHA91], the cells of a given stream i with bandwidth allocation Λ_i are allocated a time stamp on arrival and all cells in the multiplex queue are served in increasing order of time stamp. The time stamp of cell number n+1 is equal to the greatest of the time of cell n plus the maximum intercell interval $1/\Lambda_i$ and the current time. A cell is served as soon it reaches the head of the queue.

4.3.1.4 Virtual Spacing (VS)

The VS [ROJ94] realises the GRR queue discipline. As in the Virtual Clock algorithm, cells destined to a given output multiplex are attributed a time stamp which determines their order of service. However, only one cell per connection is stamped at any time, the stamp being attributed to a new cell only after the previous cell has been transmitted. The cells of any given connection are stored in a FIFO queue: when the first cell in the queue is transmitted at a certain time t, the next cell is attributed the time stamp $t + 1/\Lambda$ and will be served as soon as no other cell for the same multiplex has a time stamp of smaller value. If, when a cell is transmitted, no further cells of the same connection are queued, the next cell to arrive will be attributed a time stamp equal to the maximum of the current time and the value $t + 1/\Lambda$. If the VS would determine the new time stamp from the previous time stamp rather than the actual transmission time, we would have the VC algorithm.

4.3.1.5 Jitter Earliest-Due-Date

After service, a cell is stamped with the difference between its deadline and its actual finish time. The next switch will hold this packet for an extra amount of time equal to the calculated difference [VER91].

4.3.1.6 Stop and Go Queueing

Stop and Go Queueing [GOL90] consists of imposing a synchronised frame structure on the network guaranteeing the availability of transmission slots at the appropriate times for periodically arriving cell streams with real time constraints.

4.3.2 Spacing Algorithms

Conformance to the traffic contract at the network entry point does not imply that the traffic offered by the connection respects the negotiated PCR. It has been shown that the pick-up policing functions previously described do not prevent clusters of cells from entering the network and therefore cannot protect the network from congestion under all conditions. This is due to the jitter tolerance which has to be introduced in order to accommodate for the random delays introduced on the cell flow in successive multiplexing stages. The policing function is not able to decide whether short bursts that violate the specified peak cell rate are caused by delay jitter or by misbehaving customers. This problem can be avoided if the policing function not only discards excess cells but also delays cells so that their inter-departure times from the policing device between cells of one connection are never below a minimal value which is chosen according to the negotiated PCR. Several implementations of such devices which combine a pick-up policing function with a spacer have been [BOY92a], [WAL91].

4.4 PRIORITY CONTROL AND SELECTIVE CELL DISCARDING

The header of each ATM cell contains a Cell Loss Priority (CLP) bit. This bit is used to indicate a loss priority. The network may decide to selectively discard cells with low priority in favour of high priority cells. Priority marking can be performed either on a connection basis or on a cell basis. We give two examples of potential use of priority control in ATM networks.
(i) Different classes of QoS: By using the CLP bit on a connection basis, the network may distinguish two different classes of QoS: traffic for which CLP=0 and traffic for which CLP=1. In this case the network may guarantee different cell loss rates according to the class a connection belongs to and it must provide selective discard mechanisms in order to handle the different classes. Examples of such mechanisms are push-out, partial buffer sharing

[KRO90], [SUM88]. The increase in complexity of network elements due to these mechanisms may be compensated by the possible increase in accepted load in the network due to the existence of different classes of QoS with different cell loss ratio guarantees.

(ii) Cell tagging: When the UPC function detects a cell which violates the contract it may discard the cell or tag it as non-conforming (for a detailed discussion see Section 4.1.5). In the later case, the CLP bit may be used to indicate whether a cell is conforming or not. As soon as congestion occurs in the network, the CLP bit may then be used to selectively discard non-conforming cells.

4.5 FAST RESOURCE MANAGEMENT

Statistical multiplexing may lead to a more efficient use of the network resources at the expense of additional traffic control functions. A typical example of this principle may be found in Fast Resource Management, where control is performed on the time scale of the round-trip propagation delay of an ATM connection. Let us give an example of such a control mechanism.

In order to obtain a statistical multiplexing gain, the network should not allocate the peak bit rate for the whole duration of the connection for Variable Bit Rate (VBR) traffic. In addition, typical services generating VBR traffic, e.g. data services, tolerate a certain delay. These observations lead to the notion of the *Fast Reservation Protocol* (see [BOY92b]). The idea is to allocate the necessary bandwidth to a connection for the duration of a burst only. By means of Reservation Request Cells, a source indicates the desire to increase its bit rate. Two variants exist: *Fast Reservation Protocol with Delayed Transmission* (FRP/DT), where the source waits for an acknowledgement from the network (by means of a Reservation Accepted Cell) before increasing its activity and the *Fast Reservation Protocol with Immediate Transmission* (FRP/IT), where the burst is transmitted immediately after the request cell. In the later case, the whole burst is discarded in case the reservation fails. These schemes implement the ABT Service Category.

4.6 NETWORK RESOURCE MANAGEMENT

Network Resource Management (NRM) is a subset of traffic and congestion control functions related to resource configuration and allocation. The main networking technique is the use of VPCs. Managing these virtual path connections may involve [BUR90], [BUR9I] allocating capacity based on anticipated demand, re-routing traffic in times of congestion, changing allocations of capacity to cater for changing demand. By reserving capacity on VPCs, the processing required to establish individual VCCs is reduced:

individual VCCs can be established by making simple connection admission decisions at nodes where VPCs are terminated. VPCs can be used to [I371] :

- simplify CAC
- implement a form of priority control by segregating traffic types requiring different QoS
- aggregate user-to-user services such that the UPC can be applied to the traffic aggregation.
- efficiently distribute messages for the operation of traffic control schemes.

This use can lead to the following advantages: a reduced load on control equipment; lower call establishment delays; additional means of providing service protection to improve network availability; an additional means of controlling network congestion.

5. CONGESTION CONTROL FOR BEST EFFORT SERVICES

5.1 ABR FLOW CONTROL SCHEMES

The ATM Forum has proposed a number of congestion control mechanisms for the ABR service class. The two most important classes of proposals are the credit-based schemes and the rate-based schemes. Eventually the ATM Forum [ATM95] selected a rate-based, closed-loop, per-connection control which uses the feedback information from the network to regulate the rate at which the sources transmit cells. The transmission rate of each connection is controlled by means of special control cells called Resource Management (RM) cells. RM-cells flow from the source end system (SES) to the destination end system (DES) and return along the same path carrying congestion information (Figure 4.3). Depending on the congestion information received in the RM-cell, the SES increases or decreases its transmission rate. The standard specifies the source and destination behavior and several methods that a switch can implement to control congestion.

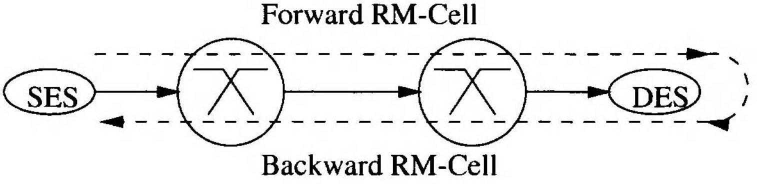


Figure 4. 3. ABR flow control

5.1.1 SES and DES behavior

At the connection set up the source negotiates the maximum and minimum rate at which it may transmit (PCR and MCR); the initial cell rate (ICR) at which it may start transmitting; the number of cells per RM-cell (Nrm); the rate increase factor (RIF) and the rate decrease factor (RDF). The flow chart in Figure 4.6 shows the source behavior. The SES starts transmitting with the agreed ICR. Each Nrm-1 data-cell transmissions, the SES sends an RM-cell with the following fields: Explicit Rate (ER) set to PCR ; Current Cell Rate (CCR) set to the Allowed Cell Rate (ACR) of the source ; Congestion Indication Bit set to 0 (no congestion); No Increase (NI) bit set to 0 (no increase) and Direction (DIR) bit set to forward. The ACR value establishes an upper bound to the transmission rate of the source. The source may transmit at the ACR while not becoming idle or rate-limited.

The cells are received by the DES which must store the Explicit Forward Congestion Indication Bit (EFCI) of the last Data-Cell received. On receiving a forward RM-cell it must change the CI bit to congested state depending on the EFCI bit stored, change the DIR to backward and send the RM-cell back to the SES along the same path.

On receiving a backward RM-cell the SES adjusts the ACR. When a backward RM-cell is received with CI = 0 and NI = 0, the SES is allowed to increase its rate (ACR) by no more than RDF*PCR. On receiving an RM-cell with CI = 1, the SES must decrease the ACR by at least RDF*ACR. Finally the ACR must be set at most to the ER field. The ACR cannot be reduced below the MCR or increased above the PCR. The actions marked as “Rescheduling option” are an optional behavior which allows to reschedule the transmission time of a cell in order to take advantage of an increase in the ACR. The actions marked as “ADTF adjustment” (ACR Decrease Time Factor) are used to control the ACR during the idle periods of the source. After such a period the source could start transmitting at the full ACR, resulting in a harm for the network if the last computed ACR was too high. The ADTF adjustment consists of measuring the elapsed time between two forward RM-cell transmissions. If this time is higher than the ADTF, the ACR is reduced down to the ICR. We note that if a source becomes rate-limited but not idle, it could also start transmitting at the full ACR and the ADTF adjustment would not work. To avoid this the ATM Forum establishes the so called “use-it-or-lose-it” optional behavior which consists of reducing the ACR in order to maintain it reasonably close to the transmission rate of the source.

The actions marked as “CRM adjustment” constrain the source to reduce the ACR in case of absence of backward RM-cells reception. This condition could be caused by a heavy congestion state of the network. If the number of forward RM-cell transmissions since the last backward RM-cell reception is higher or equal to CRM, the ACR must be reduced by, at least, ACR*CDF.

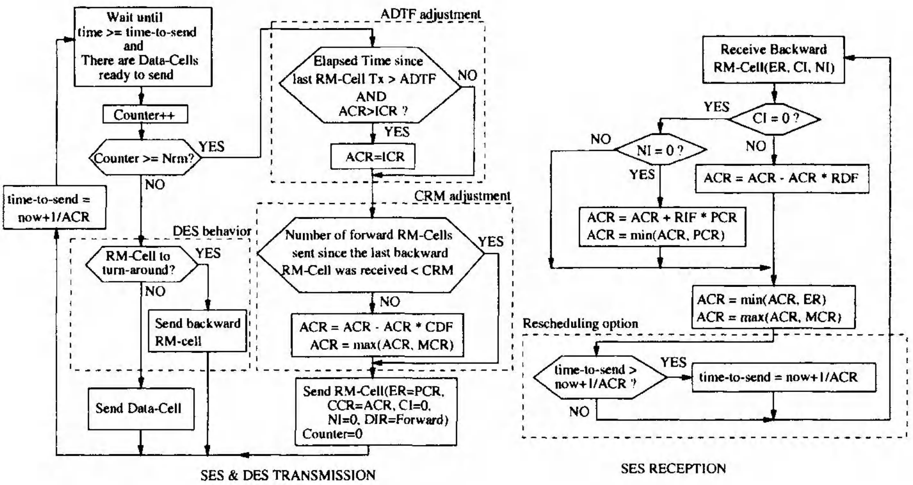


Figure 4.4. Source behaviour

5.1.2 ABR Switch Mechanisms

A switch shall implement at least one of the following methods to control congestion: set the EFCI bit of the data cells; set the CI or NI bit in forward and/or backward RM-cells; reduce the explicit rate (ER) field of forward and/or backward RM-cells. The switches that set the EFCI or CI bit to indicate a congestion state are known as binary switches. Switches that modify the ER field are called ER switches.

Several switch mechanisms compatible with the ATM Forum specifications have been proposed. They differ on the congestion monitoring criteria and the feedback mechanism used. We describe three of them which are well known to show the different degrees of performance and complexity that can be achieved.

5.1.2.1 EFCI Switch

The simplest switch mechanism [YIN94] marks the EFCI bit in data cell headers when congestion is detected. The switch monitors its queue length and detects congestion when it exceeds a given threshold. The feedback delay can be reduced by setting CI = 1 of backward RM cells during the congested state instead of setting the EFCI.

The main drawback of this switch mechanism is its lack of fairness. For example, RM-cells of a VC going through a higher number of congested links will be set to congested more often than those of VCs going through fewer congested links. This undesirable effect (known as the "beat down problem") will result in a lower rate for such VCs.

5.1.2.2 EPRCASwitch

The Enhanced Proportional Rate Control Algorithm (EPRCA) [ROL94] is an enhanced version of the original rate-control algorithm. The switch computes an heuristic approximation of a fair rate, equal to the link capacity minis the capacity of the constrained VCs over the non constrained VCs (max-min criterium). The fair rate (MACR in the figure) is computed during the uncongested periods as an exponential average (MACR = MACR + (CCR - MACR) AV) over all the VCs whose CCR is larger than MACR*VCS. AV is the averaging factor and VCS is a VC separator used to distinguish between VCs constrained by the switch and otherwise constrained VCs (see Figure 4.7). To avoid the "beat down problem", the switch just reduces during congested periods the ER field of the backward RM-cells with a CCR greater than MACR*DPF. The ER is reduced to MACR*ERF. The Down Pressure Factor (DPF) is used to cause the rate setting control when the ACR reaches a value slightly lower than the MACR. The Explicit Reduction Factor (ERF) is used to set the explicit rates slightly below MACR so that the switch will stay uncongested. The switch is considered congested when the queue length (Q) is greater than a threshold (Qth). If Q is greater than another threshold QD, the switch is considered very congested and ER is reduced in all backward RM-cells to MACR*MRF (MRF is a major reduction factor).

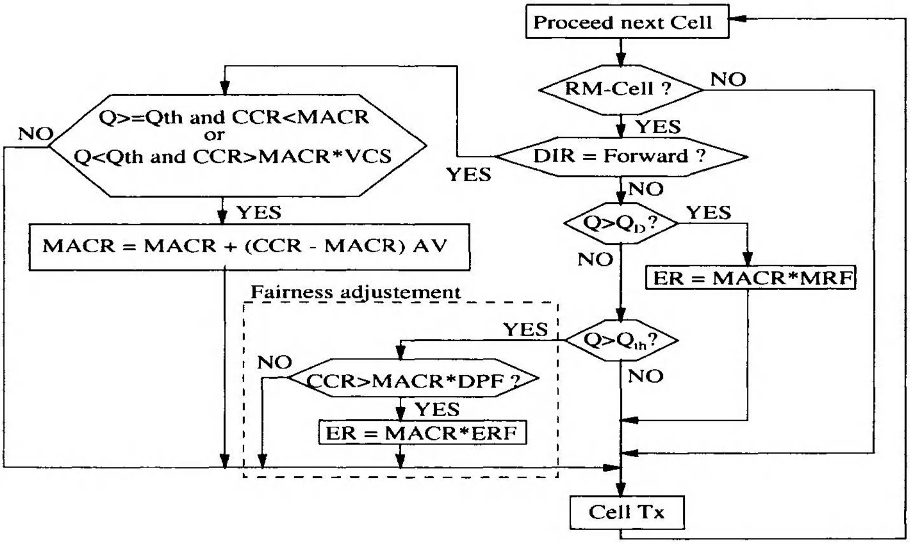


Figure 4.5. EPRCA Switch

5.1.2.3 ERICA Switch

The objective of the Explicit Rate Indication for Congestion Avoidance (ERICA) algorithm [JAI95] is to keep the queue length low and achieve max-min fairness (see Figure 4.6). Whereas In the previous mechanisms the detection of a congested state is based on a queue length threshold, in the ERICA proposal, the switches measure the input rate (IR) and compare it with a target cell rate (TCR, set to 85-95% of the link bandwidth) to compute the overload factor OF = IR/TCR. The ER field of backward RM-cells is then reduced by the OF in order to avoid the congestion state.

To compute the IR, the switch measures the time T until N cells arrive. Then it computes IR = N/T and starts another measuring interval. During each measuring interval, the switch also counts the number of active VCs in order to compute the fair share (FS) as FS = TCR/Number of VCs seen during the measuring interval. When receiving a backward RM-cell the switch computes the explicit rate ER2 based on load and fairness (ER2 = max(CCR/OF, FS)) and stores the value NER = min(TCR, ER2). If the ER field of the cell is higher than NER, the field is replaced by the computed value. To reduce the feedback delay the ER computation uses the CCR seen in the last forward RM-cell of the same VC. Therefore, this value must be stored in a VC table when a forward RM-cell is received.

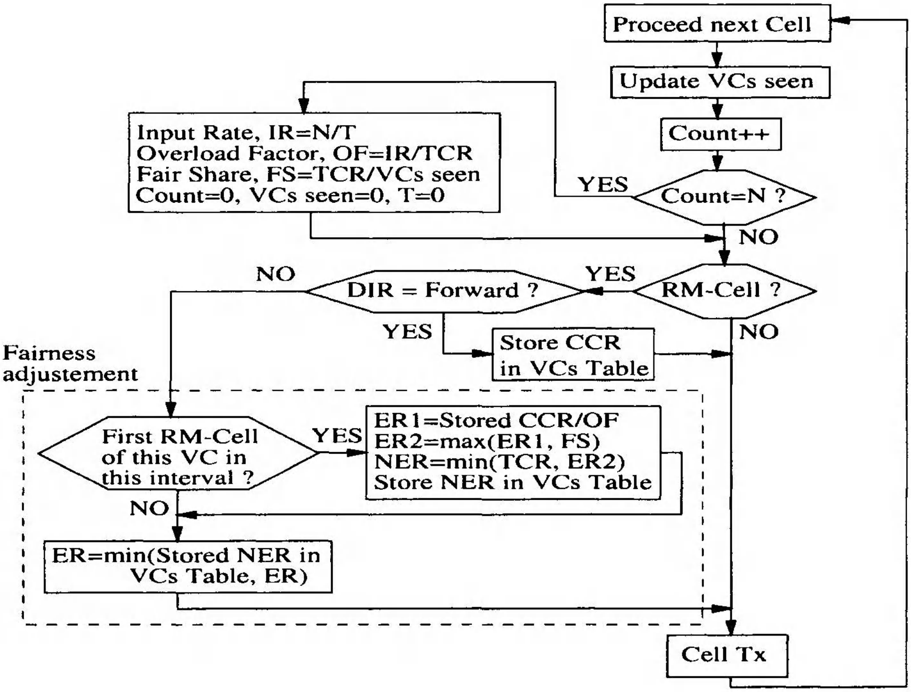


Figure 4.6. ERICA Switch

5.1.2.4 Comparison of the switch mechanisms

EFCI is the simplest switch mechanism. It only monitors the queue length and marks backward RM-cells when higher than a threshold. However it has been shown that high queue length can be reached and fairness cannot be guaranteed.

The EPRCA switch mechanism, which computes an average ACR reading the CCR field of RM-cells and modifying the ER field of backward RM-cells, achieves a better performance in terms of link utilisation, queue length and fairness than the EFCI switch [EXP96].

The ERICA switch mechanism is the most complex. It requires measuring the input rate of each buffer and accessing to a VC table each time a forward or a backward RM-cell is received. However it achieves a high degree of fairness and a tight queue length control. Another advantage compared to the EPRCA is the reduced number of parameters to be tuned (the target utilisation and the measuring interval in cells).

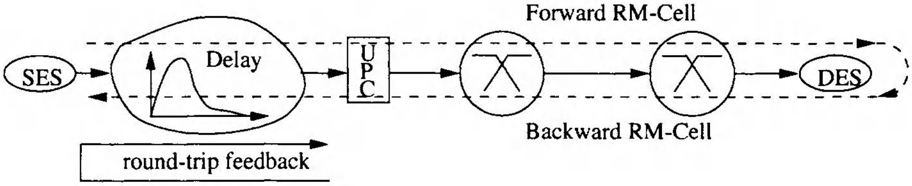


Figure 4.7. Policing ABR traffic

5.1.3 ABR Conformance Definition and Policing

To control ABR sources and to check whether or not they respond to the feedback information, a conformance definition is introduced in standardisation. An example of a conformance definition, for an ABR connection based on the Dynamic Generic Cell Rate Algorithm (DGCRA), has been defined by ITU-T [I371] and the ATM Forum [ATM95]. The conformance definition is a part of the traffic contract which defines a reference algorithm used to define whether the cells passing a measuring point located at the UNI are conforming or not. A network operator may use a Usage Parameter Control (UPC) which, based on the conformance definition, defines whether a connection is compliant or not. The UPC may mark or discard non-conforming cells.

What is new for ABR connections is the variable lag which exists between the moment a rate change is communicated to the source and the time this change is observed at the interface. This can be seen in Figure 4.7. Forward RM-cells generated by the SES are inserted in the data flow and contain a value for the ER at which the source would like to transmit. These RM-cells are looped back by the DES to the SES. Nodes in between can access the ER field and lower the ER in case of congestion. Depending on the distance between source and interface and on the background traffic, it will take a variable time before the RM cells arrive at the interface. In order to have the most recent value of the ER, the policing function must also check the flow of RM cells in the backward direction of the connection. This is also new compared to policing CBR and VBR traffic where only the direction from source to destination has to be monitored.

In order to cope with the variable lag between source and interface two time constants τ_2 and τ_3 are introduced which are respectively an upper bound and a lower bound of this round trip delay. Because of the variable available bandwidth to which the source must adapt itself, the source traffic characteristics will be altered during the lifetime of the connection. The GCRA (used for the conformance definition for CBR and VBR) is a static algorithm in the sense that its two parameters, the increment value I (e.g. the inverse of the PCR) and the limit L (e.g. the CDV tolerance) are not allowed to vary (without re-negotiation of the traffic contract). The DGCRA has the same two parameters as the GCRA but the increment parameter is allowed to vary (without re-negotiation of the traffic contract) between the inverse of the PCR and the inverse of the MCR. The computation of this varying increment is not an easy task because a rate change conveyed by a backward RM-cell received at the interface at a given time may be applied to the forward cell flow after a variable delay t ($\tau_3 \leq t \leq \tau_2$).To be on the safe side, the DGCRA schedules rate increases conveyed in the backward RM-cell flow after a delay of τ_3 and rate decreases after a delay of τ_2.

Two algorithms have been proposed to compute the variable increment. Algorithm “A” provides the tightest conformance according to the delay bounds but is rather complex to implement and a simpler algorithm “B” has been defined which is much less accurate. The tightness of the rate conformance of the DGCRA may therefore be reduced due to the difficulty of following the rate changes at the measuring point. Moreover, the algorithm does not perform a CCR conformance while switches use the CCR to estimate the VC rates (e.g. EPRCA, ERICA). Therefore, the absence of a CCR conformance can lead to misbehaviour of the feedback control of switches that make use of the CCR if a source does not properly set the CCR to the ACR. The other problem is that the algorithms A and B only keep track of the ER of the backward RM cells. However, such an ER-conformance algorithm will be inappropriate for a binary switch that conveys the congestion information by means of the CI bit [CER96].

5.1.4 ABR Experiments

Several trials have been performed to check the ability of the ABR flow control to adapt to changing bandwidth availability and the performance of TCP traffic over ABR. In [CER98] and [BCN98] a first set of experiments using an ABR switch from Able Communication show that if we want to achieve a high ABR efficiency in a fast varying available bandwidth the frequency of RM cell generation needs to be at least that of the bandwidth variation. This can have an implication for CAC as we need to reserve a certain amount of bandwidth for ABR. The experiments carried out in a WAN environment show that in that case high buffer capacities are needed even with a switch based on ERICA.

From the experimental study of TCP over ABR, we may conclude that even a simple threshold based ABR Explicit Rate notification implementation like the one used in [BCN98] assures nearly full TCP efficiency with ATM buffers much smaller than those required for UBR. The ABR buffer requirements show to be independent of the number of TCP connections carried by the ABR VC, which is an interesting behaviour if LAN traffic needs to be carried. To make the ABR mechanism robust with respect to the VBR background profile, a minimum amount of buffering is required. For very small ABR buffers and long VBR ON periods, the ABR flow control mechanism fails and TCP throughput collapse is observed similar to UBR.

5.2 INTELLIGENT PACKET DISCARD SCHEMES FOR UBR

The absence of any QoS guarantee for the UBR Service Category may lead to a low throughput. The loss of a single cell in the network of an AAL5 PDU frame inevitably leads to the discard of the whole frame at the destination. Delivering such a corrupted frame leads to an important waste of network resources and may drop the effective throughput drastically. This is in particular true in a broadband network where the high data rate and the long distances force the retransmit/recovery mechanisms to retransmit a high number of frames already sent out since the corrupted cell (depending on the window size). Another disadvantage of the absence of any traffic and congestion control mechanism in the UBR Service Category is the lack of fairness between the different connections that share the bandwidth using UBR. Indeed, connections which lose cells will be forced by the transport protocol (e.g. TCP) to slow down their transmission rate, allowing other connections to use more bandwidth and buffer space. These observations have lead to the introduction of intelligent packet discarding mechanisms. In what follows, we discuss Partial Packet Discard, Early Packet Discard, Per VC Accounting and Per VC Queueing.

5.2.1 Partial Packet Discard

Partial Packet Discard (PPD) (see e.g. RF95]) (also called Packet Tail Discard [TUR96], Partial Packet/Frame Drop [LNO96], Drop Tail [FCH94], Tail Dropping [KKT96]) is a packet discarding scheme which drops the remainder of a frame, apart from the End of Message (EOM) cell, as soon a a cell loss occurs. The EOM cell is needed by the destination end station to delineate the beginning of a new frame. The scheme is called partial, as only those cells of a frame arriving after the occurrence of a loss are dropped. There is no de-queueing of already accepted cells. The scheme recognizes the beginning of a new frame by inspecting the Payload Type Identifier (PTI) field in the header of incoming cells. The PTI contains an indication of the EOM.

5.2.2 Early Packet Discard

This scheme enforces a network element to drop an entire frame (i.e. all its cells) when the first cell of that frame arrives at a buffer which exceeds a predefined threshold. There are several methods to set the threshold. In [RF95], a fixed threshold is proposed, while in [KKT96] a variable threshold is used based on the number of already accepted frames. The Random Early Detection (RED) algorithm proposed in [LNO96] uses the observed average queue size to define a probability by which a frame is dropped in case of congestion. Although the throughput may be increased considerably using EPD, the fairness problem still remains, due to the fact that EDP does not consider the number of already accepted frames per VC when dropping a newly arriving frame. The following methods try to solve this problem.

5.2.3 Per VC Accounting

In a per VC accounting scheme, frames are discarded not only on basis of the buffer occupation level, but also based on the origin of the frames in the buffer. Apart from the buffer occupation threshold, the scheme computes a Fair Buffer Share (FBS) defined to be FBS = K × (Total Buffer Occupation / number of active VCs). The constant K is a factor for the buffer occupation and is chosen as $1 < K < 2$. As soon as the threshold of the buffer occupation is reached, then all frames of overloading connections (i.e. connections which use more than the number FBS of buffer places) are dropped. Frames of non-overloading connections are still accepted. This mechanism has the advantage that it co-operates better with the transport protocol (e.g. TCP). Indeed, when a cell of a connection is lost, the transport protocol will enforce this connection to drop its frame generation rate, implying a lower cell arrival rate at the buffer, resulting in a highly probable situation where the buffer occupation of that connection is below the FBS. Hence, a connection that has experienced a cell loss has a higher probability of having its next frame(s) accepted, even if the congestion is persisting.

5.3 GUARANTEED FRAME RATE

The mapping of the GFR frame level guarantee onto an appropriate cell level guarantee is achieved based on the identification of the frames to which the service guarantees apply. This can be done by using a modified GCRA(1/MCR, Burst Tolerance(MBS)+CDVT), where MBS = 2*CPCS-SDU size (in cells). Three sample GFR implementations are proposed in [ATM99]:

(i) *GFR implementation using Weighted Fair Queuing (WFQ) and per-VC accounting*

This implementation serves an individual VC at a rate of at least MCR using a WFQ scheduler. The buffer management is based on a per-VC accounting so that each ATM connection can have its own part of available bandwidth and buffer.

(ii) *GFR implementation using tagging and FIFO queue*

In that case the cell rate guarantee of GFR cannot be provided by the service discipline and a tagging function is needed to identify cells eligible for service guarantee. The modified GCRA(1/MCR, Burst Tolerance(MBS)+CDVT) is used to determine which cells to tag.

(iii) *GFR implementation using Differential Fair Buffer Allocation*

DFBA [GOY98] uses per-VC accounting together with static and dynamic thresholds in a FIFO buffer which estimate the bandwidth used by the connection. If the buffer occupancy of each active VC is maintained at a desired threshold, then the output rate of each VC can also be controlled.

As GFR looks promising with respect to the efficient transport of TCP traffic, the performance of TCP over GFR has thoroughly being investigated. Several simulations indicate that the GFR implementation based on FIFO queuing and tagging is not able to provide the cell rate guarantee to a TCP source while the other implementation allows to provide satisfactory performance of TCP over GFR as long as enough buffers are provided. It has been shown [BON97], [CEN98] that WFQ cannot guarantee reserved bandwidths with limited buffers. DFBA can guarantee TCP throughputs in proportion to the fraction of the average buffer occupied by each VC. Nevertheless this throughput is only achievable for low buffer allocation.

6 CONCLUSIONS

In this paper, an overview of the various Service Categories and Transfer Capabilities in ATM networks is presented, together with the traffic control and congestion control mechanisms that support the QoS guarantees offered by these categories. Today, a number of concepts and mechanisms are specified or even standardised, such as the traffic parameter definition using the GCRA, the leaky bucket algorithm for UPC, the rate based flow control scheme for ABR traffic, etc. Other mechanisms, such as CAC, are system dependent and remain still today a topic of intensive research and competition in commercial ATM products. Another development that has heavily influenced the traffic management architecture is the world-wide use of Internet and the related protocols and applications. In particular the possibility to offer QoS guarantees is an important topic in the discussion on Internet and ATM. Studies and experiments have shown that, in order to carry Internet traffic in an efficient and economical way over an ATM network, new control mechanisms that operate on layers above ATM are needed (for example to guarantee the goodput of AAL5 PDUs). The current development of the GFR Service Category illustrates this trend.

In spite of the fact that already a lot of effort has been put in Traffic Management research and development activities, there remain many questions unanswered and further research effort is needed in this area.

References

[ATM95] ATM Forum Technical Committee Traffic Management Working Group, "*ATM Forum Traffic Management Specification Version 4.0*", ATM Forum, October 1997

[ATM99] Draft Traffic Management Specification Version 4.1, ATM Forum btd-tm-02.02, 1999.

[BCN98] C. Blondia, O. Casals, J. Nelissen, "*Evaluation of the Available Bit Rate Category in ATM Networks*", IEEE Workshop on Communications, Oxford (USA), October 1998.

[BON97] O. Bonaventure, "*Simulation Study of TCP with the Proposed GFR Service Category*", Conference on High Performance Networks for Multimedia Applications, Dagstuhl (Germany), June 1997.

[BOY92a] P.E. Boyer, F.M. Guillemin, M.J. Servel and J-P. Coudreuse, "*Spacing Cells Protects and Enhances Utilization of ATM Network Links*", IEEE Network Magazine, Vol. 6, No. 5, September 1992.

[BOY92b] P.E. Boyer, D. Tranchier, "*A Reservation Principle with Applications to the ATM Traffic Control*", Computer Networks and ISDN Systems, 24, North Holland, 1992, pp. 321-334

[BUR90] J. Burgin, "*Dynamic Capacity Management in the BISDN*", Int. Journal of Digital and Analog Communication Systems, Vol. 3, pp. 161-165, 1990.

[BUR91] J. Burgin and D. Dorman, "*Broadband ISDN Resource Management: The Role of Virtual Paths*", IEEE Comm. Mag., Vol. 29, No. 10, pp. 44-48, 1991.

[BUT91] M. Butto, E. Cavallero, A. Tonietti, "*Effectiveness of the "Leaky Bucket" Policing Mechanism in ATM Networks*", IEEE JSAC, Vol. 9, No. 3, pp. 335-342, 1991.

[BS97] F. Brichet and A. Simonian, "*Measurement-based CAC for video applications using SRB service*", Proceedings of te PMCCN Conference, IFIP WG 6.3 and &.2, Tsukuba (Japan), November 1997, p.285-304.

[BS98] F. Brichet and A. Simonian, "*Conservative Gaussian models applied to measurement-based admission control*", Proceedings of the 6th IEEE/IFIP International Workshop on Quality of Service 98, Napa (USA), May 1998.

[CER96] L. Cerda, O. Casals, "*Improvements and Performance Study of the Conformance Definition for the ABR Service in ATM Networks*", ITC Specialist Seminar on Control in Communications, Lund, Sweden, September 1996.

[CEN98] F. Cerdan, O. Casals, "*A Per-VC Global FIFO Scheduling Algorithm for Implementing the New ATM GFR Service*", IFIP/IEEE Int. Conf. On Management of Multimedia Network and Services'98, Versailles (France), November 1998.

[CER98] L. Cerda, O. Casals, "*Experimental Analysis ofan ER Switchfor the ABR Service in ATM Networks*", Actas IV Jornadas de Informatica, Las Palmas de Gran Canaria, (Spain), July 1998.

[COST242] J. Roberts, U. Mocci and J. Virtamo (Eds), "*Broadband Network Teletraffic*", Final Report of Action COST 242, Springer Verlag 1996

[CSZ 92] D.D. Clark, S. Shenker and L. Zhang, "*Supporting Real Time Applications in an Integrated Services Packet Network: Architecture and Mechanisms*", ACM SIGCOM'92, 1992.

[COH79] J. Cohen, "*The multiple phase service network with generalized processor sharing*", Acta Informatica, 12, 1979, p.245-284

[DB98] T. Daniels and C. Blondia, "*Asymptotic behaviour of a discrete-time queue with long range dependent input*", Proceedings IEEE INFOCOM '99.

[DKS89] A. Demers, S. Keshav, S. Shenker, "*Analysis and Simulation of a Fair Queueing Algorithm*", ACM SIGCOM'89, 1989.

[EM93] A. Elwalid and D. Mitra, "*Effective bandwidth of general Markovian traffic sources and admission control of high speed networks*", IEEE/ACM Trans Networking, 1, June 1993

[EMW95] A. Elwalid, D. Mitra and R. Wentworth, "*A new approach to allocating buffers and bandwidth to heterogeneous regulated traffic in an ATM node",* IEEE J. Selected Areas in Comm., 13(6), August 1995, p. 1115-1128

[EXP96] Deliverable 6 of the ACTS Project AC094 EXPERT, "*Specification of Integrated Traffic Control Architecture*", September 1996.

[FCH94] C. Fang, H. Chen and J. Hutchins, "*A simulation study of TCP performance in ATM networks*", Proceedings of IEEE INFOCOM '94, Vol.2, San Francisco, 1994, p. 1217-1223

[GK97] R. Gibbens and F. Kelly, "*Measurement-based connection admission control*", Proceedings ITC 15, Washington June 1997, Teletraffic Contributions for the Information Age, Eds. V. Ramaswami and P. Wirth, Elsevier, 1997, p.879-888.

[GAN91] R. Guerin, H. Ahmadi and M. Naghshineh, "*Equivalent capacity and its application to bandwidth allocation in high speed networks*", IEEE J. Selected Areas in Comm., 9, 1991,p.968-981

[GIL91] H. Gilbert, O. Aboul-Magd, V. Phung, "*Developing a Cohesive Traffic Management Strategyfor ATM Networks*", IEEE Comm. Mag., Vol. 29, No. 10, 1991.

[GOL90] S. Golestani, "*Congestion-free Transmission of Real-Time Traffic in Packet Networks",* Proc. IEEE Infocom´90, pp. 527-542, San Francisco, CA, June 1990.

[GOY98] R. Goyal, R. Jain, S. Fahmy and B. Vandalore, "Providing Rate Guarantees to TCP over the ATM GFR Service", Proceedings 23rd Annual Conference on Local Computer Networks 1998, Lowel, MA, October 1998, pp. 390-398.

[GUE96] R. Guerin, J. Heinanen, "*UBR+ Service Category Definition*", ATM Forum contribution No. 96-1598, December 1996.

[GUI92] F. Guillemin, P. Boyer and L. Romoeuf, "*The spacer-controller : architecture and first assessments*", Proc. IFIP Workshop on Broadband Communications, Estoril, Portugal, 1992.

[I371] CCITT Draft Recommendation I.371 (now ITU-T 1.371), "*Traffic Control and Resource Management in B-ISDN*", Melbourne, Dec. 1991.

[JAI95] R. Jain et al., "*A Sample Switch Algorithm*", ATM Forum contribution No. 95-0178R1, February 1995.

[JGK96] R. Jain, R. Goyal, S. Kalyanaraman, S. Fahmy and F. Lu, "*TCP/IP over UBR",* ATM Forum contribution 96-0179

[KEL91] F. Kelly, "*Effective bandwidths at multi-class queues*", Queueing Systems, 9, 1991, p.4-15

[KKT96] K. Kawahara, K. Kitajima, T. Takine and Y. Oie, "*Performance evaluation of selective cell discard schemes in ATM networks*", Proceedings of IEEE INFOCOM '96, Vol.3, San Francisco, March 1996, p. 1054-1061

[KMI92] C.R. Kalmanek, S.P. Morgan, R. C. Restrick III, "*A High-Performance Engine for ATM networks*", ISS'92, 1992.

[KRO90] H. Kroner, "*Comparative Performance Study of Space Priority Mechanisms for ATM Channels*", IEEE Infocom'90, San Francisco, June 1990.

[LNO96] T. Lakshman, A. Neidhardt and T. Ott, "*The drop from front strategy in TCP over ATM*", Proceedings of IEEE INFOCOM '96, Vol.3, San Francisco, March 1996, p. 1242-1250

[NIE90] G. Niestegge, "*The Leaky Bucket Policing Method in ATM Networks*", Int. Journal of Digital and Analog Communication Systems. Vol. 3, pp. 187-197, 1990.

[RAG91] E.P. Rathgeb, "*Modeling and Performance Comparison of Policing Mechanisms for ATM Networks*", IEEE JSAC, Vol. 9, No. 3, pp. 325-334, 1991.

[RF95] A. Romanov and S. Floyd, "*Dynamics of TCP traffic over ATM networks*", IEEE J. Selected Areas in Comm., 13 (4), 1995, p.633-641

[ROB93] J. Roberts (Ed), "*Performance evaluation and design of multiservice networks*", COST 224, Commission of the European Communities, October 1992, Final Report

[ROB98] J. Roberts, "*Realising quality of service guarantees in multi-service networks*", Proceedings of the PMCCN Conference, IFIP WG 6.3 and &.2, Tsukuba (Japan), November 1997, p.271-283.

[ROJ94] J.W. Roberts, "*Weighted Fair Queueing as a Solution to Traffic Control Problems*", COST 242 MID-TERM Seminar, L'Aquila, Italy, Sep. 1994.

[ROL94] L. Roberts, "Enhanced Proportional Rate Control Algorithm (EPRCA)", ATM Forum contribution n° 94-0735Rl, August 1994.

[RV91] J, Roberts and J.Virtamo, "*The superposition of periodic cell arrival streams in an ATM multiplexer*", IEEE Trans. Comm., 39(2), February 1991, p.298-303,

[SKL94] A. Skliros, "*Characterizing the Worst Traffic Profile passing through an ATM-UNI* ", Proceedings of the 2nd IFIP Conference on Performance Modelling and Evaluation of ATM Networks, Bradford (U.K.), 1994.

[SUM88] S. Sumita, T. Ozawa, "*Achievability of Performance Objectives in ATM switching Nodes* ", Int. Seminar on Performance of Distributed and Parallel Systems, pp. 45-46, Kyoto, Japan, Dec. 1988.

[TUR86] J. Turner, "*New Directions in Communications (or which way in the information age?)",* Zurich Seminar on Digital Communications, pp. 25-32, March 1986.

[TUR96] J.S. Turner, "*Maintaining high throughput during overload in ATM switches*", Proceedings of IEEE INFOCOM '96, Vol.1, San Francisco, March 1996, p.287-295

[UNI3.1] ATM Forum, ATM User-Network Interface Specification, September 1993.

[VR89] J.T.Virtamo and J.W. Roberts, "*Evaluating buffer requirements in an ATM multiplexer*", Proceedings IEEE Globecom 89, 1989

[WAL91] E. Wallmeier, T. Worster, "*A Cell Spacing and Policing Device for Multiple Virtual Connections on one ATM Pipe*", Proc. RACE R1022 Workshop on ATM Network Planning and Evolution, London, 1991.

[VER91] D. Verma, H. Zhang and D. Ferrari, "*Guaranteeing Delay Jitter Bounds in Packet Switching Networks* ", Proc. Tricomm´91, Chapel Hill, NC, pp. 35-46, April 1991

[YIN94] N. Yin and M. G. Hluchyj, "*On Closed-Loop Rate Control for ATM Cell Relay Networks*", IEEE Infocom´94, pp.99-108.

[ZHA91] L. Zhang, "*Virtual Clock: A New Traffic Control Algorithm for Packet Switching Networks*", ACM Transactions on Computer Systems, Vol. 9, No. 2, pp. 101-124, 1991.

Chris Blondia obtained his Master in Science and Ph.D in Mathematics, both from the University of Ghent (Belgium) in 1977 and 1982 respectively. In 1983 he joined Philips Belgium, where he was a researcher between 1986 and 1991 in the Philips Research Laboratory Belgium (PRLB) in the group of Computer and Communication Systems. Between August 1991 and end 1994 he was an Associate Professor in the Computer Science Department of the University of Nijmegen (The Netherlands). In 1995 he joined the Department of Mathematics and Computer Science of the University of Antwerp, where he is a professor and head of the research group "Performance Analysis of Telecommunication Systems". His main research interests are related to traffic modelling, switching architectures, traffic management, medium access control protocols, etc... He has published a substantial number of papers in international journals and conferences on these research areas.

Olga Casals obtained her Ph.D in Telecommunication Engineering in 1986 from the Polytechnic University of Catalonia in Barcelona (Spain). In 1983 she joined the Computer Architecture Department of the Polytechnic University of Catalonia. In 1994 she became Full Professor and since 1998 she is head of the Department. She is also leading a research group on "Broadband Communications". Her main interests are related to performance analysis and traffic management. She has published a substantial number of papers in international journals and conferences on these research areas.

Chapter 5

A COMPARATIVE PERFORMANCE ANALYSIS OF CALL ADMISSION CONTROL SCHEMES IN ATM NETWORKS

Khaled Elsayed
Department of Electronics and Communications Engineering
Faculty of Engineering, Cairo University, Giza, Egypt 12613
E-mail: khaled@ieee.org

Harry Perros
Department of Computer Science
North Carolina State University, Raleigh, NC 27695, USA
E-mail: hp@eos.ncsu.edu

Abstract: Connection Admission Control (CAC) is one of the primary mechanisms for preventive congestion control and bandwidth allocation in ATM networks. A substantial number of CAC schemes have been proposed. In this paper, we review the salient features of some of these algorithms. We also provide a comparative study of the performance of CAC schemes devised to meet certain quality of service requirements expressed in terms of cell loss probability and maximum delay.

Keywords: Call Admission Control, Traffic Management, ATM Networks, Quality of Service, Effective Bandwidth, Diffusion Approximation, PGPS, EDF, SP

1. INTRODUCTION

In recent years, there has been a tremendous growth in the development and deployment of ATM networks. One area which is of significant importance to ATM networks is traffic management. Congestion control is one of the primary mechanisms for traffic management. The primary role of a network congestion control procedure is to protect the network and the user in order to achieve network performance objectives and optimize the usage of network resources. In ATM-based B-ISDN, congestion control should support a set of ATM quality-of-service classes sufficient for all foreseeable B-ISDN services.

Congestion control schemes can be classified into preventive control and reactive control. In preventive congestion control, one sets up schemes which prevent the occurrence of congestion. In reactive congestion control, one relies on feedback information for controlling the level of congestion. Both approaches have advantages and disadvantages. In ATM networks, a combination of these two approaches is currently used in order to provide effective congestion control. For instance, CBR and VBR services use preventive schemes and ABR service is based on a reactive scheme.

Preventive congestion control involves the following two procedures: connection admission control (CAC) and bandwidth enforcement. ATM is a connection-oriented service. Before a user starts transmitting over an ATM network, a connection has to be established. This is done at connection set-up time. The main objective of this procedure is to establish a path between the sender and the receiver. This path may involve one or more ATM switches/routers. On each of these ATM switches, resources have to be allocated to the new connection.

The connection set-up procedure runs on a resource manager (which is typically a workstation attached to the switch). The resource manager controls the operations of the switch, accepts new connections, tears down old connections, and performs other management functions. If a new connection is accepted, bandwidth and/or buffer space in the switch is allocated for this connection. The allocated resources are released when the connection is terminated.

Call admission control deals with the question as to whether a switch can accept a new connection or not. Typically, the decision to accept or reject a new connection is based on the following two questions:

1. Does the new connection affect the quality-of-service of the connections that are currently being carried by the switch?
2. Can the switch provide the quality-of-service (QOS) requested by the new connection?

The answer to these questions is a function of the connections' traffic characteristics, the QOS requested, and the network state.

Call admission control schemes have been developed so that they satisfy a particular quality of service. In packet networks, the two major QOS attributes are packet loss and packet delay. A new connection may request from the network a certain bound on packet loss and packet delay. Moreover, these bounds can be deterministic or statistical. For deterministic

QOS, a new connection would request a maximum end-to-end packet/cell delay or a maximum threshold on the value of packet/cell loss probability. On the other hand, for statistical QOS, a connection would request that its packets experience, for example, a mean end-to-end delay or a mean packet/cell loss probability.

Call admission control schemes may be classified into a) non-statistical allocation, or peak bandwidth allocation, and b) statistical allocation. Non-statistical allocation can be used to enforce deterministic bounds on the requested QOS of a connection. Statistical allocation can be used to enforce either deterministic or statistical QOS bounds. Below we examine the two types of call admission control. The advantage of peak bandwidth allocation is that it is easy to decide whether to accept a new connection or not. The disadvantage of peak allocation is that unless connections transmit at peak rates, the output port link may be grossly under-utilized.

In statistical allocation, bandwidth for a new connection is not allocated on per peak rate basis. Rather, the allocated bandwidth is less than the peak rate of the source. As a result, the sum of all peak rates may be greater than the capacity of the output link. Statistical allocation makes economic sense when dealing with bursty sources, but it is difficult to carry out effectively. This is because of difficulties in characterizing the arrival process of ATM cells and the lack of understanding as to how this arrival process is shaped deep in the ATM network.

Another difficulty in designing a connection admission control algorithm for statistical allocation is that decisions have to be done on the fly, and therefore they cannot be CPU intensive. Typically, the problem of deciding whether to accept a new connection or not may be formulated as a queueing problem. The connection admission control algorithm has to be applied to the buffer of each output port. If we isolate an output port and its buffer from the rest of the switch, we will obtain the queueing model shown in figure 5.1. This type of queueing structure is known as an ATM multiplexer. It represents a number of ATM sources feeding a finite capacity queue, which is served by a server (the output port). The service time is constant equal to the time it takes to transmit an ATM cell.

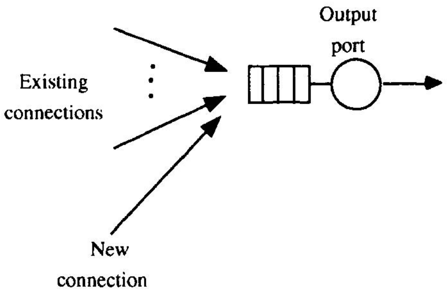


Figure 5.1. An ATM multiplexer

Now, let us consider the cell loss probability as the requested QOS, and let us assume that the cell loss probability of the existing connections is satisfied. The question that arises is whether the cell loss probability will still be maintained if the new connection is added. This can answered by solving this ATM multiplexer with the existing and new connections. However, the solution to this problem is very difficult and CPU intensive (see for example Elsayed and Perros [8] and Li [16]). It gets even more complicated if we assume complex arrival processes. In view of this, a variety of different bandwidth allocation algorithms have been proposed which are based on different approximations, or different types of schemes which do not require the solution of such a queueing problem.

In this paper, we will examine some of the connection admission control algorithms that have been proposed for statistical allocation. Before we proceed, however, we review briefly the various traffic models that have been proposed in the literature.

1.1 CHARACTERIZATION OF AN ARRIVAL PROCESS

Prior to the advent of ATM networks, performance models of telecommunication systems were typically developed based on the assumption that arrival processes are Poisson distributed. That is, the time between successive arrivals is exponentially distributed. In some cases, such as in public PSTN switching, extensive data collection actually supported the Poisson assumption.

Over the last few years, we have gone through several paradigm shifts regarding our understanding as to how to model an ATM source. Following the first performance models which were based on the Poisson assumption, or the Bernoulli assumption, it became apparent that these traffic models did not capture the notion of burstiness that is present in traffic resulting from applications such as moving a data file and packetized encoded video. Thus, there was a major shift towards using arrival processes of the on/off type.

The ATM Forum has defined a standard mechanism for specifying a connection's traffic [1]. A connection is specified by the tuple (PCR, CDVT, SCR, MBS) where PCR is the peak cell rate, CDVT is the cell delay variation tolerance, SCR is the sustainable cell rate, and MBS is maximum burst size. Using the peak rate and the cell delay variation, one can effectively police the peak rate. Also, using the maximum burst length, one can estimate a cell delay variation that can be used to police the sustainable rate. These parameters can be enforced using the GCRA algorithm of the ATM forum, which is equivalent to a dual leaky-bucket mechanism [1].

Most of the CAC schemes use the tuple of parameters (PCR, SCR, MBS) of the existing and new connections when making a decision on accepting or rejecting a connection. The parameter CDVT is a function of the user and network equipment and has little effect on traffic characterization of the connection. The tuple (PCR, SCR, MBS) can be used to specify a variety of traffic models. A model that introduces statistical variation into the model specified by (PCR, SCR, MBS) is the on/off source model. A popular instance of on/off sources is the Interrupted Poisson Process (IPP) or its discrete-time counterpart the Interrupted Bernoulli Process (IBP). In an IPP, there is an active period during which arrivals occur in a Poisson fashion, followed by an idle period during which no arrivals occur. These two periods are exponentially distributed, and they alternate continuously. An IBP is defined similarly, only the arrivals during the active period are Bernoulli distributed, and the two periods are geometrically distributed. Another way of describing a source is using the fluid approach. Here arrivals occur with a continuous rate during the active period. This defines an on/off fluid source or equivalently an Interrupted Fluid Process (IFP).

1.2 CLASSIFICATION OF CONNECTION ADMISSION SCHEMES

In this paper we consider two main categories of CAC schemes: a) schemes for bounding cell loss probability for connections, and b) schemes for bounding cell delay. A variety of different connection admission schemes have been proposed in the literature. Some of these schemes require an explicit traffic model and some only require traffic parameters such as the peak and average rate. In this paper we review some of these schemes. For presentation purposes, the schemes have been classified as follows:

- CAC schemes based on the cell loss probability. These include
    1. Effective Bandwidth (Equivalent Capacity)
    2. Diffusion Approximation
    3. Upper Bounds of the cell loss probability
- CAC schemes based on cell delay. These are usually associated with certain scheduling methods. Our study includes

1. Weighted Fair Queueing (WFQ) or Packet-by-Packet Generalized Processor Sharing (POPS) scheduling
2. Delay-Earliest Deadline First (EDF) scheduling
3. Static Priority (SP) scheduling

This classification was based on the underlying principle that was used to develop a CAC scheme and its targeted QOS objective. The remainder of this paper is organized as follows. In section 2, we review the salient features of the four CAC schemes mentioned above that are based on the cell loss probability. Extensive numerical comparisons between three of these schemes are then given in subsections 2.5 to 2.11. In section 3, we review the CAC schemes mentioned above that are based on the cell delay. Numerical comparisons between three of these schemes are given in section 3.5. Other CAC schemes are described in section 4.

2. CAC SCHEMES FOR THE CELL LOSS PROBABILITY QOS

2.1 EFFECTIVE BANDWIDTH/EQUIVALENT CAPACITY

Let us consider a single source feeding a finite capacity queue. Then, the effective bandwidth of the source is the service rate of the queue that corresponds to a cell loss of ε. The effective bandwidth for a single source can be derived as follows (see Guerin, Ahmadi, and Naghshineh [13]). Each source is assumed to be an IFP. Let R be its peak rate, r the fraction of time the source is active, and b the mean duration of the active period. An IFP source can be completely characterised by the vector (R, r, b). Let us now assume that the source feeds a finite capacity queue with constant service time. Let K be the capacity of the queue. The effective bandwidth e is given by:

$$e = \frac{a - K + \sqrt{(a-K)^2 + 4Kar}}{2a} R \tag{5.1}$$

where $a = \ln(1/\varepsilon)b(1-r)R$.

In the case of N sources, and given that the buffer has a capacity K, the effective bandwidth is again the service rate e which ensure that the cell loss for all sources is less than or equal to ε. Guerin, Ahmadi, and Naghshineh [13] proposed the following approximation for multiple sources:

$$e = \min(\rho + a'\sigma, \sum_{i=1}^{N} e_i) \tag{5.2}$$

where e_i is the effective bandwidth of the *i*th source calculated using expression (5.1), and $\sum_{i=1}^{N} e_i$ is the sum of all the individual effective bandwidths, ρ is the total average bit rate, i.e. $\rho = \sum_{i=1}^{N} \rho_i$, where $\rho_i = r_i R_i$ is the mean bit rate of the *i*th source, $\sigma^2 = \sum_{i=1}^{N} \sigma_i^2$, where σ_i^2 is the variance of the bit rate of the *i*th source, $\sigma_i^2 = \rho_i(R_i - \rho_i)$, and $a' = \sqrt{-2\ln(\varepsilon) - \ln 2\pi}$.

Some studies (see Choudhury, Lucantoni, and Whitt [4] and Elsayed and Perros [8]) have clearly indicated the inaccuracy of effective bandwidth methods in some situations. In particular, the effective bandwidth method fails when a bufferless system subject to an input traffic has a small probability that the traffic load exceeds the link capacity. In the effective bandwidth approach, this probability is assumed to be close to one (and is taken as one in the calculations). Rege [21] compares various approaches for effective bandwidth and proposes some modifications to enhance the accuracy of the scheme. Elwalid et al. [9] proposed a method in which they combined Chernoff bounds with the effective bandwidth approximation to overcome the shortcomings of the effective bandwidth. This method permits better accuracy than effective bandwidth for the case of a bufferless (or a small buffer for that matter) multiplexer that can achieve substantial statistical gain. However, in some other cases, the method does not improve the accuracy of the effective bandwidth.

Kulkarni, Gun, and Chimento [15] considered the effective bandwidth vector for two-priority on/off source. Chang and Thomas [3] introduced a calculus for evaluating source effective bandwidth at output of multiplexers and upon demultiplexing or routing. On-line evaluation of effective bandwidth have been proposed by De Veciana, Kesidis and Walrand [23]. Duffield et al. [7] proposed maximum entropy as a method for characterizing traffic sources and their effective bandwidth.

2.2 DIFFUSION APPROXIMATION

Gelenbe, Mang and Onvural [12] proposed a scheme that uses statistical bandwidth obtained from a closed-form expression based on the diffusion approximation models. Specifically, a diffusion process with absorbing boundaries and jumps was used to analyze approximately a discrete-time ATM multiplexer with N IFP sources. Two models are used: one for a finite buffer (FB) ATM multiplexer and the other for an infinite buffer (IB) ATM multiplexer. In the IB model, the cell loss probability is estimated by the overflow probability, which is the overall probability of exceeding the actual buffer capacity (*K*) in a system with an unlimited buffer size. The cell loss probability calculated from these two models is:

$$L_{FB} = \frac{1}{\sqrt{2\pi}} e^{\frac{2K}{\alpha}(\rho - C)} e^{-\frac{(\rho-C)^2}{2\sigma^2}} \tag{5.3}$$

$$L_{IB} = \frac{\sigma}{\rho\sqrt{2\pi}} e^{\frac{2K}{\alpha}(\rho - C)} e^{-\frac{(\rho-C)^2}{2\sigma^2}} \tag{5.4}$$

For *N* IFP sources with parameters (R_i, r_i, b_i), we have $\sigma^2 = \sum_{i=1}^{N} \sigma_i^2$ is the total variance where $\sigma_i^2 = \rho_i(R_i - \rho_i)$ and $\rho_i = r_i R_i$, $\rho = \sum_{i=1}^{N} \rho_i$, $\alpha = \sum_{i=1}^{N} \rho_i CV_i^2$ is the instantaneous variance of the arrival process where $CV_i^2 = \frac{1-(1-\beta_i T_i)^2}{(\beta_i T_i + \gamma_i T_i)^2}$ and $T_i = \frac{1}{R_i}$, $\frac{1}{\beta_i} = b_i$ the mean on period, and $\frac{1}{\gamma_i}$ is the mean off period of the *i*th source. Finally, C^{1} is the time required to transmit one cell.

Let us define the statistical bandwidth as the bandwidth that needs to be allocated for the multiplexed connections in order to keep the cell loss probability below ε (the required cell loss probability). We get two expressions (one for the FB and the other for the IB model respectively) for the statistical bandwidth:

$$C_{FB} = \rho - \delta + \sqrt{\delta^2 - 2\sigma^2\omega_1} \tag{5.5}$$

$$C_{IB} = \rho - \delta + \sqrt{\delta^2 - 2\sigma^2\omega_2} \tag{5.6}$$

where $\delta = \frac{2K}{\alpha}\sigma^2$, $\omega_1 = \ln(\varepsilon\ \sqrt{2\pi})$, and $\omega_2 = \ln(\varepsilon\rho\sqrt{2\pi}) - \ln(\sigma)$. It is possible to take:

$$C_{df} = \max(C_{FB}, C_{IB})$$

as the (worst-case) estimate of the statistical bandwidth. The procedure to admit or reject a new connection is then summarized as follows:

1) At any time keep a record of the quantities $\rho = \sum\rho_i, \sigma^2 = \sum\sigma_i^2, \alpha = \sum\rho_i CV_i^2$ of the existing connection
2) When a new connection arrives, update $\rho = \sum\rho_i, \sigma^2 = \sum\sigma_i^2, \alpha = \sum\rho_i CV_i^2$ to include the new connection
3) If the resulting $C_{df} < C$, then admit the new connection.
4) Else reject the new connection and update $\rho = \sum\rho_i, \sigma^2 = \sum\sigma_i^2, \alpha = \sum\rho_i CV_i^2$ to exclude the rejected connection effect.

2.3 UPPERBOUNDS OF THE CELL LOSS PROBABILITY

Several connection admission schemes have been proposed which are based on an upper bound for the cell loss probability. Saito [22] proposed an upper bound based on the average number of cells that arrive during a fixed interval (*ANA*), and the maximum number of cells that arrive in the same fixed interval (*MNA*). The fixed interval was taken to be equal to *D*=2, where *D* is the maximum admissible delay in a buffer. Using these parameters, the following upper bound was derived. Let us consider a link serving *N* connections, and let $p_i(j), i$ = 1,2,•••,*N*, and *j* = 1,2,•••, be the probability that *j* cells belonging to the *i*th connection arrive during the period *D=2*. Then, the cell loss probability *CLP* can be bounded by:

$$CLP \leq B(p_1, p_2, \cdots, p_N; D/2) = \frac{\sum_{k=0}^{\infty}[k - D/2]^+ p_1 * p_2 * \cdots p_N(k)}{\sum_{k=0}^{\infty} k\, p_1 * p_2 * \cdots p_N(k)}$$

where $*$ is the convolution operation. Let $\theta_i(j)$be the following functions:

$$\theta_i(j) = \begin{cases} ANA_i \,/\, MNA_i, & j = MNA_i, \\ 1 - ANA_i \,/\, MNA_i, & j = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then it can be shown that

$$\begin{aligned} CLP \;&\leq\; B(p_1, p_2, \cdots, p_N; D/2) \\ &\leq\; B(\theta_1, \theta_2, \cdots, \theta_N; D/2) \\ &=\; \frac{\sum_{k=0}^{\infty} [k - D/2]^{+} \theta_1 * \theta_2 * \cdots \theta_N(k)}{\sum_{k=0}^{\infty} k\, \theta_1 * \theta_2 * \cdots \theta_N(k)} \end{aligned} \qquad (5.7)$$

A new connection is admitted if the resulting $B(p_1, p_2, \bullet\bullet\bullet, p_N; D/2)$ is less than the allowable cell loss probability. Saito proposes a scheme for calculating $\theta_1 * \theta_2 * \cdots \theta_N$ (k) efficiently. He also obtained a different upper bound based on the average and the variance of the number of cells that arrive during $D/2$.

The main disadvantage of this method is the absence of the burst size in the calculation and thus a worst case behaviour is assumed for the source. This method works well in the case when the actual source behaviour is close to the worst case behaviour assumed in the above calculation.

For other upper bounds on the cell loss probability see Rasmussen et al. [20], Castelli, Cavallero, and Tonietti [2], Doshi [6] and the closely related work by Elwalid, Mitra, and Wentworth [10].

2.4 COMPARATIVE PERFORMANCE ANALYSIS OF THE LOSS-ORIENTED CAC SCHEMES

In this section, we provide a numerical comparison among the following CAC schemes: a) the method proposed by Guerin, Ahmadi, and Naghshineh [13] for calculating the effective bandwidth (hereafter referred to as the "equivalent capacity method"), b) the diffusion approximation method proposed by Gelenbe, Mang and Onvural [12] for calculating the statistical bandwidth (hereafter referred to as the "diffusion approximation method), and c) Saito's upper bound of the cell loss probability [22] (hereafter referred to as the CLP upper bound). These schemes were selected since they use the same set/subset of traffic descriptors. Namely, the peak bit rate, mean bit rate, and mean burst length of a call $(R; \rho; b)$, (Note that the CLP upper bound scheme only utilizes the mean and peak bit rate information.) Before presenting the results, we define some necessary terms.

We will consider an ATM multiplexer consisting of a finite capacity queue of size K. This queue is served by a server (the outgoing link) of capacity C. The connections handled by this are classified into M classes, namely classes 1 through M. In this work, for illustration purposes, we limit M to 2. All the connections in the same class i have the same traffic descriptor (R_i, ρ_i, b_i), where R_i is the connection's peak rate, ρ_i is the connection's average bit rate, and b_i is the connection's mean burst length.

Admission Region: This is the set of all values of $(n_1 ; n_2)$ for which the cell loss probability is less than a small value ε, where n_i is the number of allocated class i connections, $i = 1; 2$. In other words, this is the set of all combinations of the connections from the 2 classes for which the required cell loss probability ε is achievable. In the numerical results given below, we obtain the outermost boundary of the region. All points enclosed between the boundary and the axes represent combinations of connections from each class which lie within the admission region.

Statistical Gain: Let $N_{\min_i}$ be the number of class i connections admitted using peak rate allocation. So, $N_{\min_i} = \lfloor C / R_i \rfloor$. Likewise, define $N_{\max_i}$ to be the number of class i connections that can be admitted using mean rate allocation. So, $N_{\max_i} = \lfloor C / \rho_i \rfloor$. The statistical gain for a particular traffic class is defined as the maximum number, N_i, of connections admitted by a CAC scheme divided by the number of connections that can be accepted using peak rate allocation ($N_{\min_i}$), i.e. $N_i / N_{\min_i}$ when a single class of calls is exclusively using the multiplexer. In order for a CAC scheme to be effective it should be able to provide some statistical gain when possible, i.e. achieve $N_i / N_{\min_i} > 1$

Each of the three CAC schemes was implemented separately. The performance of these schemes relative to each other was compared for various regions of input traffic parameters, buffer size, and required cell loss probability. Also, operating regions for which a particular scheme provides statistical gain over peak rate allocation were identified. We fixed the link speed at 150 Mbps and choose two classes of traffic with parameters given in Table 5.1.

Table 5.1. Traffic parameters for the two classes

	R (Mbps)	ρ (Mbps)	b (Cells)
Class 1	10	1	340
Class 2	2	0.1	2600

2.5 CASE 1: RELATIVELY SMALL BUFFER SIZE

We consider the admission control of two classes assuming a relatively small buffer. The system parameters were chosen as follows. We set the required cell loss probability ε is equal to 10^{-6} and the buffer size K equal to 618 cells (32 Kbytes). The minimum, $N_{\min_i}$, and maximum number, $N_{\max_i}$, of connections for class 1 and 2 are respectively: $(N_{\min_1}, \mathrm{N}_{\max_1}) = (15;\ 150)$ and$(N_{\min_2}, N_{\max_2}) = (75;\ 1500)$.

The admission regions obtained for the three CAC methods are shown in figure 5.2.

The diffusion approximation provides the largest admission region for this example. For this method, the statistical gain for classes 1 and 2 is respectively 7.3 and 14.16. For the equivalent capacity method the gain is 6.13 and 11.37 respectively. For the equivalent capacity method, we note that the admission region is approximately bounded by the intersection of two regions bounded by two almost-linear boundaries: one is obtained by the Gaussian approximation and the other by the effective bandwidth calculation (the intersection near the (25,410) point). The CLP upper bound scheme provides a conservative admission regions yielding a statistical gain for classes 1 and 2 of 2.86 and 11.55 respectively. We note that for the case when the majority of connections belong to class 2, the CLP method is superior to equivalent capacity. However, this scheme is in general conservative with respect to the other schemes. It is obvious that for class 2 which has a much smaller mean to peak ratio the achieved gain for any of the methods is much higher than class 1 although it has a much longer on period. In general the larger the ratio of link capacity to the mean rate of connections, the larger the achieved statistical gain.

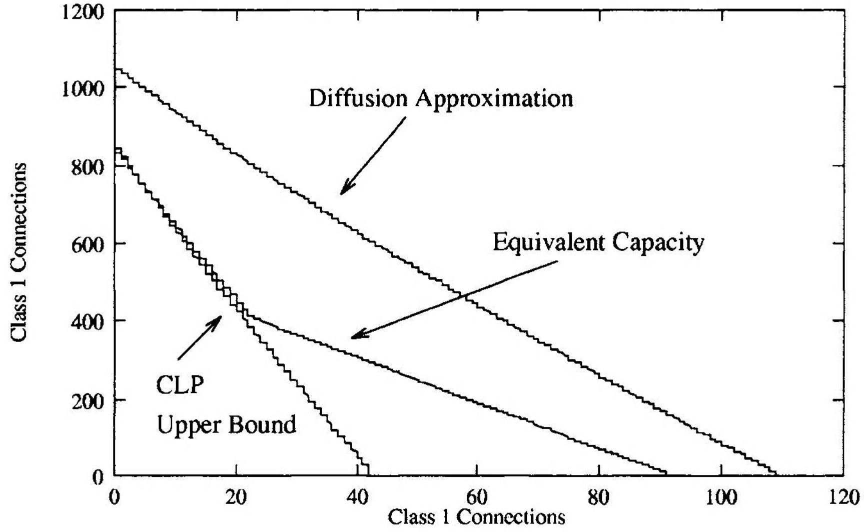


Figure 5.2. Admission regions for the CAC schemes, K=618 cells, ε=10^{-6}

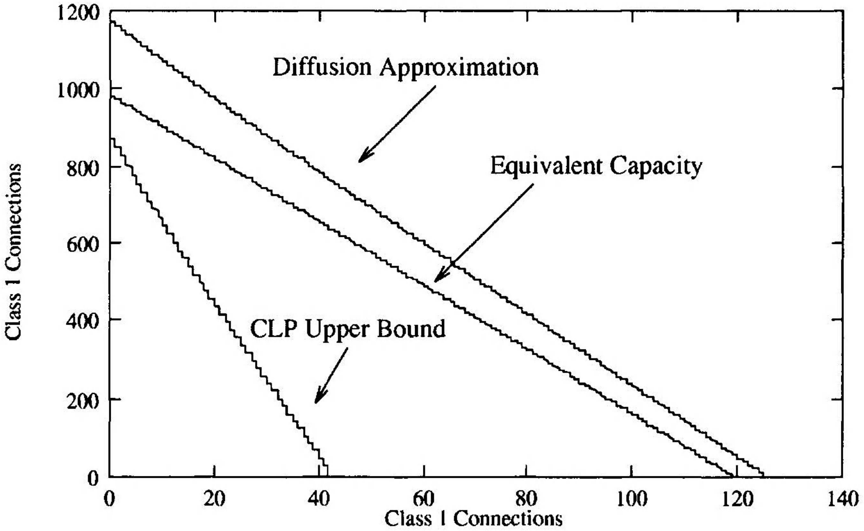


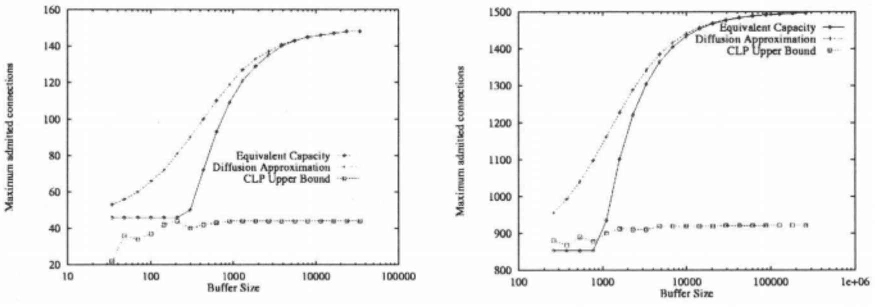
Figure 5.3. Admission regions for the CAC schemes, K=1236, ε= 10^{-6}

2.6 CASE 2: RELATIVELY LARGE BUFFER SIZE

The buffer size K was doubled to 1236 cells (64 Kbytes). The obtained admission regions for the three schemes are shown in figure 5.3.

Since the buffer size is increased to 1236, the admission region of all schemes increases. The diffusion approximation provides the largest admission region. When a single class share the multiplexer, the statistical gain that the diffusion approximation yields for classes 1 and 2 are respectively 8.4 and 15.75. For the equivalent capacity method the gain is 8 and 13.19 respectively. In this case, only the effective bandwidth calculation affects the admission region of the equivalent capacity.

For the CLP upper bound scheme, we observe that the maximum number of admitted connections from each class does not increase appropriately when doubling the buffer size. The achieved gains are 2.86 and 11.95 for class 1 and 2 respectively. The maximum number for class 1 remains at 43 while the maximum for class 2 increases slightly from 866 to 896. The reason for this is that class 2 has a lower peak rate and average rate than class 1. We note that in order for this scheme to yield a statistical gain, we need to have traffic sources with small peak and average rate relative to the link capacity.



(a) (b)

Figure 5.4. Varying the Buffer Size, $\varepsilon = 10^{-6}$

2.7 EFFECT OF THE BUFFER SIZE

Assuming that only class 1 or class 2 connections are transported, we obtain the maximum number of admitted connections as a function of the buffer size. The buffer size is increased from a value of $b_i/10$ to 100 $b_{i,}$ where b_i is the mean burst length of class i, while the required cell loss probability is ε fixed at 10^{-6}. The results are shown in figure 5.4. The figure indicates that the diffusion approximation scheme and the equivalent capacity scheme asymptotically admit the same number of connections as the buffer size approaches infinity.

We observe that for small buffer sizes, the equivalent capacity method admits a fixed number of connections obtained through the Gaussian approximation (bufferless approximation). Furthermore, for class 1, the number of connections admitted by equivalent capacity is smaller than those admitted by the CLP upper bound for small buffer sizes.

The CLP upper bound scheme is less sensitive to the increase in buffer size. For this scheme, a temporary drop occurs to the maximum number of connections that can be admitted as the buffer increases. This is due to the effect of dividing *ANA* by *MNA* where *MNA,* a function of the buffer size and peak rate, must be an integer. So, by increasing the buffer size we get different values of *ANA/MNA.* We note also that increasing the buffer size beyond a specific value does not cause any increase in the number of admitted connections.

2.8 EFFECT OF THE REQUIRED CELL LOSS PROBABILITY

Assuming that only class 1 or class 2 connections are transported, we obtain the maximum number of admitted connections as a function of the required cell loss probability. We fix the buffer size at 1236 cells and increase the cell loss probability from 10^{-9} to 10^{-3}. The results are shown in figure 5.5.

From this figure, we observe that for class 1 the diffusion approximation and the equivalent capacity scheme exhibit low sensitivity to the cell loss probability. In this particular example, the buffer size is large enough so that the two schemes admit a large number of connections even for a very small value of the required cell loss probability. For the diffusion approximation, the increase in the cell loss probability caused the maximum number of connections for class 1 to only increase from 118 to 138, not even reaching the maximum number of admittable connections, 150. The equivalent capacity scheme is more sensitive to the required cell loss probability than the diffusion approximation scheme. The maximum number of connections that can be admitted increased from 105 to 136 exhibiting higher sensitivity. This sensitivity is of course a function of buffer size as well. In general both methods become more sensitive when buffer sizes are small.

The CLP upper bound method is the most sensitive to the cell loss probability. In this example, the increase in the maximum number of connections is from 25 to 100 for class 1 and from 740 to 1320 for class 2. Since the sensitivity of the CLP upper bound method to buffer size is small, it seems, that the required cell loss probability affects the admission region and the achievable statistical gain.

Therefore, for the diffusion approximation and the equivalent capacity methods, if the buffer size is large their sensitivity to CLP is small whereas the CLP upper bound scheme is usually quite sensitive to the cell loss probability.

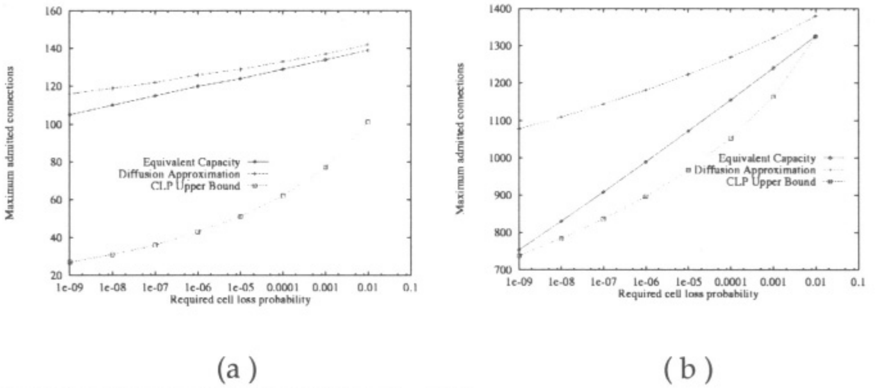


Figure 5.5. Varying the required CLP ε, K = 1236

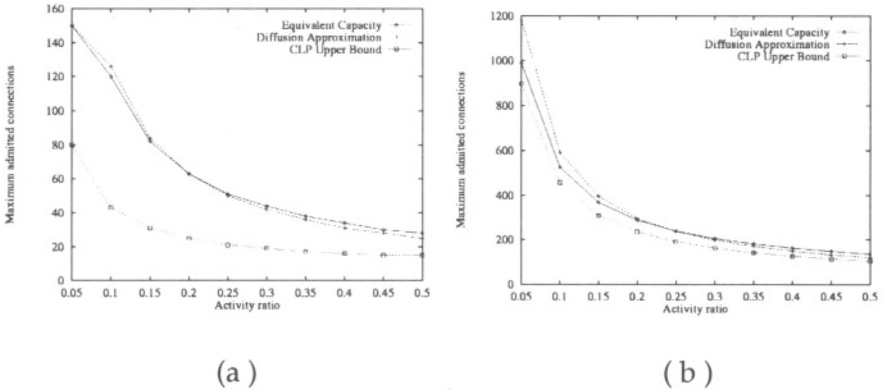


Figure 5.6. Varying the Activity Ratio r, $\varepsilon = 10^{-6}$

2.9 EFFECT OF THE ACTIVITY RATIO

In this section, we study the sensitivity of the three CAC schemes to changes in the activity ratio $r_i = \rho_i / R_i$. Assuming that only class 1 or class 2 connections are transported, we obtain the maximum number of admitted connections as a function of r_i , as r_i increases from 0.05 to 0.5. We fix the buffer size at 1236 cells and the required cell loss probability at 10^{-6}. The results are shown in figure 5.6.

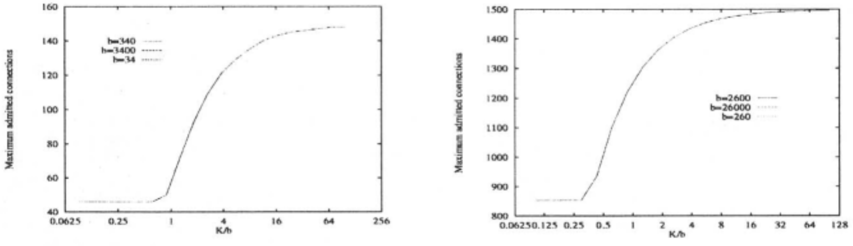
We observe a strong dependence of all methods on the activity ratio. For class 1, when the activity ratio is 0.05, the two methods provide the maximum possible admitted number of connections (i.e. 150). The admitted number of connections drops sharply to 28, 25, and 15 respectively for the equivalent capacity, diffusion approximation, and CLP upper bound methods. The same behaviour is also observed in the case of class 2. The sensitivity to the activity ratio is greatest for the diffusion approximation and it is larger for the class with the smaller peak rates. We note that the equivalent capacity methods admits more connections than the diffusion method when the activity ratio exceeds 0.25 for both class 1 and class 2.

2.10 EFFECT OF THE RATIO OF THE BUFFER SIZE TO THE MEAN BURST LENGTH

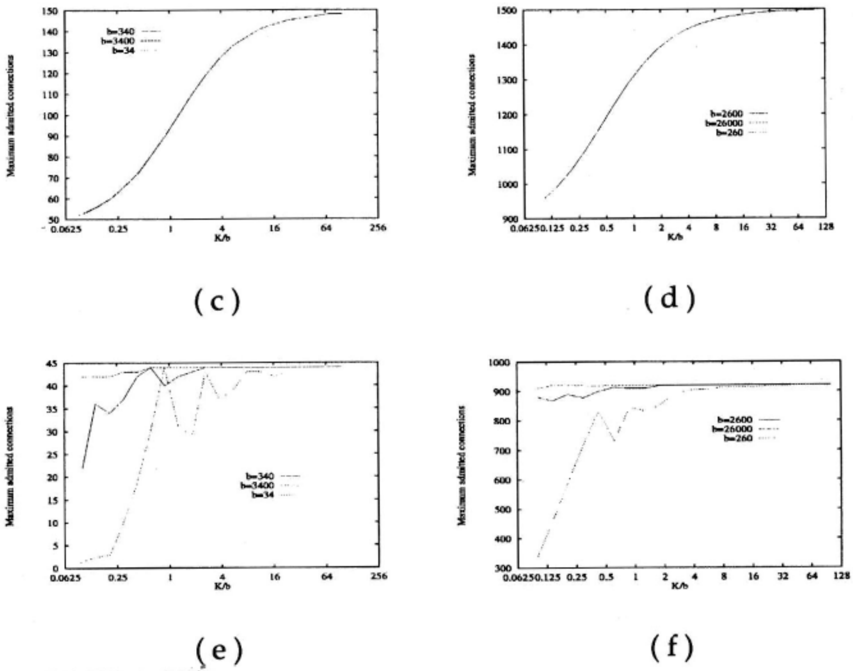
We have already observed that the diffusion approximation scheme and the equivalent capacity scheme behave similarly when the buffer size is large. In this section, we study the effect of the ratio of the buffer size to the mean burst length of a connection, while keeping all other parameters fixed. We consider a multiplexer with either class 1 or 2 connections. The peak and average rates are given in Table 5.1 while the mean burst size b was varied. For each value of *b*, the buffer size *K* was varied so that the ratio *K/b* varied from 0.1 to 100.

The results for the equivalent capacity, the diffusion approximation, and the CLP upper bound schemes are shown in figure 5.7. We note that for the equivalent capacity and the diffusion approximation methods, as long as the ratio *K/b* is kept constant, the maximum number of admitted connections is almost the same regardless of the value of the mean burst length *b*. This observation can be used in order to approximate the solution of a multiplexer with a large buffer size by that of a multiplexer with a smaller buffer. The mean burst length of the source must be scaled down accordingly in order to keep the ratio *K/b* constant.

The CLP upper bound scheme does not behave similarly, since it does not use any information about the burst length of the connection. This is reflected in figures 5.7(e) and 5.7(f). In this case, for each given value of *b* and *K/b* we get a new value of K. Since b is not taken into account in the calculation, the number of connections does not scale as in the other two schemes. As has already been observed this scheme's sensitivity to buffer size is poor.



(a) (b)



(c)

(d)

(e)

(f)

Figure 5.7. Effect of K/b

3. CAC SCHEMES BASED ON THE CELL DELAY

For real-time applications, the network must be able to provide timely delivery of packets. For many applications, packets must be delivered within a bounded delay and/or bounded delay jitter. In this case, the CAC process has to ensure that the network will meet the required end-to-end delay and/or delay jitter for a new connection. Also, the CAC must insure that admitting the new connection would not affect those connections already in progress.

For delay bounded connections, we have two major categories of QOS: deterministic and statistical. For deterministic QOS, a connection requests that all its packets reach their destination within some finite delay D. Such a connection will be called a guaranteed service connection. For statistical QOS, a connection requests, for example, that the probability that the delay of a packet is smaller than a given bound *D* must be greater than a given value Δ. Such a connection will be called a predictive service connection. In this paper, we concentrate on CAC schemes for guaranteed service connections.

CAC schemes for the cell delay are closely associated with the packet scheduling mechanism implemented in the network switches. The scheduling mechanism determines to a large extent the packet queueing delay at each switch. A lot of work has been done in the area of calculating packet delays for various scheduling disciplines such as First-In-First-Out [10, 11], Static Priority [10, 11], Weighted Fair Queueing [19], and Earliest Deadline First [18]. When comparing scheduling disciplines it is necessary to evaluate the following aspects:

- Admission/schedulability region: how many connections from each class can be admitted without violating their requested delay bounds?
- Isolation and fairness among connections
- Ease of implementation and complexity of the calculation needed to perform the admissibility/schedulability test

In our model, we assume that connections are constrained by a leaky-bucket like traffic filter and each connection *i* has the traffic descriptor (R_i, ρ_i, b_i) where R_i is the peak rate, ρ_i the average rate, and b_i is the maximum burst size. With this traffic model, it is possible to calculate the worst-case end-to-end delay for many scheduling disciplines.

3.1 PACKET-BY-PACKET GENERALIZED PROCESSOR SHARING

Weighted fair queueing (WFQ) and packet-by-packet generalized processor sharing (PGPS) are approximations of the Generalized Processor Sharing (GPS) discipline. WFQ and PGPS are identical so we will only consider WFQ. In GPS, packets are served as if they are in separate logical queues, the server visits each nonempty queue in turn and serves an infinitesimally small amount of data from each queue, so that, in any finite time interval, it can visit each logical queue at least once, independent of the number of queues. The scheduler in WFQ works as follows: compute the time that a packet would finish its service if the packet is served by a GPS server; then serve packets in order of their finishing times. The calculation of the packet finishing times under (weighted) GPS is illustrated in Keshav [14].

To determine the worst-case end-to-end packet delay, consider a connection constrained by (b_i, ρ_i) passing through *L* schedulers, where the *l*th scheduler has a link rate C_1 . Let $g_{i,l}$ be the service rate assigned to that connection at the *l*th scheduler. Let $g_i = \min_l g_{i,l}$, where $g_i \geq \rho_i$ for stability of the queues. Let $P_{\max_i}$ be the largest packet from connection *i*, and assume that $P_{\max}$ is the largest size of packet allowed in the network. Then, the end-to-end network delay d_i for a packet from connection *i* satisfies [14, 19]:

$$d_i \le \frac{b_i}{g_i} + \sum_{l=1}^{L-1} \frac{P_{\max_i}}{g_{i,l}} + \sum_{l=1}^{L} \frac{P_{\max}}{C_l} \tag{5.8}$$

independently of the behaviour of other connections.

It is very important to note that, when the link speed is very large compared to $P_{\max}$, the above bound of d_{i} simplifies to $\frac{b_i}{g_i}$, i.e. packetization is very important for providing small end-to-end delay.

A CAC scheme based on WFQ scheduling works as follows. When a connection is setup, the connection parameters $(b_i, \rho_{i,}, d_{\mathrm{i}})$ are signaled to the network. The network calculates the required g_i to satisfy the delay constraint using equation (5.8). If $g_i \ge \rho_i$ and the sum of g_i plus the reserved bandwidth of the existing connections is smaller than C_l and the sum ρ_i of plus the overall average rate of the connections is smaller than C_l at all intermediate switches, the connection is admitted; otherwise it is rejected.

3.2 DELAY EARLIEST-DEADLINE-FIRST SCHEDULING

In earliest-deadline-first (EDF) schedulers, each packet is assigned a deadline and the scheduler serves packets in order of their deadline. Delay-EDF is an extension of EDF that describes how a scheduler assigns deadlines to packets. At connection setup time, the connection declares a peak rate and a desired delay bound for worst-case delay. The scheduler performs a schedulability test to ensure that every connection meets its delay bound even when they are transmitting at peak rate.

A delay-EDF scheduler needs to sort packets in order of their deadline, which is also done by WFQ. The scheduler also needs to store finishing times as in WFQ. The main advantage of delay-EDF over WFQ is that its delay bound is independent of the allocated bandwidth to the connection at the expense of peak bandwidth allocation (this, however, can be relaxed for connections constrained by a leaky-bucket). EDF has been proven to be an optimal scheduling discipline in the sense that if a set of connections is schedulable under any scheduling discipline then the set is also EDF-schedulable in the single node case.

Consider leaky-bucket constrained connections with traffic descriptor $(b_i, \rho_{i,})$ and a delay bound d_{i} at scheduler l. Assume that two connections i and j are ordered such that $d_i^l < d_j^l$ if $i < j$. Then as long as $\sum_{i=1}^{N} \rho_i < C_l$, we have the following schedulability condition at scheduler *l* (due to Libeherr, Werge, and Ferrari [18])

$$d_j^l \geq \frac{b_j + \sum_{i=1}^{j-1}(b_i - \rho_i d_i^l) + \max_{k>j} P_{\max_k}}{C_l - \sum_{i=1}^{j-1}\rho_i} \qquad (5.9)$$

The schedulability test of delay-EDF schedulers is complex since the check for condition (5.9) is computationally expensive.

Liebherr and Werge [17] simplified the implementation of EDF scheduling by discretizing the range of packet deadline values. The search time for the next packet to schedule is brought to O(1). Firiou, Kurose, and Towsley [11], suggested an efficient algorithm for schedulability testing given that connections are constrained by (R, ρ, b). The complexity is *O(N),* where *N* is the number of admitted connection at the time of invocation of the schedulability test.

A possible CAC scheme based on EDF scheduling is the following:

1. A set-up message for connection *i* is sent along the connection's selected path. The set-up message contains connection's i traffic descriptor (R_i, ρ_i, b_i) and its end-to-end delay bound d_i. A variable d_i is initialized to zero and included in the set-up message.

2. At each intermediate scheduler *l*, a minimum value for the maximum delay d_i^l that can be assured for connection *i* is calculated. The variable d_i is incremented by d_i^l. At the same time, CAC checks if $\sum \rho < C_l$ for all connections passing through the link including the new connection.

3. At the destination node, CAC checks if $d_i \leq d_i$. If yes, the connection is accepted.

4. On the reverse path, a local delay bound d_i^l is calculated. This is the local deadline of connection *i* at link *l*.

3.3 STATIC PRIORITY SCHEDULING

A static-priority (SP) scheduler assigns each connection to a fixed priority level *p*, where $1 \leq p \leq P$, where *P* is the number of priority levels. All connections in priority level *p* will have the same delay bound d_p, with $d_p < d_q$ for $p < q$, i.e. the priority of a connection is high if its delay bound is low. The SP scheduler always selects the first arriving packet packet from the highest priority backlogged queue. It is fairly easy to implement an SP scheduler since it consists of a fixed number of FIFO queues, one for each priority level. For leaky-bucket constrained connections, Cruz [5] has derived necessary and sufficient schedulability conditions for SP schedulers to satisfy a given delay bound.

Consider connections that are (ρ, b) constrained, where ρ is the average rate and b is the maximum burst size. Let $P_{\max_p}$ be the largest packet size for connections belonging to priority level p. Assuming only one connection in each priority level and that the minimum packet size is zero. Let P be the number of sessions, (ρ_p, b_p, d_p) be the traffic descriptor and delay bound for connection p, where $1 \leq p \leq P$, then the set of connections is schedulable at link l if

$$d_p \geq \frac{\sum_{q=1}^{p} b_q + \max_{r>p} P_{\max_r}}{C_l - \sum_{q=1}^{p-1} \rho_q} \tag{5.10}$$

for $1 \leq p \leq P$. A CAC scheme for SP schedulers can be devised in a similar manner to the EDF-based CAC as shown in section 3.2.

3.4 COMPARATIVE PERFORMANCE ANALYSIS OF THE CAC SCHEMES FOR CONNECTIONS WITH A MAXIMUM DELAY BOUND

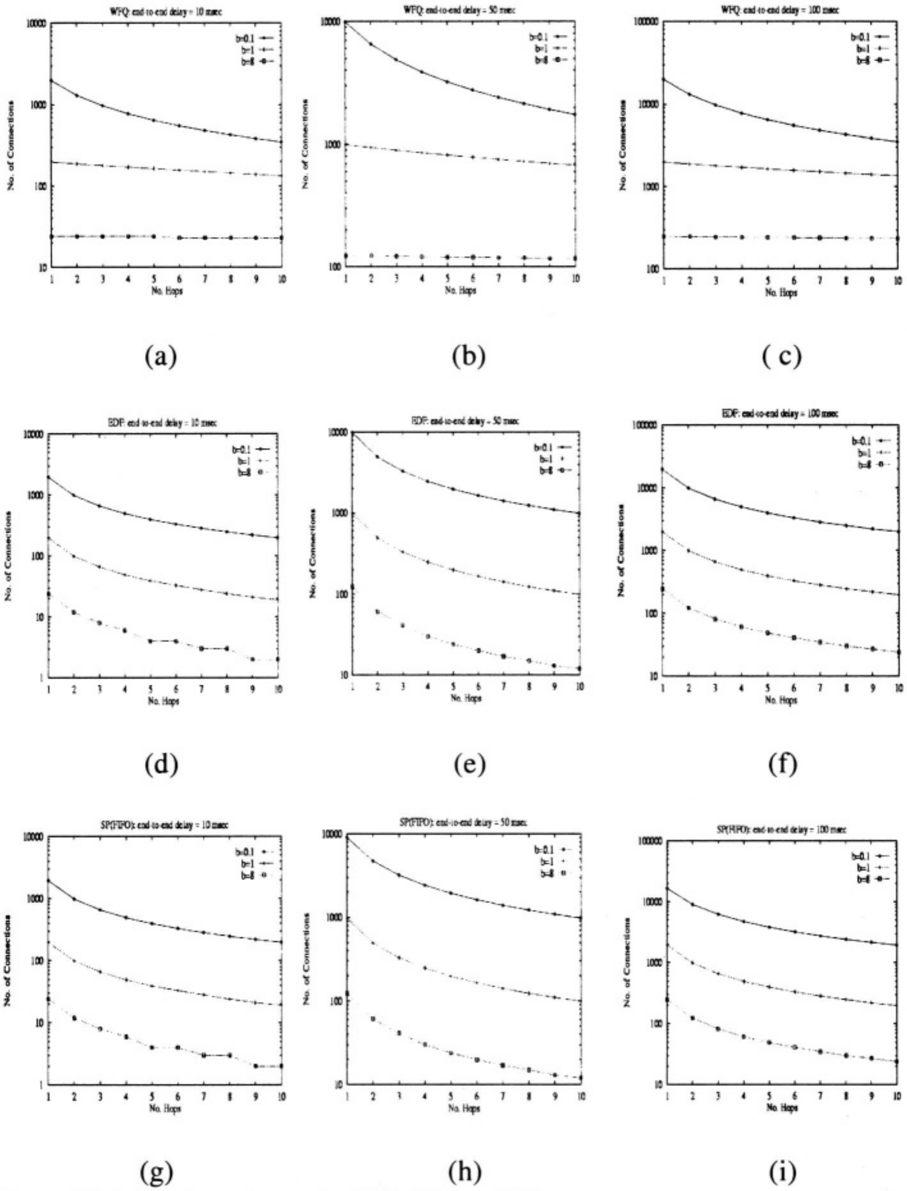
In this section, we carry out a comparative study of the WFQ, EDF, and SP scheduling disciplines for connections with guaranteed service for delay. We consider a source/destination pair interconnected by a path of L hops. The nodes are homogeneous and have a fixed link capacity C. We also assume that all connections are identical, and that they all traverse the same set of nodes and links from source to destination. We evaluate the maximum number of connections that can be admitted in the network without violating the delay bound, for various traffic characteristics and different values of the delay bound. Certainly, the above does not describe a realistic network configuration, since in a real network there will be more than one source destination pair, connections would traverse different paths, and connections would have varying traffic descriptors and delay bounds. This fictitious configuration, however, is useful in illustrating the basic behavior of the scheduling disciplines and their performance relative to each other.

In all experiments we set the link capacity C to 155 Mbps and all packets are constrained to be of fixed length of 53 bytes. We assume a connection of characterised by the descriptor (ρ, b), where ρ is the average rate and b is maximum burst size. We fix ρ at 3550 bps such that the condition $N\rho < C$, where N is the maximum number of connections admitted by a specific scheduling discipline, is not violated in all the cases considered. This is deliberately chosen such that the number of connections is constrained by the deadline schedulability conditions of a particular scheduling discipline and not the stability condition of the system. The number of hops traversed by the connections is varied from 1 to 10, the burst size b takes values from the set {0.1, 1, 8}Kbytes, and the delay bound d takes values from set {10, 50, 100} msec.

Figure 5.8 shows the maximum number of connections as a function of the number of hops, burst size, and delay bound for WFQ, EDF, and SP scheduling disciplines. Since we have only one class of connections, SP becomes equivalent to the FIFO discipline.

For the three disciplines, there is a clear strong dependence of the number of admitted connections on the burst size and the number of hops traversed. As the burst size increases, the number of admitted connections decreases. Likewise, as the number of hops increases, we observe a decrease on the number of admitted connections. The burst size has the strongest influence on WFQ. Also, in the case of WFQ, the increased number of hops does not affect the number of connections when burst size is very large compared to the maximum packet size (in the reported results, b=8 kbytes, and $P_{\max}$ = 53 bytes). In general for small burst sizes (i.e. b=0.1 Kbytes), the number of admitted connections decreases rapidly as the number of hops increases. The influence of the number of hops on the number of admitted connections decreases as b increases (i.e. when the sources become burstier).

In figure 5.9, we compare the performance of the three scheduling disciplines, assuming an end-to-end delay of d = 50 msec and b = 0.1 and 8 Kbytes. For the examples considered, both WFQ and EDF always provide the same number of connections for L = 1. It is obvious that WFQ performs consistently better than EDF in a multi-hop network. Does this violate the fact that EDF is optimal? No, the reason behind this is that in WFQ, the reserved bandwidth is calculated using a methodology that takes into account the network as a whole. In EDF, however, local delays are added up in each node without taking into account how the connection's traffic is distributed among the multi-hop path. This shows that there is a need to modify the schedulability conditions of EDF (and SP) schedulers to take the distribution of traffic among the network nodes. It is also clear that by increasing the number of hops, EDF and SP become identical. Also, for large burst size, there is no difference between EDF and SP.



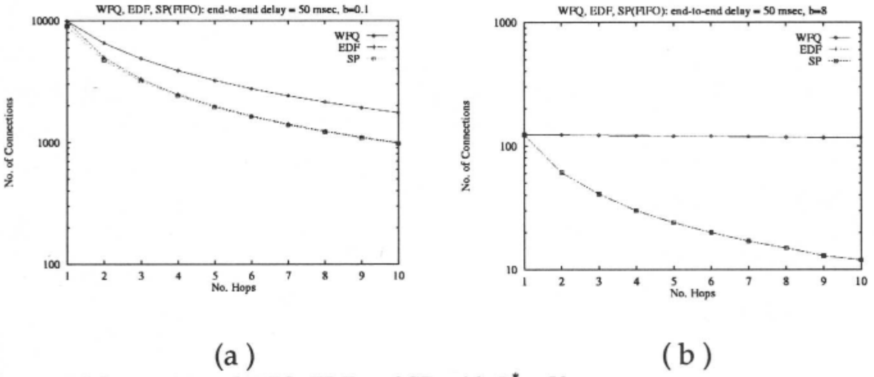
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.8. Admission region under WFQ, EDF, and SP

Figure 5.10 shows the dependence on burst size for the case where $L = 1$ (note that WFQ and EDF provide identical results for L=l). The burst size is varied from 50 bytes (about one ATM cell) to 50 Kbytes (about 1000 ATM cells). For large bursts, the scheduling disciplines provide identical results and the number of admitted connections decreases dramatically. Since most real-time applications have some knowledge about the required end-to-end delay but they can not tell in advance their burst size, we suggest that the network should be able to negotiate a burst size with the application requesting guaranteed service. For predictive service, we suggest adjusting the burst size dynamically at run-time to allow for better network performance.



(a) (b)

Figure 5.9 Comparison of WFQ, EDF, and SP, with D^* = 50 mses

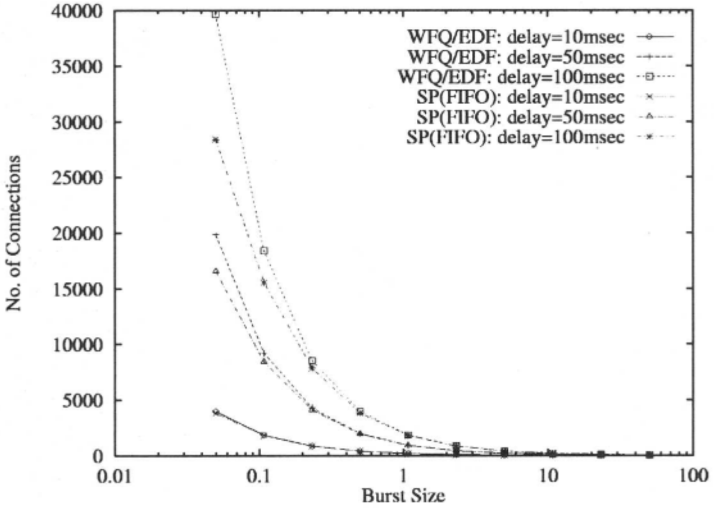


Figure 5.10. Effect of burst size on admission region

4. CONCLUSIONS

In this paper, we have provided an evaluation of CAC schemes for cell loss and delay sensitive services in an ATM network. For cell loss sensitive services, we evaluated the equivalent capacity, the diffusion approximation and the CLP upper bound methods. Below, we summarize the findings reported in this paper:

Performance: In most of the scenarios considered in this paper, the diffusion approximation has outperformed the other methods in terms of providing a larger admission region than the other two schemes. The CLP upper bound scheme is usually a pessimistic scheme but still provides some statistical gain over peak rate allocation. The CLP upper bound scheme as the activity ratio decreases and approaches the same performance levels attained by the other two schemes.

Complexity of CAC decision making: The CLP upper bound method is the most complex to implement because of the expensive and computationally intensive convolution operation (note that a simplification for the evaluation exists). The diffusion approximation method is computationally more expensive than the equivalent capacity method since C_{FB} and/or C_{IB} needs to be computed each time a new connection is negotiated with the system. The calculation of C_{FB} and C_{IB} involves the computation of a square root. If the Gaussian approximation (see equation (5.2)) is used to complement the equivalent capacity basic calculation, then the equivalent capacity and diffusion approximation schemes are approximately equivalent from the complexity point of view.

Suitability for implementation in a real-life network: The CLP upper bound method is more suitable for implementation in a real-life scenario since it requires only two traffic parameters which are peak rate and average rate. In scenarios where traffic is shaped, it would be possible to get estimates for the burst size and the other schemes (equivalent capacity and diffusion approximation) will provide better performance than CLP.

Sensitivity: All schemes have exhibited some dependency on the variations of the system parameters. The equivalent capacity and diffusion approximation methods show significant dependency on buffer size and the buffer size to mean burst size ratio. The CLP upper bound scheme does take the mean burst size into account and is therefore not sensitive to uncertainties in this parameter. All schemes have exhibited large sensitivity to the activity ratio. Therefore any change in the average/peak rate values will affect the accuracy of the decision made by the CAC scheme applied. This calls for the need to do apply dynamic bandwidth allocation and calculation when applying these schemes to assure the accuracy of the CAC decision in real-life networks.

In the second part of the paper, we considered delay sensitive services. Specifically, we discussed CAC schemes associated with three scheduling disciplines: WFQ, EDF, and SP scheduling. We believe WFQ to be the best available candidate for deployment in real-life networks due to its superiority when it comes to the issues of complexity, implementation and performance. However, WFQ is the most sensitive scheme to burst size uncertainties. EDF can be competitive to WFQ in the case of networks with small number of hops.

References

[1] ATM Forum User Network Interface (UNI) 4.0 Specification, 1996.

[2] P. Castelli, E. Cavallero, and A. Tonietti, Policing And Call Admission Problems in ATM Networks, in: A. Jensen and V.B. Iversen (Eds.), Teletraffic and datatraffic in a period of change, (North-Holland, 847-852, 1991.

[3] C.-S. Chang and J. A. Thomas. Effective Bandwidth in High Speed Networks. IEEE Journal on Selected Areas in Communications, 13:1091--1100, 1995.

[4] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, On the Effective Bandwidths for Admission Control in ATM Networks. In Proceedings of 14th International Teletraffic Congress (ITC), pages 411--420, 1994.

[5] R. Cruz, A calculus for network delay: part I: Network elements in isolation, IEEE Trans. Inform. Theory, vol. 37, 114--131, 1991.

[6] B. T. Doshi. Deterministic Rule Based Traffic Descriptors for Broadband ISDN: Worst Case Behavior and Connection Acceptance Control. In Proceedings of 14th International Teletraffic Congress (ITC), 591-600, 1994.

[7] N. G. Duffield, J. T. Lewis, N. O'Connel, R. Russell, and F. Toomey. Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters. IEEE Journal on Selected Areas in Communications, 13:981-990, 1995.

[8] K. Elsayed and H. G. Perros. Analysis of an ATM Statistical Multiplexer with Heterogeneous Markovian On/Off Sources and Applications to Call Admission Control. Journal of High Speed Networks, vol. 6 no. 2, pp. 123-139, 1997.

[9] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing. IEEE Journal on Selected Areas in Communications, 13:1004-1016, 1995.

[10] A. Elwalid, D. Mitra, and R. H. Wentworth. A new Approach for Allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node. IEEE Journal on Selected Areas in Communications, 13:1115-1127,1995

[11] V. Firoiu, J. Kurose, and D. Towsley, Efficient admission control for EDF Schedulers, IEEE Infocom'97, 1997.

[12] E. Gelenbe, X. Mang, and R. Onvural, Diffusion based statistical call admission control in ATM, Performance Evaluation, vol. 27, 411-36, 1996.

[13] R. Gu’erin, H. Ahmadi, M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE Journal on Selected Areas in Communications, vol. 9, 968-981, 1991.

[14] S. Keshav, An Engineering Approach to Computer Networks, Chapter 9, Addison-Wesley, 1996.

[15] V. Kulkarani, L. G¨un, and P. Chimento, Effective bandwidth vector for two-priority ATM traffic, INFOCOM '94 , 1056-1064, 1994.

[16] S.-Q. Li. A General Solution Technique for Discrete Queueing Analysis of Multimedia Traffic on ATM. IEEE Transactions on Communications, 39:1115-1132, 1991.

[17] J. Liebherr and D. E. Werge, Design and analysis of a high-performance packet multiplexer for multiservice networks, with delay guarantees, Technical report CS-94-03, University of Virginia, 1994.

[18] J. Liebherr, D. E. Werge, and D. Ferrari, Exact admission control for networks with a bounded delay service, IEEE Trans. on Networking, vol. 4, 885-901, 1996.

[19] A. K. Parekh and R. G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The multiple node case, IEEE Trans. on Networking, vol. 2, 137-150, 1994.

[20] C. Rasmussen, J.H. Sørensen, K.S. Kvols, and S.B. Jacobsen, Source-independent call acceptance procedures in ATM networks, IEEE JSAC 9,351-358,1991.

[21] K. M. Rege. Equivalent Bandwidth and Related Admission Criteria for ATM Systems-A Performance Study. International Journal of Communications Systems, 7:181--197, 1994.

[22] H. Saito, Call admission control in an ATM network using upper bound of cell loss probability, IEEE Trans. Comm. vol. 40, 1512-1521, 1992.

[23] G. De Veciana, G. Kesidis, and J. Walrand. Resource Management in Wide-Area ATM Networks Using Effective Bandwidth. IEEE Journal on Selected Areas in Communications, 13:1081--1090, 1995.

Chapter 6

TRAFFIC CONTROL IN ATM: A REVIEW, AN ENGINEER'S CRITICAL VIEW & A NOVEL APPROACH

Nikolas M. Mitrou
National Technical University of Athens,
Heroon Polytechneiou 9,
157-73, Zografou, GREECE
mitrou@softlab.ntua.gr

Abstract This paper has a twofold objective. First, it aims at reviewing the basic ATM traffic control functions, as defined by the standards, addressing popular methods that are used for performance analysis of ATM multiplexing and, then, bending critically over controversial issues in this area; among these issues is the suitability of statistical models for traffic control, the effectiveness of rate-based control schemes as well as the advantages and disadvantages of using effective rates as traffic descriptors. Motivated by the results of this critical analysis, the second part of the paper is devoted to a new approach to the ATM traffic control problem. Taking into account the foreseen convergence between IP and ATM, it focuses on a burst-level modelling through the classical *M/G/1* model, which essentially ignores cell-level details within bursts, including the cell rate itself, and exploits the buffering gain, assuming a large buffer_space/burst_size ratio.

Keywords: ATM, Traffic Control, Statistical Gain, Effective Rate, M/G/1

1. INTRODUCTION

The Asynchronous Transfer Mode (ATM) has been chosen for use in the Broadband Integrated Services Digital Networks (B-ISDN) as a set of layer-2 functions in the respective protocol stack providing the upper layers with a universal transport service, independent of application-specific characteristics and QoS requirements. Combined with a high-speed transmission at the physical layer, ATM is expected to remove the burdens of the many specialized infrastructures and offer a real integrated environment for information exchange.

Among the unique features that are upgrading the ATM from a simple switching technique, as initially conceived, to one of the key technologies for the information age, are: (i) the small, fixed size of the ATM packet (*cell*), resulting in a small packetization delay as required by real-time applications; (ii) the relatively small set of functions in the ATM layer, encoded into the 5-byte cell header, which allows for a fast processing at switching nodes (hardware-implemented fast packet switching); (iii) the possibility of allocating resources to connections according to their instantaneous demand (practically, with no bandwidth granularity), which can lead to a high resource utilization through statistical multiplexing.

On the debit side, complexity in the control & management planes is the critical feature of ATM, since a rich set of respective functions must be developed to cope with the different traffic profiles (from Constant Bit Rate - CBR - to multirate and Variable Bit Rate - VBR) and the divergent Quality-of-Service (QoS) requirements imposed by the different applications (voice, data, video, multimedia).

From its conception mid eighties [37], a great deal of effort has been devoted to ATM world-wide, covering all range of activities: building components and systems, writing operation and conformance testing software, analyzing the performance of key elements or network-scale configurations, developing traffic engineering methods and tools, and so on. Work in Standardization bodies and Specification fora (ITU-T, ETSI, ATM Forum etc.) has proceeded rapidly, and a lot of research projects have completed their work in the field. A large portion of this activity has been devoted to the traffic analysis and control, since the latter has been identified as one of the key issues for the success of ATM [2, 4, 13, 24,...]. Despite, however, the great advances that have been achieved, some important issues of this field are still open and are not expected to be fixed soon. The main reason for this is that the ATM B-ISDN is a network open to new services, the features and quality requirements of which can hardly be anticipated.

The primary role of *traffic control* is to guarantee network performance and to offer the contracted QoS to the customers. A second objective is the improvement of resource utilization in order to reduce the communication cost. In this respect, the evolution towards a full development of the ATM may pass through a first stage of a circuit-switching-like operation, where connections will be allotted resources according to their peak demand; then, by developing suitable traffic control mechanisms, the gain from statistical sharing of network resources (links, buffers) among many bursty information sources will be exploited.

This paper aims at addressing some key issues of ATM traffic control with references to related standards and performance analysis work. It is not intended to be a bibliography review, nor a tutorial in the usual sense, since a lot of material remains uncrystalized on the air and many subjects are still open. Instead of presenting yet another review, the paper aims at giving the basic notions of the subject, with emphasis on the respective traffic modelling and performance analysis from an engineer's viewpoint and also presenting the author's opinion about open or insufficiently covered topics (section 2). A new burst-centric modelling and control paradigm is presented in section 3, exploiting mainly the buffering gain and yielding delay-oriented QoS figures. In the same section the main traffic control mechanisms (shaping, aggregation, splitting, conformance testing) are established within the new framework, where the effective rate of *M/G*-type streams is used as the basic traffic parameter under control.

2. FUNCTIONS, STANDARDS AND ANALYSIS METHODS FOR ATM TRAFFIC CONTROL

2.1 BASIC TRAFFIC & CONGESTION CONTROL (T&CC) FUNCTIONS

In this sub-section, definitions and preliminary explanations on some basic notions encountered throughout this paper are given, in order to provide the reader with a concise reference material and a guide on the paper's main issues. Although these definitions refer to connection-oriented networks, as is the case of ATM, most of them can apply to connectionless environments as well. For a formal description of the generic Traffic and Congestion Control functions in ATM networks (defined so far) the reader is referred to the ITU-T and ATM Forum respective documents [2,13].

Congestion addresses the state of Network Resources in which the network is not able to meet the negotiated Performance Objectives for the already established connections. Possible reasons for congestion include unpredictable statistical fluctuations of traffic streams, which may lead to buffer overflow and excessive cell losses or unacceptable delays, or possible faults in network components, e.g. links, which may congest the network.

Traffic & Congestion Control (T&CC) is the set of actions taken by the network and/or the user to avoid congestion (preventive traffic control) or to minimize the effects of congestion once occurred (congestion control). Two types of control actions are distinguished:

Open-loop control refers to those control functions which are applied without any feedback from the network. Closed-loop control, on the other hand, refers to those control functions that utilize some knowledge about the network resource occupancy, or its performance. Note that, by definition, congestion control is always of the closed-loop type. In [40] the reader can find a more general taxonomy of control algorithms in packet switching networks (there referred to collectively as Congestion Control algorithms).

Network Resource is any component of the network, hardware or software, that is used to support the communication needs. Two generic resources are spread all over the network: The *time resources* (link rate or bandwidth) and the *space resources* (buffer space).

Statistical multiplexing (SM) of traffic addresses the sharing of a network resource among many statistically varying traffic streams, the peak demand of which may exceed the available resource capacity.

Statistical Gain (referred mainly to bandwidth) is defined by the ratio of the actual resource utilization by a group of statistically varying traffic streams and the respective figure with peak demand allocation to each stream in the group.

Network Provisioning (NP) is the set of the long-term control actions that determine the physical quantities of the resources (buffers, links etc.) to be placed in the network. The challenge with NP is to ensure that sufficient resources are available to accept all potential service demands with cost-effective configurations.

Given the network resources, as determined by suitable network provisioning, *Network Resource Management (NRM) is the set of functions related to resource configuration/allocation, which are performed by the network in order to achieve the basic performance and utilization objectives, with simple CAC and Routing procedures. Setting logical configurations through using Virtual Paths (VPs) and allocating resources to VPs are the principal NRM functions in an ATM Network* [5].

An ATM Traffic Descriptor is a set of measurable and controllable traffic parameters that can be consistently used by T&CC functions (e.g. for Connection Admission Control).

Connection Admission Control (CAC) addresses a set of actions taken by the network at connection set up phase (or during connection re-negotiation phase) in order to establish whether a Virtual Channel Connection or a Virtual Path Connection can be accepted or rejected.

The information required by CAC, called also the *Traffic Contract,* is

- the connection Traffic Descriptor
- the QoS class

Apart from accepting or rejecting a connection, the *CAC* algorithm may also determine which traffic parameters must be policed by the Usage Parameter Control, which routing functions are required, and also whether any re-allocation of network resources is necessary by calling the appropriate NRM functions.

Usage Parameter Control (UPC) & Network Parameter Control (NPC) address the set of functions performed by the network to monitor and control existing connections crossing a User-Network Interface (UNI) or a Network-Network Interface (NNI), respectively. The requirements from a *UPC* or *NPC* algorithm include the capability of detecting any illegal traffic situation, the rapid response to parameter violations and the simplicity of implementation. *Possible actions* against illegal performance are: (i) cell discarding, (ii) cell tagging (CLP bit: 0→1), (iii) cell re-scheduling (when combined with traffic shaping)

Traffic shaping is the function of altering the traffic characteristics of a stream in order to enforce certain parameter value(s). Peak-rate enforcement or burst spacing are typical examples of traffic shaping.

Priority Control refers to allocating resources to connections, bursts or cells according to a certain priority scheme.

2.2 APPLICATIONS, ATM-LAYER SERVICES AND RESPECTIVE TRAFFIC AND QUALITY-OF-SERVICE PARAMETERS

In order to devise an efficient set of service classes within a protocol layer, the overlaying applications should be carefully analyzed first, in terms of their traffic characteristics and Quality of Service (QoS) demands. Although this is not fully attainable for a B-ISDN, due to its open nature and the lack of exact knowledge of future application characteristics, some general application types have been identified and respective ATM-layer service categories have been specified. Table 6.1 gives such a classification.

Table 6.1: Classification of Communication Applications and ATM layer Services

	Classification attribute			
Application	Time constraints	real - time		non-real-time
Type	Traffic profile	CBR	VBR	-
ATM Service type		CBR	rt-VBR	nrt-VBR, ABR, UBR

The two fairly general application categories are *real-time* and *non-real-time.* Considering the configuration of fig. 6.1 (a) for information exchange between an *information provider*, A, and an *information consumer*, B, real-time applications are those requiring the information produced by A to be consumed at B within certain (usually tight) time limits, i.e. $D_{\mathrm{it}} < \max D_{\mathrm{it}}$. Information units that violate this condition are considered as lost. There are two different cases where such a requirement arise: (i) *in interactive applications* (e.g. voice communication, conferencing applications), and (ii) *in applications of non-storable data exchange.* The second type is faced in cases where not enough storage is available (at the consumer or within the Transfer System) to accommodate the information produced within time intervals greater than $\max D_{\mathrm{it}}$ (e.g. in live video distribution). Obviously, the actual time and data-storage constraints in the above definitions are case-dependent.

In the real-time type of applications (and only in that) there is a traffic profile associated with the application. Constant Bit Rate (CBR) and Variable Bit Rate (VBR) are the two general profiles. For non-real-time applications, the lower layers in the communication protocol stack (including the ATM layer) are free to make up any traffic profile desired through *traffic shaping.*

Focusing now on the ATM layer, we may reasonably assume that the upper layers contribute to the total delay with a constant amount, spent for processing. Thus we consider the adjacent upper layer (AAL) at the two sides as the *provider* and the *consumer* of the information, respectively, in the sense of fig. 6.1 (a). A typical probability distribution of the information transfer delay is shown in fig. 6.1(b). The delay still consists of two parts: a constant part, $\min D_{\mathrm{it}}$, and a variable part, due to queuing within the multiplexers and the switches. In [2] the parameters D_{it}, minD_{it}, and minD_{it} are referred to as CTD (Cell Transfer Delay), Fixed Delay and maxCTD (maximum Cell Transfer Delay), respectively, while the quantity (maxD_{it}-minD_{it},) is named peak-to-peak CDV (peak-to-peak Cell Delay Variation), for a specific delay percentile p: $\Pr\{D_{it}>\max D_{it}\}=p$.

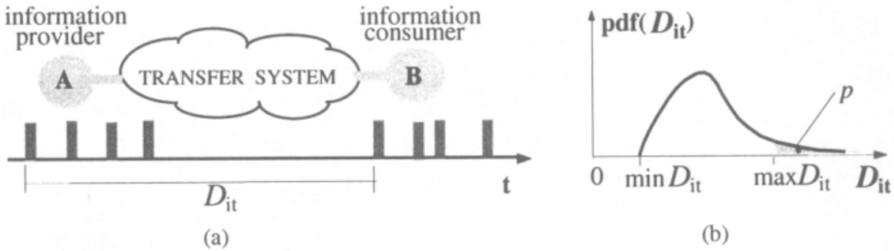


Figure 6.1: (a) information exchange configuration (b) Probability density function of end-to-end delay

In this context, four main ATM layer Services have been specified by the ATM Forum [2], as listed in Table 6.1. Three of them, namely the nrtVBR, the ABR (Available Bit Rate) and the UBR (Unspecified Bit Rate) are intended to support non-real-time applications. ITU-T specified in [13] only a CBR Service and a nrtSBR (non-real-time Statistical Bit Rate) Service, similar to the nrtVBR, while in this I-371 draft there is no equivalent to UBR. In the sequel we will be confined to the ATM Forum specifications, since they are in a more complete stage than the respective of the ITU-T.

For each of the above services two sets of parameters are specified: Traffic Parameters, constituting the *Traffic Contract*, and *Quality of Service* (*QoS*) parameters.

2.2.1 Traffic Parameters

For all ATM-layer Services listed in the Table 6.1, two traffic parameters are always specified: the Peak Cell Rate (PCR) and a Cell Delay Variation Tolerance (CDVT). These parameters are defined through using the Generic Cell Rate Algorithm (GCRA) [2,19]. In short, PCR is the maximum allowable cell rate, with some tolerance in the intercell distance dictated by the CDVT. For the rtVBR and the nrtVBR services an additional pair of parameters, the Sustainable Cell Rate (SCR) with a respective CDVT, or, equivalently, a Burst Tolerance (BT) is defined. The SCR is the maximum mean rate that can be sustained in the connection, again with a tolerance defined by using the GCRA.

2.2.2 QoS Parameters

Three negotiable Quality of Service parameters have been defined for the ATM layer;

- peak-to-peak CDV
- maxCTD
- Cell Loss Ratio: CLR = (Lost Cells)/(Total Transmitted Cells)

According to the standards, the first two are specified only for the real-time oriented Services (CBR, rtVBR), while CLR is specified also for the nrtVBR and, optionally, for the ABR Service.

2.3 STATISTICAL MULTIPLEXING / BUFFERING

2.3.1 Preliminaries

Referring to the capacity of a link (*time resource*), statistical multiplexing results in a finite probability that the sum of the instantaneous rates of the multiplexed connections exceeds the link capacity. To avoid losses, the portion of traffic in excess of the available capacity may be stored in a buffer (*space resource*) and transmitted later, when link capacity becomes available again. This implies delays in the transmission of information units and a respective degradation of the offered transport service quality. Moreover, there is a finite probability that the buffer overflows, if the link overload persists, resulting in cell losses. Thus, statistical multiplexing and buffering cause *delays* and *cell losses*, which are the two basic performance issues of interest. This is the penalty paid for the higher resource utilization achieved and the consequent service cost reduction. The bandwidth savings from statistical multiplexing and buffering is quantitatively described by the *statistical gain*, as defined in section 2.1.

Statistical multiplexing is not a brand-new concept in telecommunication networks. A trunk in a Plain Old Telephone Network (POTN) is shared among many subscribers, with connection requests arising at random. The bandwidth of the trunk is not dimensioned, of course, to cover the peak demand, since not all subscribers are active at the same time on the same route. There is, therefore, a non-zero probability that no bandwidth will be available for a new connection request, resulting in call blocking, estimated by the well-known *Erlang's loss formula*. In traditional packet-switching data networks packets share link bandwidth and buffer space also in a statistical manner. Coming to the ATM, statistical multiplexing is performed with the fine granularity of cells. In that respect ATM networks resemble packet-switching data networks. There are however substantial differences between the two cases, mainly in the performance metrics related to statistical multiplexing: In the traditional data networks, mean delay and mean load (throughput) have been the performance measures of interest. In the ATM networks, however, with the demand of supporting services with divergent traffic profiles and QoS requirements, including real-time services, higher order statistics of the cell losses and the delay are of the main interest.

Independently of the level of statistical multiplexing and the related performance measures, some fairly general principles can be stated. According to the law of large numbers, the larger the population of the statistics, the better the quality of the prediction that can be made from these statistics. In the case of statistical multiplexing it means that the larger the number of multiplexed connections, the more accurate the prediction of their combined characteristics can be and the larger the statistical multiplexing gain that can be achieved is.

Deterministic multiplexing is also possible with ATM, addressing the peak-rate allocation to each connection (at each link the sum of the peak rates of the multiplexed connections does not exceed the link capacity). As already mentioned, this is the likely operation of an ATM network in the first stage of its introduction, in order to obtain the required performance with simple traffic control mechanisms. ATM deterministic multiplexing does not mean, however, that no losses or no delays occur. Concurrent arrivals of cells from different input lines may result in a short-term link overload (cell-level congestion). So, even for a peak-rate allocation, small buffers are necessary to absorb cell-level congestion (see next section).

From the preceding discussion it becomes obvious that the multiplexer is the key element in an ATM network, which determines the basic performance metrics. Other network elements, like switches, or end-to-end paths can be modelled as suitable configurations of simple multiplexers. Devising suitable models of the multiplexed traffic that are able to capture the essential features related to statistical multiplexing and developing appropriate analysis tools is indispensable to applying effective traffic control mechanisms. In the following sub-section some commonly used traffic modelling and multiplexing analysis methods are discussed. Further important facets of statistical multiplexing are highlighted through performance analysis examples given in this section.

2.3.2 ATM Traffic Modelling and Multiplexing Analysis

A*n ATM traffic model is a set of rules (mathematical or other) that govern the generation of new connection requests and/or the generation of ceils or groups of cells (bursts) within a particular connection.*

According to this definition, three different levels of traffic description arise in an ATM environment: *Connection*, *burst* and *cell level*. The discussion here focus on the last two levels, since connection level models do not differ substantially from those in traditional circuit-switching networks.

2.3.2.1 Cell-level models

Since the physical transmission underlying the ATM layer is a slotted, synchronous process, an exact ATM traffic model would consist of a *discrete point process* describing the location of ATM cells on the discrete time lattice.

Markovian processes keep a prominent position in queuing theory. They offer sufficient flexibility and substantial analytical tractability, due to their memoryless property. A rather general and versatile Markovian point process is described and analysed in [34], while the respective queuing model (denoted by *N/G/1* for a general service time distribution, *N* standing for the arrival process from Neuts, the name of its proposer) is handled in [35]. Many classical point processes are derived as special cases of the N-type process, like the Phase-renewal process (Erlang-type processes belong to this category), the Markov Modulated Poisson Process (*MMPP*), etc. The N-type process is a continuous-time process, but the definitions and the derivations in the two papers cited above are extendible to point processes on the discrete time lattice. Thus, the *MMBP* (Markov Modulated Bernoulli Process) is the discrete version of the *MMPP*.

Coming now to our analysis problem of an ATM multiplexer, two difficulties arise when using such point processes for modelling ATM streams:

- The superposition of many such streams feeding the multiplexer gives rise to a very complex arrival process which is difficult to model. The reader should think of the superposition of renewal processes, which is not necessarily a renewal [7].
- Even if we succeed in modelling the aggregate input process, the resultant state space of the discrete-time queuing model is large (especially for large buffer sizes, if a finite-buffer model is considered). Special decomposition techniques, if applicable, can reduce the order of computations and alleviate the numerical instability or ill-conditioning problems usually associated with large state-space models. In [8] such a decomposition method is used to analyze the multiplexing problems of *MMPP*s.

2.3.2.2 Phase- or Burst-level models

An alternative approach to ATM traffic modelling and the associated multiplexing analysis problem is the one that ignores the cell-level details and concentrates on the phase- or burst-level statistics. A class of this type consists of the so called *fluid-flow models*, according to which groups of cells with constant or "almost" constant inter-cell distance (a random delay with a small - compared to the inter cell distance - variance may be added to the latter) is substituted by an equivalent continuous fluid-like flow of information. The fluid version of an *MMDP* is called, following [36], Markov Modulated Rate process (*MMRP*), where, at each state of the underlying Markov chain, a constant information flow is assumed.

A burst-level modelling method, as opposed to an exact, cell-level discrete-time modelling, has the following arguments on its side, when used for traffic control:

- The cell-level fluctuations will be absorbed by the buffers (necessary anyway to absorb cell-level congestion) in a well-dimensioned system, so only the burst level statistics are of real interest for the traffic control issues (note that for dimensioning both scales are necessary).
- Traffic control mechanisms, especially those applied real-time, are starving for simple models and analysis algorithms, not affording the computational load implied by most of the cell-level techniques.
- Even if an exact and easy-to-handle cell-level model was available for an ATM traffic stream at a certain point of the connection, this would not be valid after some multiplexing stages, due to the random queuing delay cells are experienced

through these stages. In contrast, burst-level statistics are likely to be roughly maintained in a connection all its way through the network.

- Most of the traffic sources in the all-service networks of the future will be bursty, thus a burst-level modelling is more suitable to capture the source dynamics and describe an end-to-end connection

In the rest of this section some basic fluid-flow models and respective methods for multiplexing analysis are reviewed.

Multiplexing of Markov modulated fluid streams have been extensively studied in the literature. The most general treatment is found in [36], which considers general time-reversible Markov modulating chains and exploits the properties of such chains arising from traffic superposition, to decompose the global system into smaller and, consequently, easier to solve subsystems.

A two-state *MMRP*, alternating between On and Off exponentially distributed phases, is the simplest model featuring a bursty behaviour and allowing for the qualitative interpretation of the derived results. Burstiness is defined in a clear way (the peak-over-mean ratio) and the few (namely three) parameters required for the model description are easily measurable. In [1] a closed form solution for the homogeneous traffic multiplexing problem is derived, while Kosten [22] extended these results to the heterogeneous case. In [19] the results of the more general theory in [36] are connected with the particular structure of the On/Off streams to give even more powerful, stably computable and easily implementable results.

The On/Off type of traffic seems to be of the most frequently encountered in the forthcoming ATM networks. By realizing that multirate or other VBR traffic streams are difficult to control and to get a profit out of their statistical multiplexing, transforming them into On/Off streams, through suitable shaping at the transmitter, offers a promising approach to handling bursty traffic without sacrificing the multiplexing gain. Moreover, the worst-case traffic streams that pass through standardized shaping or conformance monitoring mechanisms (e.g. the GCRA algorithm [2,13,14]) is of the On/Off type. However, the case of exponentially distributed On and Off periods, which is accurately modelled by a two-state *MMRP,* will be only an exception. Peak-rate shaping, for example, tends to form hyper-exponentially distributed bursts (consecutive information blocks may merge after shaping, forming larger bursts, while burst-size shaping (e.g. imposing a maximum size) is likely to give a hypo-exponential burst size distribution. In [39] it was found that the cell streams within the switching network can be modelled with hyper-exponential distributions for both On and Off periods. In such cases the simple exponential On/Off model fails to give acceptable results and models of a higher dimension are necessary [e.g. 27].

The issue of heterogeneous traffic multiplexing is dominant in an integrated-service environment, like an ATM network. The resulting aggregate streams feature dynamics in more than one time scale giving rise to increased modelling and analysis complexity. Fortunately, in many practical cases the different time scales are ... very different and Nearly Completely Decomposable (NCD) Markovian fluid models may apply. Such cases include mixing of slow and fast traffic sources or sources featuring different time scales internally (e.g. video sources featuring line, frame and scene changes). The NCD structure gives rise to analysis techniques, based on model decomposition, that, not only lead to efficient and tight approximations for the global system behavior, but also clearly identify the important contributors to the small- and large-buffer dynamics, thus enabling a better insight into the multiplexer's performance [20]. When the NCD structure originates from mixing slow and fast sources, the relevant results take a particularly simple form (see [30] for a brief review and experimental validation).

2.3.3 Basic performance metrics

The various traffic models and the respective multiplexing analysis methods reviewed in the previous section can give a variety of results, from simple mean figures (e.g. mean buffer content, mean delay) to detailed buffer occupancy distributions. Since, as mentioned earlier, the B-ISDN service requirements from the ATM layer are usually expressed in terms of higher order statistics (e.g. the peak-to-peak CDV), analysis methods that can give such statistics are appropriate. We focus here on the fluid-flow methods, although similar results can in principle be obtained by cell-level modelling methods too. We also exclude from our discussion here connection-level performance, which would require appropriate traffic modelling and also the involvement of traffic control functions at this level, like connection admission control. Finally, we assume a single-class system without priorities; the reader can refer to [23] for concepts and analysis methods of queuing systems with priorities.

2.3.3.1 Performance measures derived through a buffering model

First we consider a buffering model. The base performance measure that is assumed to be obtained by a multiplexing analysis method, is the *Complementary Probability Distribution Function (CPDF)* of buffer occupancy, defined as $G_b(x) \equiv Pr\{buffer\_conent > x\}$. A first approximation that is usually made in calculating $G_b(x)$ is related to the boundary conditions. If, instead of taking the actual buffer size, we assume an infinite buffer, we get an upper bound to $G_b(x)$, name it $G(x)$. The larger the actual buffer size, the tighter that bound is (in a normalized scale). The infinite buffer assumption, apart from rendering the analysis problem much easier to solve, gives also a solution usable for *any buffer size.* In what follows we deal with $G(x)$, keeping in mind that it is an upper bound to the actual occupancy distribution.

$G(x)$ is the "average" buffer occupancy distribution observed at a random time instance. Obviously, this is not what the cells of a bursty traffic stream (contributing to $G(x)$) see. In [19] a simple relationship between $G(x)$ and the buffer occupancy seen by the cells of the multiplexed traffic streams is established. In particular, it is shown that the latter is given by

$$H(x) = \frac{G(x)}{\rho}, \qquad (6.1)$$

where ρ is the average (normalized) load of the multiplexer.

Moreover, if the multiplexed traffic is heterogeneous, different buffer occupancy distributions are seen by the different traffic classes. In [19] the per-class distributions are also calculated. It is worth noting the special case of CBR-plus-*MMRP* traffic multiplexing, where the following simple formulae hold:

$$H_{CBR}(x) = G(x) \qquad H_{MMRP}(x) = \frac{(1-\rho_{CBR})}{\rho_{MMRP}} G(x) \tag{6.2}$$

Experimental results validating the above, theoretically derived relationships are found in [30].

Having obtained the buffer occupancy distribution $H(x)$, seen by a particular traffic stream, one can directly derive estimates of the basic QoS figures of the ATM layer, namely the cell losses and the cell delay: With the assumption that no priorities are implemented in the multiplexer (see next paragraph for prioritized schemes), $H(x)$ gives also the cell delay distribution, while its value at x equal the buffer size gives an estimate of the Cell Loss Ratio (CLR) for that particular traffic stream. It should be emphasized once more that what we get are upper bounds to the respective actual figures for both the delay and the CLR.

Now, since $H(x)$ is also the distribution of the delay that a cell is experienced through the multiplexer, equation (6.1) can be interpreted as a "fluidified" form of the Little's formula concerning distributions. Note also that for highly loaded systems (ρ close to 1) $H(x)$ does not differ too much from $G(x)$.

2.3.3.2 CLR estimation using a bufferless fluid model

When the buffer size is small, compared to the size of the multiplexed bursts, or when we do not have any indication about the burst size distribution, a safe estimate of the CLR can be obtained by considering as lost all the traffic that overflows the link capacity. If a stationary flow-rate distribution, $f_x(x)$ can be calculated for the multiplexed traffic, CLR is given by

$$CLR = \Pr\{link\_overflow\} = \frac{\int_C^{\infty}(x-C)f_x(x)dx}{\int_0^{\infty}xf_x(x)dx} \tag{6.3a}$$

We can identify in the numerator of (6.3a) the average rate of the input traffic above the link capacity, while the denominator is simply its average rate (see fig. 6.2). In the case of homogeneous multiplexing of On/Off traffic streams, each featuring a peak rate equal to *c* and an average rate equal to *r*, the above equation takes the form

$$CLR = \frac{\sum_{n=\left\lceil \frac{C}{c} \right\rceil}^{N} (nc - C)\binom{N}{n}\left(\frac{r}{c}\right)^{n}\left(1-\frac{r}{c}\right)^{N-n}}{Nr} \tag{6.3b}$$

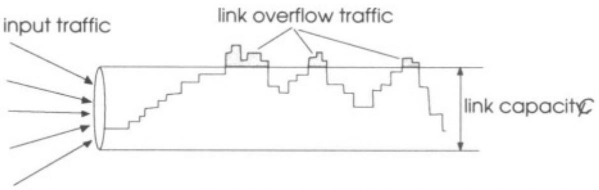


Figure 6.2: A buffer-less fluid model

For large *N,* approximations of the overflow probability can be obtained through a *Gaussian distribution* (see, e.g. [11]) or the *tilted probability distributions* [12]. These approximations allow also for solving the inverse problem of traffic control, e.g. for deriving the required bandwidth in order to ensure certain CLR bound.

2.3.4 A performance analysis example - Comments on Statistical multiplexing & Buffering

Here is a performance analysis example with homogeneous traffic multiplexing of On/Off Markovian streams. Eight such streams are multiplexed, each featuring the following traffic parameters (drawn from [28]):

normalized peak rate (constant in the On period): *c*=0.25 (one cell every 4 slots)
mean burst volume: *V*=10 cells (mean burst duration: *V/c*=40 slots)
normalized mean rate: *r*=c/5=0.05 (burstiness = 5)

Fig. 6.3 summarizes the multiplexer's performance, as obtained by different analysis methods: *Exact results*, derived by discrete-time cell-level simulation with a buffer size equal to 20 cell spaces, are shown in thick solid lines. The lower curve is the average (over time) complementary distribution of buffer occupancy, while the upper curve is the distribution seen by the arriving cells. Respective curves, derived with a buffer size equal to 1000 cell spaces (practically infinite) are shown in dotted lines. Shown in thin solid lines are the corresponding curves derived by the *fluid-flow analysis* with the infinite buffer boundary condition. The theoretical distance of the two curves *(G(x)* and *H(x)*, according to the terminology of section 2.3.3) is given by (6.1), which in log scale is equal to logl0(1/(8*r*))=0.397. We can observe the agreement of the fluid-flow results with the respective simulation results for a large buffer size in the burst-congestion region (>3 cells in this case). In the cell-congestion region (<=3 cells) the *M/D/1 curve* gives a better approximation to the actual performance.

H(x) curves are also delay distributions through the multiplexer, given that no priorities are implemented. As far as the CLR is concerned, the simulation gives a number around -3.81 (log scale), the fluid-flow upper bound (with an infinite buffer assumption) is -3, while the bufferless model through the formula (6.3b) gives a value -2.53.

Three more curves, derived by the fluid-flow approximation, are shown in fig. 6.3. The upper, thin solid line corresponds to streams with the same parameters as before, except for *V*, which is taken equal to 50 cells now. The lowest two curves correspond to another peak rate, equal to 0.143 (one cell every 7 slots), with *V*=10 and 50, respectively.

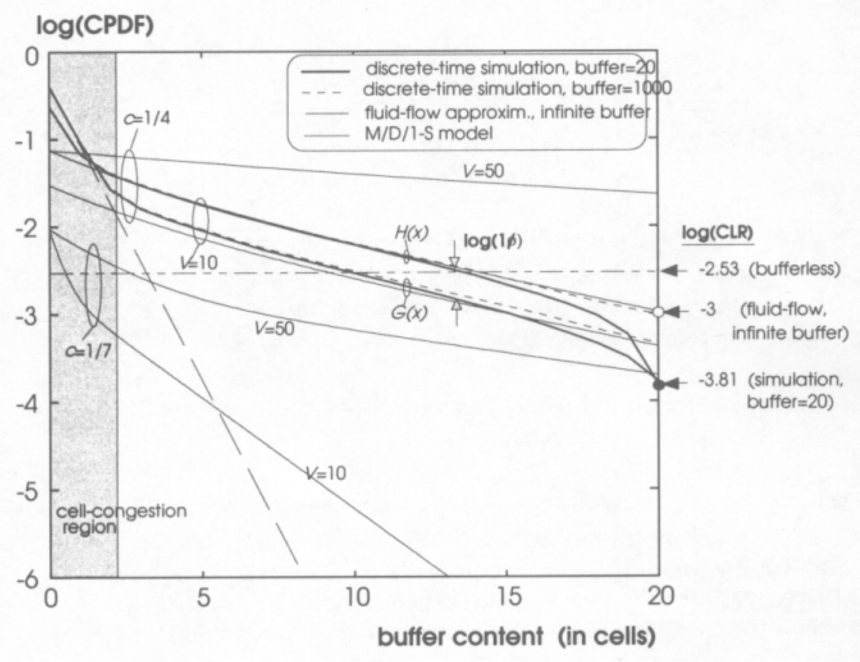


Figure 6.3: Multiplexing performance analysis, homogeneous On/Off traffic

Despite the very specific type of traffic used here (On/Off, homogeneous), some more general results concerning the statistical multiplexing and its analysis by the various methods can be drawn from this example:

- The cell congestion region is bounded by the maximum number of multiplexed streams, but usually is much smaller.
- In the cell-congestion region the *M/D/1* model can provide a good approximation.
- The CPDF obtained with an infinite buffer assumption is an upper bound to the actual curve of the finite-buffer model.
- The fluid-flow analysis gives a very good approximation to (upper bound estimate of) the CPDF, apart from the cell-level congestion region. The buffer occupancy CPDF seen from the arriving cells, when calculated at the buffer size, gives an upper bound to the Cell Loss Ratio. For low losses, this bound can be a tight one.

- The size of the bursts (compared to the buffer size) is critical for the multiplexer's performance (as discussed in section 2.3.2.2, the distribution of the burst size is also critical). For large bursts, a simple estimate (upper bound) of the CLR is obtained by the bufferless model.
- The peak rate is also a very critical parameter for the performance. A large sensitivity of the results with respect to this parameter is observed in the region close to deterministic multiplexing (i.e. close to the value: [link rate]/[number of streams]).

2.4 STATISTICAL GAIN AND EFFECTIVE BANDWIDTHS

2.4.1 Preliminaries

The inverse of the analysis problem, examined in the previous sections, consists in fixing certain QoS demands, in terms of either a CLR or a delay percentile, and finding the minimum amount of resources (bandwidth and/or buffer) that can maintain the required QoS. As already observed, there are two distinct mechanisms that lead to bandwidth savings when serving statistically varying traffic streams: the mutual *rate compensation* that takes place in multiplexing of many such streams on the same link (referred also to as *statistical multiplexing*) and the *buffering*. The first prevails in the cell-scale region and is effective even in a bufferless system; the second prevails in the burst-scale region and is effective even for a single source. Both of them are combined to give the so called *Statistical Gain*.

The ultimate goal of most modelling and analysis approaches is to devise simple albeit as accurate as possible formulas or algorithms that can solve the traffic control problem. Towards this end, a lot of work has been devoted to finding a simple scalar quantity that would describe the bandwidth effectively consumed by a stream or by a group of streams in the context of the previous paragraph. The sought quantity is named *Effective Bandwidth* or *Effective Rate* or *Equivalent Capacity*. With reference to fig. 6.4(a), showing a simple multiplexer (without priorities), the Effective Bandwidth of the N connections multiplexed on the link is equal to the required link capacity, C, so that the buffer occupancy distribution is marginally under a specified point (p, V_b) (corresponding to the desired delay percentile); a similar demand can alternatively be set for the overflow probability (corresponding to the desired bound to the CLR).

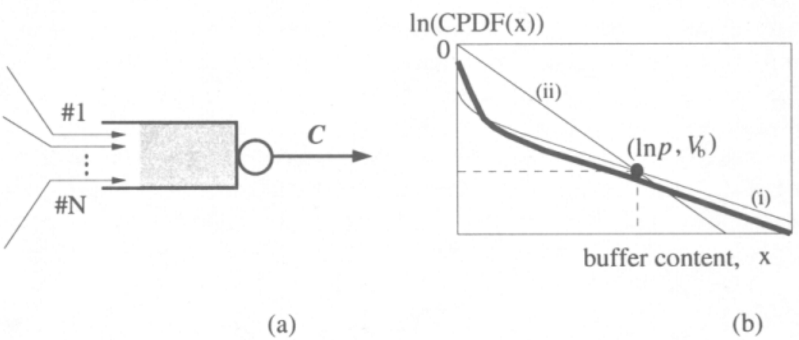


Figure 6.4: Effective Bandwidth definition, (a) multiplexer (b) performance plane

2.4.2 Variants and properties of Effective Bandwidth

Additivity of Effective Bandwidths:

A property of the *EB* that, if held, would greatly simplify the traffic control problem, is *additivity.* An *EB* definition possesses the additivity property iff

$$EB\{\text{group\#1+group\#2}\}=EB\{\text{group\#1}\}+EB\{\text{group\#2}\}, \tag{6.4}$$

where the same QoS figure is used for both groups of connections.

The most stringent definition of the Effective Bandwidth, that exploits all of the statistical gain potentially available, would be obtained by demanding the "exact" or marginal satisfaction of the specified QoS figure (thick line in 6.4(b) passing through the QoS point). Since, however, "exact" performance is difficult to calculate and only approximations are in most cases available, approximate figures for the Effective Bandwidth are derived. We distinguish the following approaches:

A. Effective Bandwidths based on bufferless models

As we have seen, bufferless models are suitable in cases where buffering does not contribute to the statistical gain (e.g. the bursts have size comparable to or larger than the available buffering space). In such cases CLR is the only performance measure applicable. Fixing a CLR and using appropriate approximations for the stationary rate distribution so that equation (6.3) may be solved for the link capacity *C*, can yield the sought *EB*.

Example: For two-state Markov streams, not necessarily identical, multiplexed on a bufferless link, a good approximation of the required link capacity that respects a small overflow probability bound, *p*, can be obtained through the Gaussian distribution as [11]:

$$\hat{C} = m + \alpha'\sigma, \quad m = \sum_{i=1}^{N} m_i, \quad \sigma^2 = \sum_{i=1}^{N} \sigma_i^2, \quad \alpha' = \sqrt{-2\ln p - \ln(2\pi)}, \tag{6.5}$$

where *m* is the mean and σ the standard deviation of the aggregate bit rate (m_i, σ_i are the respective quantities of the i^th^ stream). A finer result can be derived through the *Bahadur-Rao* approximation, expressing the tail of the stationary bit rate distribution in an exponential form [12]. The result is not given in closed form, however, and the increased numerical complexity is the penalty paid for the approximation improvement obtained.

Advantages: Simplicity of the result (at least with the Gaussian approximation).

Disadvantages: Buffering gain not exploited. Delay-expressed QoS cannot be handled. Additivity does not hold, even for groups of similar traffic streams , i.e. $\hat{C} / N$ in (6.5) depends on *N*.

B. Effective Bandwidths based on buffering models

In this approach, any of the fluid models, as those addressed in section 2.3.2, can be used to yield a close approximation to the actual buffer occupancy distribution in the burst-scale region [e.g. line (i), approximating the thick line in fig. 6.4(b)]. After translating the QoS requirement into a specific point on this plane, we can (approximately) compute the required link capacity that marginally respects this point. Unfortunately, no closed form solutions exist, so numerical methods are employed in a trial-and-error procedure.

Advantages: Both mechanisms yielding a statistical gain, i.e. multiplexing and buffering, are exploited

Disadvantages: Difficult computation. Additivity does not hold.

C. Effective Bandwidths based on asymptotic approximations

This is what is usually met in the literature. It was originally proposed by [10,11,15] and was applied to Markovian On/Off streams and other simple models. Then it was extended to a more general Markovian framework [9] and other stationary sources [16]. A review of the Effective Bandwidth theory with further references and a consideration from the statistical mechanics point-of-view can be found in [6]. Recently, the Effective Bandwidth of a class of non-Markovian fluid streams was proved to exist and explicitly determined [21].

The asymptotic definition of the *EB* takes into account only the slope of the buffer occupancy *CPDF,* enforcing a line of this slope to pass through zero and the QoS point [line (ii) in fig. 6.4(b)]. In mathematical terms, $G(V_b)$ is approximated by $e^{y_o V_b}$, which is enforced to become equal to p, or $y_o = \ln p / V_b$ (see e.g. 3.2). In this equation y_o is the largest (negative) eigenvalue of the system.

Advantages: Calculations are confined to the determination of only the dominant eigenvalue of the system. Additivity holds.

Disadvantages: The gain from statistical multiplexing is not exploited.

The only definition of the effective bandwidth that features the additivity property is the last one (case *C*). The asymptotic Effective Bandwidth (*aEB*), being additive, does not depend on the multiplexing environment (e.g. on the number and the profiles of other streams multiplexed together), but only on the characteristics of the individual stream it refers to and the QoS point. Indeed, a direct definition of the *aEB* of a cell arrival process is given (under some conditions, see e.g. [15,16]) by

$$aEB(p) \equiv \frac{\lim_{t\to\infty} t^{-1} \ln E\{e^{pA(0,t)}\}}{p} \tag{6.6}$$

where A(0,t) is the number of cell arrivals in the interval (0,*t*].

One problem arising in determining the *EB* from delay constraints is that the latter cannot be directly transformed into a specific point on the buffer occupancy - probability plane, as shown in fig. 6.4(b), since such a transformation involves the unknown serving capacity (equal to *EB*): content *x* ➔ delay *x/EB*. To overcome this problem a recursive procedure which would successively approximate the sought *EB* may be necessary, even in cases where closed-form performance analysis results are available.

2.5 A CRITICAL VIEW ON ATM SERVICE SPECIFICATIONS, PERFORMANCE ANALYSIS METHODS AND RELATED TRAFFIC CONTROL APPROACHES

Nothing more true than the saying: “the colour of the surrounding world is the colour of the glasses we wear”. If this colour happens to be pleasant to our eyes, we are not willing to put the glasses down! In the case of deterministic system analysis this colour is named “linearity”, while in the stochastic case it becomes “stationarity”, “Poisson”, or “Markovian”. In this section we’ll try to look at some of the traffic control issues addressed in the previous sections with the naked eye.

2.5.1 A debate on the ATM-layer services and QoS

How many ATM transfer capabilities do we need? Are all services listed in Table 6.1 really necessary, or they have been specified just to compromise the different trends, given the lack of concrete knowledge about the real future needs? The main argument supporting a positive answer to the above question is, of course, the usual pluralistic statement: "let them be there and the most appropriate ones will be selected and used". This is true, no doubt, but a multitude of other control and management functions should be developed before the capabilities of each of the ATM service categories could be fully exploited. In some cases (as for the ABR service) a special hardware must also be implemented to support the service.

Another questionable issue is the specified QoS parameters. We have seen (section 2.2) that, according to the standards [2,13], the *Cell Loss Ratio* (CLR), the *Cell Delay Variation* (peak-to-peak CDV) and the *Maximum Cell Transfer Delay* (maxCTD) are the primary ATM-layer QoS parameters. It should be realized, however, that except for the maxCTD, the other two parameters are probabilistic in nature and, as such, they should be evaluated "over the long term and over multiple connections with similar QoS commitments" [2]. For the CLR in particular, a statistically stable evaluation of a value in the region e.g. 10^{-7}, requires a number of transmitted cells around 10^{10}, or 500 Gbytes! This is not the order of transmitted data over a typical connection.

For non-real-time applications cell losses are handled by the upper layers which request for retransmission of any lost or corrupted data units. Thus, a lost cell is perceived by the end user as an extra delay in receiving the so-corrupted data unit. Probably it is not even observed, if most data units arrive with a considerable delay due to queuing through the network nodes. For real-time applications, excessive delays render the received information units useless and, therefore, practically lost. These "effective losses" may be much higher than the cell losses within the network switches; the latter may be kept arbitrarily rare through using appropriately large buffers.

Given the difficulty in guarantying and evaluating CLR figures on an individual connection basis and the observations of the preceding paragraph, the following statement sounds reasonable:

The primary QoS issue that traffic control has to deal with is delay. Cell losses might be handled mainly at the network (buffer) dimensioning level.

2.5.2 Suitability of statistical models for traffic control

Most of the ATM performance analysis approaches found in the literature (some of them addressed in section 2.3) are based on two basic assumptions: (a) *traffic source stationarity* and (b) *applicability of some type of Markovian model.* There are, however, some drawbacks with these assumptions in a traffic control framework. First, nobody can assure stationarity of a traffic source (it is the exception rather than the rule). Second, it is rather impossible to enforce a process to follow a Markovian behavior by simply controlling (or shaping) a set of parameters. Any deterministic action (control) takes us away from randomness, exponential behavior, Markov models and the like, as by definition. More dramatically, even simple statistical parameters, like the mean or the variance of a random variable, which are ensemble quantities, cannot be adequately enforced on sample processes, e.g. on individual traffic streams (stationarity is again a sought property). For example, how can one enforce a specific mean rate on an ATM stream? If the answer is by enforcing averages over a time window, then how long must the averaging window be, and what is the influence of possible non-stationarities within the chosen window? More complex statistical parameters of ATM streams are much harder to estimate and control (see, for example, [38] for estimation of Effective Bandwidths).

It is specified by the standards that the control functions should be based on parameters included in the traffic contract and that any parameter in the contract should be understandable by the end systems, and also be measurable and controllable. Under the observations of the previous paragraph, no statistical parameter can be used as such (only maximum values, in the case of variable quantities). Does this mean that we should forget statistical multiplexing and the gain we can get out of it? The answer is definitely no, although Markovian models seem hardly applicable. Even if we shape the traffic streams in a deterministic way, we may get a statistical gain, provided that the multiplexed streams are sufficiently many, independent to each other and multiplexed with a random phase. Such conditions may be ensured in a large network.

2.5.3 Rate shaping and statistical gain

It is instructive to bend over a simple rate shaping example. A packet stream of the *M/D* type (i.e. exhibiting Poisson arrivals of constant-size packets) feeds a rate shaper (cell spacer) producing a cell stream of a limited peak cell rate. After being shaped, a number of independent such streams are multiplexed together (fig. 6.5(a)). The multiplexer is simulated and its behavior is shown in fig 5(b) for the following data: number of multiplexed streams N=35, packet size v=50 cells, packet arrival rate λ= 1/2500 time_unit $^{-1}$, multiplexer's output rate C=1 cell/time_unit (then, average normalized load ρ=0.7).

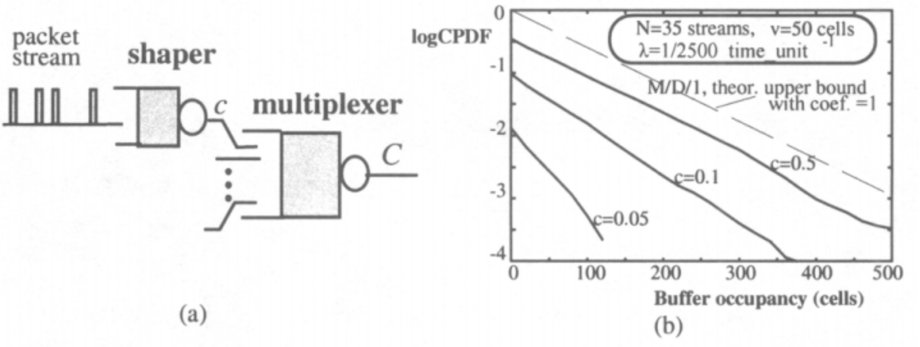


Figure 6.5: Multiplexing of rate-shaped packet streams (a) configuration (b) buffer occupancy CPDF curves

Looking at the curves of fig. 6.5(b) one can observe the following:
For relatively high peak cell rates, the performance improvement through rate shaping manifests itself mainly as a displacement of the curves towards lower values, rather than as a significant change in the slope. In other words, rate shaping for peak cell rates fairly larger than the mean rate affects mainly the coefficient and not the exponent of the dominant term of the CPDF (statistical multiplexing gain prevails).

The above mentioned improvement cannot be exploited by asymptotic effective bandwidths, since the latter are based on the dominant-term approximation with a coefficient set equal to one. Setting the coefficient equal to one is the only approximation that defines an effective rate with the nice property of additivity, but essentially it does not exploit the statistical multiplexing gain (see section 2.4.2).

The above observations mean that in order to get a significant improvement in multiplexing performance through rate shaping, which could be exploitable by effective-bandwidth based methods, a drastic compression of the peak cell rate should be performed. This, however, would transform all traffic streams to near-CBR ones, increasing the actual delay seen by the end systems. Moreover, large buffers are required at the end systems (e.g. by the shaper of fig. 6.5(a)) which increase their costs since they account for non-shared resources. We should keep in mind that the upper layers usually exchange information packet-wise and not rate-wise (entire Protocol Data Units - PDUs - are exchanged, e.g. IP datagrams or MPEG frames).

Calculate this number as a function of V_b for the following cases:

I. multiplexing (without shaping)

II. multiplexing of (rate-shaped) on/off streams; use of dominant term approximation with coef.=l (asymptotic effective bandwidth approach)

III. as in (b); use of a better approximation fitting the actual curve of CPDF, see [29].

Typical results of this experiment are found in [31] and quoted in fig. 6.6(a) and (b) below.

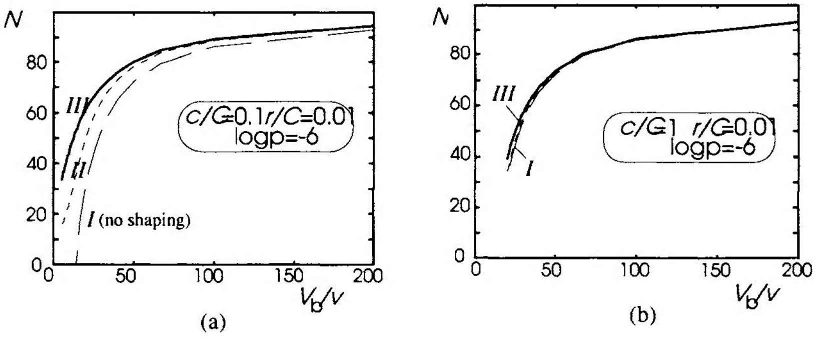


Figure 6.6: Maximum load (number of streams) respecting, $\log CPDF(V_b) < 10^{-6}$
(a)c=0.1C (b) $c=C$

An apparent observation is that for large V_b / v values (v is the average burst size of the multiplexed streams) the three curves converge to each other. Thus any performance improvement derived by rate shaping is predominated over by the buffering gain. This is perfectly in line with our physical intuition: *when purring cups ofwater into a large reservoir, it doesn't matter ifemptying the cups at once or drop by drop*! But the most desirable feature of the buffering gain is that it is exploitable by the asymptotic effective bandwidth approaches with all the nice properties exposed in section 2.4.2.

3. A NOVEL, BURST-LEVEL APPROACH TO ATM TRAFFIC CONTROL

In the previous section it was demonstrated that

- a strong rate shaping requires large buffers at the end terminals, increasing their cost and also increasing the delay seen by the higher protocol layers, while
- a light rate shaping does not give a significant performance improvement exploitable by tractable effective-bandwidth-based methods; on the other hand,
- using large buffers within the network nodes increases the buffering gain (further diminishing the contribution of rate shaping to the statistical gain).

By taking also into account that

- rate can not be associated with an end-to-end connection extending over VPs/ lines with different bandwidths/transmission speeds (a very likely situation in a global, multi-domain network), and also that
- rate may be altered drastically through successive multiplexing stages,

we propose a new paradigm of traffic shaping and control based on a burst-level modelling, which essentially ignores cell-level details within bursts (e.g. the cell rate). Cell-level processes, however, not exhibiting burst-level dynamics (e.g. CBR streams) could in principle be modelled by the proposed scheme through limiting the burst size to a single cell. The proposed approach exploits the buffering gain, assuming a large *buffer_space/burst_size* ratio and is perfectly in line with the forthcoming IP-ATM convergence; notice that in the QoS-aware IP, rate is not included in the traffic descriptor as a controllable parameter, and that a new ATM service, the *guarantied frame rate* (GFR) has been recently proposed, [3], to facilitate the desired convergence. Under these circumstances, the *M/G/1* model of the classical queuing theory comes into the foreground and could serve as the basic gear of traffic control mechanisms in such heterogeneous and multi-service environments [32]. In the rest of this section the main results from the *M/G/1* queuing will be reviewed and respective control mechanisms will be outlined.

3.1 PRECIOUS RESOURCE AND QOS METRIC

One basic assumption throughout this section is that *bandwidth* is the precious resource, even in broadband infrastructures, while buffering space can be readily made available at contention points (e.g., at backbone nodes).

Another assumption is that QoS is mainly intended for the real-time applications (like Internet Telephony, Internet TV, multimedia on WWW, etc.), thus *delay* is the prime QoS parameter of concern. This holds even for commercial services such as Virtual Private Networks (VPN). Data losses can of course affect the service quality perceived by the end user; we assume however that appropriate buffer dimensioning is sufficient to keep the respective figures at acceptable levels. Under the previous assumptions and the reasoning exposed in 2.5.1, the basic QoS metric of concern is a *delay percentile*, i.e. the probability of exceeding a certain delay threshold (either at a single node or end-to-end).

3.2 MAIN RESULTS FROM THE *M/G/1* MODELLING

Consider a traffic stream as a sequence of *Data Units* (a DU is a burst of ATM cells in our context or an IP packet, in an IP context) transmitted at certain time instances. We further assume that the time and the DU size are both quantized into integer values, expressed in *time slots* and *data atoms* (elementary DUs, e.g. ATM cells), respectively. We also assume batch transmission of an entire DU (at an infinite rate), which is the worst case for the multiplexer accepting this stream.

We will now focus on the case of data streams of the *M/G*-type (i.e. featuring Poisson arrivals and generally distributed burst (packet) size; the Poisson assumption of aggregates of shaped burst streams has been validated under certain assumptions in [32]). The queue-length distribution of an *M/G/*l is asymptotically governed by the dominant (i.e. less negative) solution, q_o , of the *M/G/1* characteristic equation,

$$q_o = \sup\{q : q = \frac{\lambda}{C}(1 - V(q))\}, \tag{6.7}$$

i.e. a Chernoff bound holds in the form

$$\gamma e^{q_o x} \leq W_c(x) \leq e^{q_o x}, \tag{6.8}$$

where λ is the mean DU arrival rate; $V(s) \equiv E\{e^{-sv}\}$: the moment generating function of the DU size (denoted by v); C is the service rate; $W_c(x) \equiv \Pr\{queue\_lengh > x\}$: the queue-length CPDF; and γ some positive number $\in (0,1)$ (see e.g., [17,18]).

In the above expressions, q_O is the asymptotic slope in logarithmic scale of the queue-size CPDF that determines the performance level seen by the packet streams at a node's queue, i.e. $q_o = \ln p/V_b$, with p the probability that the buffer content exceeds V_b data units. Extrapolating this distribution to the actual buffer size available for the queue gives a close bound to the data loss, while appropriate scaling by the allocated service rate transforms it to a delay distribution .

For a desired q, one may calculate the required service rate *C(q)* of the multiplexer, so that a specific CPDF slope is maintained. Solving (6.7) for *C*, gives:

$$C(q) = r\frac{1 - V(q)}{q\bar{v}}, \tag{6.9}$$

where $\bar{v}$ is the average DU size and $r = \lambda\bar{v}$ is the mean rate of the stream. According to the terminology of 2.4.2, the quantity *C(q)* is the asymptotic *Effective Rate*, name it *f(q)*, of the *M/G* stream for the specific QoS setting q. In [32], the following important theorem was proven, establishing the additivity of *f(q)*:

Theorem 1: *The effective rate of a stream in an M/G/1 system, calculated by (6.9), is summable, i.e. the transmission rate required by two or more independent streams to maintain a specific QoS figure q is equal to the sum ofthe rates required by the individual streams for the same QoS, independently of their packet size distributions.*

What is also important for a coherent and complete traffic control paradigm is the ability to produce streams of desired *f(q),* given a target *q*, through shaping, as well as to apply a similar function for conformance testing (policing). These significant aspects are covered in the following sub-section.

1.3 BURST-LEVEL TRAFFIC CONTROL MECHANISMS

1.3.1 Shaping

In our context, traffic shaping is implemented by suitable spacing of consecutive *DU*s (packets or bursts of ATM cells). The following mechanism and the associated theorem 2, quoted from [33], give the way.

The spacing mechanism

After each DU of size v, enforce a silence s (gap between the current and the next DU), given by:

$$s = \frac{1}{f}\frac{1-e^{-qv}}{q} \qquad (6.10)$$

The above mechanism is referred to as a *tight shaper*, in contrast to a *credit-based* (or *token-bucket*) *shaper,* which allows certain tolerance in the *intcr-DU* distance.

A credit-based (token-bucket) shaper

According to [33], a stream of *DUs* is shaped for a multiplexing quality *q* and an effective rate *f*, if

$$c \equiv \left(\sum_n \frac{1-e^{-qv^{(n)}}}{qf} \right) - T \in [c^-, c^+] \quad \forall \quad T >> (c^+ - c^-), \qquad (6.11)$$

i.e., the quantity *c*, named *credit*, remains between specific bounds $[c^-, c^+]$ for any arbitrary time window of duration $T >> (c^+ - c^-)$; in the above formula, the summation index, *n,* includes all the DUs within the considered time window.

A shaped stream, according to the above definition, remains shaped even after passing through a network with a variable (but bounded) delay, provided that we are allowed to enlarge the credit bounds by the maximum such delay.

Regarding the service rate, required by a multiplexer which serves several independent streams shaped according to this algorithm, the following theorem was proven in [33]:

Theorem 2: *Let a number N of independent data streams be multiplexed, after being shaped by the spacing law (6.10) or through a token bucket mechanism fulfilling (6.11). If the multiplexer's service rate fulfils* $C=\sum_{i=1}^{N} f_i$, *its queue-length distribution follows (6.8) with* $q_o \le q$, *where q is the one used by the shapers; equality holds in a maximum utilization scenario, i.e. when the shapers have always data to transmit and no saturation of c at* c^+ *occurs*

In addition to the above theorem, we formulate and prove two theorems with regard to the aggregation and splitting of shaped streams.

1.3.2 Aggregation of shaped streams

Multiplexing of many, independently shaped streams, results in a behaviour determined by *q*, being the target objective of the shaping operation. In the sequel, it is shown that the aggregate stream (i.e. the one resulting from the superposition of the many individually shaped streams) is conformant with the shaping definition. This is an important result, since it allows for further handling the aggregate streams as atomic ones e.g., for conformance testing, policing and further multiplexing.

Theorem 3: Suppose that streams #1 and #2 are shaped according to (6.11) with the sets of parameters $\{q, f_1, [c_1^-, c_1^+]\}$ and $\{q, f_2, [c_2^-, c_2^+]\}$, respectively. Then the aggregate stream #1+#2 is conformant (shaped) with the parameters: $\{q, f_1+f_2, [\min\{c_1^-, c_2^-\}, \max\{c_1^+, c_2^+\}]\}$

Proof: Since streams #1 and #2 are shaped, it is

$$\sum_n \frac{1-e^{-qv_1^{(n)}}}{q} \le (T+\tau_1)f_1,\ \tau_1 \in [c_1^-, c_1^+] \text{ and } \sum_n \frac{1-e^{-qv_2^{(n)}}}{q} \le (T+\tau_2)f_2,\ \tau_2 \in [c_2^-, c_2^+]$$

Summing the above inequalities side by side gives

$$\sum_n \frac{1-e^{-qv_1^{(n)}}}{q} \le T(f_1+f_2)+\tau_1 f_1+\tau_2 f_2 \le (T+\tau)(f_1+f_2),$$

$$\tau \in [\min\{c_1^-, c_2^-\}, \max\{c_1^+, c_2^+\}],$$

where now the summation in the left-hand side includes the DUs from both streams. This proves the conformance of the aggregate stream, #1+#2, with the shaping algorithm, and it can be directly extended to the aggregation of any number of streams. ■

1.3.3 Splitting of shaped streams

Another important operation is the one of splitting of a single stream into a number of conformant substreams. If the initial stream is a composite one made up of shaped streams and the splitting respects the individual stream boundaries (flow-based, as is the case of aggregation and segregation of ATM streams), then the resultant substreams being themselves aggregates of shaped streams, are still shaped. The question is what will happen in the case of DU-based splitting, as in the case of datagram routing. The answer is given by the following theorem.

Theorem 4: *Suppose that stream #1 is shaped according to (6.11) with the set of parameters* $\{q, f_1, [c_1^-, c_1^+]\}$ *and a substream is extractedfrom it, which is also shaped with the parameters* $\{q, f_2, [c_2^-, c_2^+]\}$. *Then the remaining stream #1-#2 is conformant (shaped) with the parameters*

$$\{q, f_1 - f_2, [\min\{c_1^-, c_2^-\}, \max\{c_1^+, c_2^+\}]\}$$

*.**Proof:*** Similarly as for theorem 3. ■

The splitting algorithm can be realized quite easily through maintaining the credits of the two substreams and allocating each new DU to the one with the largest credit. As a result, both credits remain close to each other and within the bounds applied to the initial stream.

1.4 APPLICABILITY TO ATM

The *M/G/1* modelling considered in the previous sections assume (i) burst (or packet) arrivals on the continuous time domain, (ii) burst size drawn also from a continuous domain and (iii) batch entry (i.e. with an infinite rate) into the multiplexer's buffer. None of these assumptions are fulfilled in the case of ATM, being defined on a discrete (cell-slotted) time lattice, having bursts of integer multiples of cells and featuring a finite cell rate , less than or equal to the speed of the link. How good, then, can the proposed models be for ATM traffic?

With respect to the continuous time approximation, one may think that typical affordable delays, being of the order of tens of msec, correspond to hundred or thousands of ATM slots on a 155.52 Mbps link. On these figures half a slot maximum discrepancy between continuous and slotted time is of no practical significance. A similar argument can be stated for the discrepancy between continuous and quantized burst size.

The third approximation, namely the one related to the rate of the entry of bursts in the buffers, is more essential. According to the discussion exposed in 2.5.4, however, it provides safe upper bounds to the desired delay figures, overcomes the problem of rate heterogeneity along an end-to-end connection, while it exploits the buffering gain and allows the use of additive effective rates.

In [32], applications of the *M/D/1* model to ATM traffic control are presented, like controlling the cell-level congestion, controlling the CDV of variable-bit-rate traffic or the CDV of CBR traffic modulated by call-level dynamics, and using adaptable *sustainable cell rate* (*SCR*) - *burst tolerance* (*BT*) pairs to preserve desired effective rates. In the same paper, the burst spacing mechanism of 3.3.1 was derived starting from an *M/D/1* analysis. Here we focus on the formulation and implementation of the burst spacer as a *Generic Effective Rate Algorithm* (in analogy to the Generic Cell Rate Algorithm -GCRA- defined by the standards) to be used both for shaping and conformance testing of ATM connections. As far as the rest of the control functions are concerned with (e.g. admission control, bandwidth allocation, ABR-type functions) these can be based on effective rates, instead of peak rates, with no essential modifications.

From what exposed in the previous subsections, it is deduced that the parameter vector $[T_e = 1/f\ ,\ q\ ,\ v_{max}]$ can be used in traffic contracts for effective-rate-shaped connections. A Generic Effective Rate Algorithm - GERA- for the enforcement or monitoring of these parameters, is given in the Appendix, with the inclusion of a fourth parameter, τ, which is used as a normalised (with respect to the burst size, v) delay tolerance, in analogy to the CDVT used in conformance testing of CBR connections.

Fig. 6.7 explains the GERA mechanism. With the arrival of the first cell in the new burst, conformance is checked through comparing its actual arrival time with the Theoretical Arrival Time (*TAT*) of the new burst reduced by the delay tolerance, $v\tau$. If the cell is conforming, cells within the newly started burst are allowed to pass, until the maximum burst size, v_{max},is reached or until no other cell arrives within the time window of the current burst (started at *TAT* and increased with the arrival of every new cell by the increment T_e-τ). Then the new *TAT* is calculated using the burst spacing law.

Obviously, for v_{max}=1, the GERA reduces to the GCRA as it is described in the ITU-T standards [13] and the ATM-FORUM specifications [2], i.e.

GERA(T_e, q, 1, τ) ≡GCRA(T_e, τ).

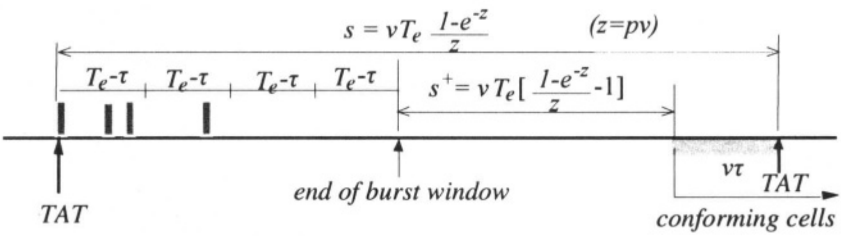


Figure 6.7: Explanations on the GERA algorithm

We conclude this section with a comparison between GERA(1/f, q, v_{max}, τ) and GCRA(1/*SCR*, *BT*) used to control nrt-VBR traffic.

GCRA(1/*SCR*,*BT*)	GERA(1/f, q, v_{max}, τ)
constant mean rate = *SCR*	variable mean rate (≤ f)
variable multiplexing behavior	constant effective rate

4. DISCUSSION

ATM traffic control is a hot topic, since it will actually determine the degree of success of ATM and whether it will eventually dominate or not over its competitors from the Internet world, namely the frame relay, the forthcoming IP WAN with the real-time-service support and the Gigabit Ethernet [25]. Simplicity and efficiency are the two, somehow conflicting, properties that are desirable by the traffic control functions to have. Towards this end, using only two ATM layer services out of the 4 or 5 currently standardized, namely, one for real-time applications and another one for non-real-time ones, could be a judicious choice. The selected services, however, should set and guarantee delay QoS figures, thus any further service differentiation in the ATM layer to be possible through setting different QoS levels.

It has been demonstrated that statistical models are not very suitable for traffic control, due to the unrealistic assumptions most of them are based upon, namely, the stationarity and the Markovian behavior of the traffic sources. On the contrary, control implies a more or less deterministic behavior, while stationarity of sources is only an exception and difficult to enforce, if absent. In this context, the new burst-level approach proposed in this paper may offer a solution. It fully exploits the buffering gain that is potentially available through using large buffers, while the simple *M/G/1* model can become the vehicle towards a simple and efficient traffic control framework. In this framework, the traffic descriptors are expressed in terms of the asymptotic effective rate associated with the *M/G/1* model, which is additive, while the QoS is expressed in terms of delay percentiles. Basic traffic control mechanisms, implementing shaping for effective rate enforcement, conformance testing, traffic aggregation & splitting, have been established.

Finally, it should be highlighted that the burst-wise approach fits better the functionality of the higher protocol layers and the applications themselves, most of them exchanging data packet-wise rather than rate-wise. In that sense, the burst-wise approach brings ATM closer to IP. From the other side, the introduction of IP flows with some QoS guarantee, closes even further the gap between ATM and IP and may help the two worlds to converge smoothly instead of competing to death.

Appendix

Generic Effective Rate Algorithm (GERA)

The GERA is proposed as a conformance definition of EfR-shaped bursty streams with respect to the traffic contract. As with the GCRA, for each cell arrival GERA determines whether the cell conforms to the traffic contract of the connection. The GERA is defined with four parameters: The *increment*, T_e, the *multiplexing quality slope, q,* the *maximum burst size,* v_{max}, and the *delay tolerance,* τ. It realization in pseudocode is shown next.

```
v:=0; TAT:=0;  t(k):=0;      /*  intialisation    */
do                           /* main loop         */
/*  wait for the first cell in the burst  */
 until(new_cell)
 do
    t(k)  := t(k)+1;
 loop
/*  check conformance  */
 $t_a(k)$  := t(k);
 if  ($t_a(k)$<TAT-vi) → non conforming cell;
 if ($t_a$ (k)>TAT)     TAT := $t_a$ (k);
 v := 1;
/*  accept new cells until end of burst window */
 while (TAT+v(Te-r)  >t(k)  &  v<$v_{max}$)
 do
    if(new_cell)  v := v+1;
    t(k)  :=  t(k)+1;
 loop
 /* calculate new TAT of next burst*/
 if(v=1)  s := $T_e$;
 else   z := qv;  s := [v$T_e$ (1-e-z) / z|;
 endif
 TAT := TAT+s;
/*  repeat forever   */
forever
```

References

[I]Anick D., Mitra D. and Sondhi M.M., (1982) "Stochastic theory of a data-handling system with multiple sources", *Bell Syst. Tech. J.* 61, 1871-1894.

[2] ATM Forum (1996), "Traffic Management Specification," Ver. 4.0, April 1996.

[3] ATM Forum (1998), BTD-TM-01.03, Traffic Management Working Group Baseline Text Document. John Kenney, ed. October, 1998.

[4]Bolotin V.,. Kappel J and Kuehn P. (1991) *eds,* Special Issue on :Teletraffic Analysis of Communications Systems," *IEEE Journal on Sel. Areas in Commun.*, Vol. 9, No. 2,4.

[5] Burgin J. and Dorman D. (1991), "B-ISDN Resource Management: The Role of Virtual Paths", *IEEE Comm. Mag.,* Vol. 29, No. 9, pp. 44-48.

[6] Chang C.-S. and Thomas J.A. (1995), "Effective bandwidth in high-speed digital networks," *IEEE JSAC,* Vol. 13, No. 6, pp.1091-1100.

[7] Cox D.R. (1962), *Renewal Theory*, Methuen & Co, London.

[8] Elwalid A.I., Mitra D. and Stern T.E. (1991), "Statistical Multiplexing of Markov Modulated Sources: Theory and Computational Algorithms", ITC *'91,* Copengagen.

[9] Elwalid A. and Mitra D. (1993), "Effective bandwidth of general Markovian sources and admission control of high speed networks,"*IEEE/ACM Trans. on Net.*, pp. 329-343.

[10] Gibbens R.J. and Hunt P.J. (1991), "Effective bandwidths for the multi-type UAS channel," *Queueing Sys.*, 9, pp. 17-28.

[11] Guerin R., Ahmadi H. and Naghshineh M. (1991), "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE JSAC,* 9, pp. 968-981.

[12] Hsu I. and Warland J. (1996), “Admission control for multi-class ATM traffic with overflow constraints,” *Comp. Net. and ISDN Sys.,* Vol. 28, pp. 1739-1751.

[13] ITU-T (1996), Recommendation 1.371, “Traffic control and congestion control in B-ISDN”, ITU-T Study Group 13, Geneva, April 1996.

[14] Jain R. (1996), “Congestion control and traffic management in ATM networks: Recent advances and a survey,” *Comp. Net. And ISDN Sys.,* Vol. 28, pp. 1723-1738.

[15] Kelly F.P. (1991), "Effective bandwidths at multi-type queues," *Queueing Syst.,* vol. 9, pp. 5-15.

[16] Kesidis G., Warland J. and Chang C.-S. (1993), "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources", *IEEE/ACM Trans. on Networking*, Vol. 1, No 4, pp. 424-428.

[17] Kingman J. (1970), “Inequalities in the theory of queues,” *J. Roy. Statist. Soc.,* Ser. B, Vol. 32, pp. 102-110.

[18] Kleinrock L. (1975), *Queueing Systems,* vol. II, ch. 2, New York, John Wiley.

[19] Kontovasilis K.P. and Mitrou N.M. (1994), "Bursty Traffic Modelling and Efficient Analysis Algorithms via Fluid-Flow Models for ATM IBCN,"*Annals of Operations Research*, Vol. 49, spec, issue on Methodologies for High Speed Networks, pp. 279-323.

[20] Kontovasilis K.P. and Mitrou N.M. (1995),"Markov Modulated Traffic with Near Complete Decomposability Characteristics and Associated Fluid Queueing Models," *Applied Probability Journal* Vol. 27, No. 4, Dec. 1995, pp. 1144-1185.

[21] Kontovasilis K.P. and Mitrou N.M. (1997), “Effective Bandwidths for a Class of Non Markovian Fluid Sources”, *ACM SIGCOMM’97 Conf,* 14-18 Sept. 1997, Cannes France.

[22] Kosten L. (1984), "Stochastic theory of data-handling systems with groups of multiple sources," in *Performance of Computer Communication Systems,* H. Rudin and W. Bux, Eds. Amsterdam, The Netherlands: Elsevier, pp. 321-331.

[23] Kroner H., Hebutern G., Boyer P. and Gravey A. (1991), "Priority Management in ATM Switching Nodes," *IEEE J-SAC,* Vol. 9, No.3, pp. 418-427.

[24] Kuehn P., Lehnert R. and Gallassi G. (1994), *eds,* Special Issue on "Teletraffic Research for Broadband-ISDN in the Race Programme", *European Transactions on Telecommunications,* Vol. 5, No. 2.

[25] Mace S. (1997), “ATM’s Shrinking Role”, *Byte,* Oct. 1997, pp. 59-62.

[26] Mitrou N.M., Vamvakos S., Kontovasilis K. (1995), "Modelling, Parameter Assessment and Multiplexing Analysis of Bursty Sources with Hyper-exponentially Distributed Bursts *Computer Networks and ISDN Systems*, Vol. 27, pp. 1175-1192.

[27] Mitrou N.M., Kontovasilis K. and Nellas V. (1994), "Bursty Traffic Modelling and Multiplexing Performance Analysis in ATM Networks: A Three-moment Approach," *2nd IFIP Intern. Workshop on on Performance Modelling and Evaluation ofATM Networks,* Bradford, 4-6 June 1994.

[28] Mitrou N..M., Kontovasilis K.P., Kroener H. amd Iversen V.B. (1994), "Statistical Multiplexing, Bandwidth Allocation Strategies and Connection Admission Control in ATM Networks,"*European Trans. on Telecommunications,* Vol. 5, No. 2, pp. 161-175.

[29] Mitrou N.M., Kontovasilis K. and Protonotarios E.N. (1995), "A closed-form expression for the effective rate of On/Off traffic streams and its usage in basic ATM traffic control problems," *Proc. Int. Teletrqffic Seminar,* St. Petersburg, pp. 423-430.

[30] Mitrou N.M., Lykouropoulos N., Nellas V. and Kontovasilis K. (1996), "Experimental Validation of Selected Results on ATM Statistical Multiplexing in the EXPLOIT Project," *European Trans. on Telecommunications,* Vol. 7, No. 5, pp. 423-431.

[31] Mitrou N.M. and Kontovasilis K. (1996), "Comparison of Three Control Laws for Statistical Multiplexing in ATM," *Journal on Communications,* Vol. XLVIII, Jan.-Feb. 1996, pp. 24-29.

[32] Mitrou N.M. and Kavidopoulos K. (1998), "Traffic Engineering using a class of M/G/1 models," *J. of Net. & Comp. Appl.* Vol. 21, pp. 239-271.

[33] Mitrou N.M. (1999), "Shaping of Traffic Streams through Data Spacing", *IEEE Communications Letters,* Vol. 3, No. 10.

[34] Neuts M. (1979), "A Versatile Markovian Point Process," *J. Appl. Prob.,* 16, 764-779.

[35] Ramaswami V. (1980), "The *N/G/1* Queue and its Detailed Analysis," *Adv. Appl. Prob.* Vol. 12 , pp. 222-261.

[36] Stern T.E. and Elwalid A.I. (1991), "Analysis of separable Markov-modulated rate models for information-handling systems", *Advances in Applied Probability*, vol. 23, pp. 105-139.

[37] Turner J. (1986), "Design of an Integrated Services Packet Network," *IEEE Journal on Select. Areas in Commun.,* Vol. 4, No 8, pp. 1373-1380.

[38] De Veciana C., Kesidis G.and Warland J. (1995), "Resource management in wide-area ATM networks using effective bandwidths," *IEEE JSAC,* Vol. 13, No. 6,pp.l081-1090.

[39] Xiong Y., Bruneel H. and Petit G. (1993), "Performance study of an ATM self-routing multistage switch with bursty traffic: simulation and analytic approximation", *Europ. Trans. on Telecomm. (ETT),* vol. 4, no. 4, July-August 1993, pp. 443-453.

[40] Yang C.-Q. and Reddy A.V.S. (1995), "A taxonomy for Congestion Control Algorithms in Packet Switching Networks," *IEEE Network,* July/August 1995, pp. 34-45.

Nikolas.M.Mitrou was born in Greece on October 5, 1957. He received the undergraduate Diploma degree in electrical engineering from the National Technical University of Athens (NTUA) in 1980, the MSc degree in Systems and Control from the UMIST, Manchester, in 1982 and the PhD degree in electrical engineering from NTUA in 1986. From 1982 to 1985 he was with the Nuclear Research Centre "Demokritos", Athens, where he was involved in signal processing projects. From 1986 to 1988 he worked at the National Defence Research Centre, Athens, for the development of a low-bit-rate voice coding system. In 1988 he joined the NTUA as a senior researcher, where he is currently full professor in the Department of Electrical and Computer Engineering. His research interests are in the areas of digital communication systems and signal processing, with emphasis on the architecture, modeling, performance evaluation and optimization of integrated networks, local area networks and mobile communication systems, digital video and multimedia, having more than 60 publications in the above fields.

Prof. Mitrou has had a leading participation in many research projects (RACE, ESPRIT, ACTS): RACE 1022 “Technology for ATD”, RACE 1043 “Mobile Telecommunication Project”, RACE 2061 “EXPLOIT”, RACE 2064 “FLASH-TV”, ACTS AC094 “EXPERT”, ACTS AC235 “WATT”. In EXPLOIT and EXPERT he was the leader of a Workpaclage devoted to Network Resource Management and Routing issues in ATM. He was also the coortdinator of the WATT project, dealing with the remote monitoring of broadband networking experiments using Web technologies.

Prof. Mitrou is a member of the IEEE and the Technical Chamber of Greece and member of the IFIP WG 6.4.

Chapter 7

VIDEO OVER ATM NETWORKS

Gunnar Karlsson
Department of Teleinformatics
Royal Institute of Technology (KTH)
Electrum 204, 164 40 Kista
Sweden
gk@it.kth.se

Abstract There will be quality requirements posed on network transfers of video information. They stem from the timing of the temporal sampling, and from perceptual limits on delay and loss that are given by the use of the transfers. Loss and delay exceeding these limits impede the information exchange between users and may render a service useless. This paper is a review of video communication over ATM networks with overviews of source coding, bit rate regulation, quality requirements, traffic characterization and ATM traffic classes.

Keywords Video coding, quality of service, bitrate control, asynchronous transfer mode

1. INTRODUCTION

One may place three general types of requirements on a network service: connectivity, quality and cost. This means that the network should connect to the parties (both people and service points) that the user may want to reach and it should offer a quality of service that is adequate for the communication at an acceptable cost. If we concentrate on quality issues, then the communication system has to limit information loss and delay according to requirements placed by the application and the information type. Video, which is the focus of this survey, may be transferred over a network for instantaneous viewing and requires timely delivery to be displayed in a perceptually continuous sequence. This requirement stems from the isochronal sampling that has to be maintained at the receiver to avoid disturbing aliasing effects (spectral distortion). There might also be bounds on the maximal transfer delay from camera to display monitor when the video is part of an interactive conversation or conference.

The information will not only be delayed in the network since errors and loss may also occur. Absolute delivery is not required for video; perceptual–quality guarantees are sufficient. For instance, information loss not detectable by the eye, cannot be considered to degrade the perceptual quality of an image in a video sequence. Yet, the absolute delivery of the information has been compromised. Perceptual criteria are relevant for all types of digitized information since the amplitude quantization following the sampling already has distorted the signal.

This survey is intended to give an overview of transfer and quality aspects for video. The contents are as follows. We describe the video signal and the video–specific functions of the communication system in Section 2. The perceptual limits in terms of delay and loss are discussed in Sections 3. The characteristics of video traffic is considered in Section 4, owing to its importance for choosing multiplexing mode. The network support is covered in Section 5 and deterministic and statistical multiplexing are distinguished for giving service–quality guarantees. Section 6 discusses the matching of video applications to the network service classes. The Chapter is summarized in Section 7.

The material in this Chapter has in parts been published previously in [46], [47] and [48].

2. DIGITAL VIDEO

Video in digital form is a three dimensional signal: it is a time sequence of equidistantly spaced two dimensional frames. The frames can be samples of a real scene captured by a camera, or they may be generated by computer graphics. The structure of a video sequence is illustrated in Figure 7.1. The frames in the sequence may be decomposed into two fields if interlaced scanning is used. The fields are composed of the even and the odd lines of pixels, respectively. Each pixel consists of three color components with a fixed number of bits (normally 8 bits per component). (See [66] for details.)

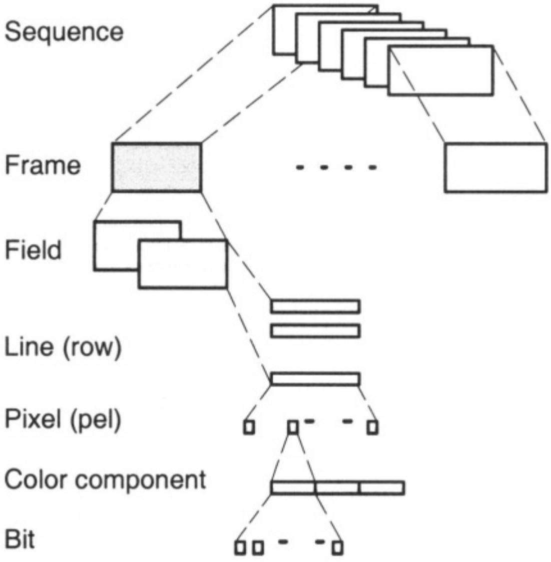


Figure 7.1 Structure of a video signal

The ISO/IEC Moving Picture Expert Group (MPEG) has defined some other structural elements of a video signal which, owing to the importance of their video coding standards, are worth mentioning. A *macroblock* consists of a square with 16 × 16 luminance samples together with one 8 × 8 block of samples for each chrominance component (this assumes a *4:2:0* format for which the chrominance components have half of the luminance component's sample rate in each dimension). A stretch of consecutive macroblocks in the scanning order together with a header is called a *slice* [65] [83].

The sender side of a video communication end–system is shown in Figure 7.2 and will be briefly outlined in this section. The receiver side performs functions which are the reciprocal of the sending functions and may compensate for imperfections during the transfer. These functions include source decoding, error handling, delay equalization, and clock synchronization.

2.1. SOURCE CODING

As Figure 7.2 shows, the digitized video is first passed to a source encoder. It is often built with three system system components: energy compaction, quantization and entropy coding.

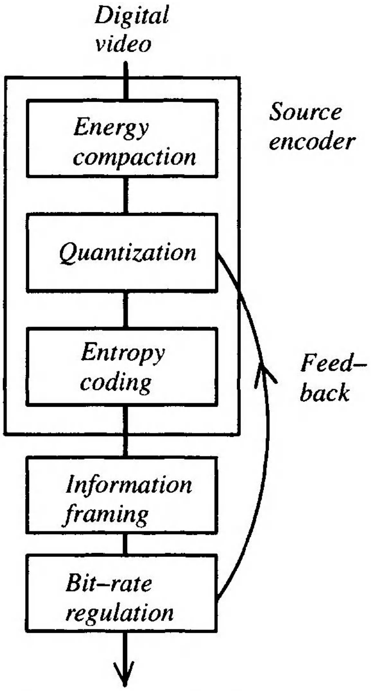


Figure 7.2 The sending side of a video system.

The energy compaction aims at putting the signal into the form most amenable to coarse quantization. Common methods for video include discrete cosine transform, subband and wavelet analysis [87], and prediction, possibly motion estimated. The quantizer reduces the number of permissible amplitude values of the compacted signal and introduces thereby round–off errors. The entropy coding, simply put, assigns a new representation to the signal which gives shorter code words using fewer bits to frequent sample values and longer code words with more bits to infrequent values. It represent the signal more efficiently but there is no longer a constant number of bits per sample, and the bit rate is therefore temporally varying (see [66] for an introduction to video coding).

MPEG-2 overview

The coding scheme of the ISO/IEC Moving Picture Experts Group warrants a short introduction [65][83]. The MPEG-2 video coding standard (ISO/IEC 13818–2:1996) has been developed jointly with ITU–T SG15 as recommendation H.262 to be used for video communication (MPEG–1 is intended for storage media, notably CD–ROM). It uses two forms of energy compaction: prediction is used to exploit temporal redundancy, and discrete–cosine transformation is applied to compact the signal spatially (it is applied to 8 × 8 blocks). The transformed blocks are quantized and the coefficients are then run–length and Huffman encoded.

There are two types of prediction, of which both are motion compensated. The first type, denoted P, is unidirectional and tries to estimate a macroblock in a frame by matching it to areas of 16 × 16 in a prior frame. The difference values between the macroblock and the best matching area are transformed, entropy encoded and transferred along with a pointer to the area used (a so called *motion vector*).

The second type of prediction is bi–directional (B). A macroblock in a frame is matched to areas in past and future frames. The prediction is delayed until the future frame is present. The bi–directional prediction is more accurate than P prediction and consequently gives a very high degree of compaction. There is also an intra coding mode (I) for which a macroblock is transformed without using prediction. Such macroblocks serve as re–start points for the prediction. All three modes of energy compaction can be used intermixed.

If a picture contains only intracoded macroblocks, then it is denoted I picture, a P picture may mix intracoded and forward–predicted macroblocks, and finally a B picture may encode macroblocks in any of the I, P and B modes. I pictures will not be used in a sequence that is meant to be transferred, with the exception of the very first picture. P pictures and possibly also B pictures will be used—delay permitting—for which a suitable number of macroblocks are intracoded to provide refresh of the prediction. The reason, of course, is to avoid the surges in bit rate that a full I picture would cause. A smoother bit rate can thus be achieved by intracoding only parts of each picture, without compromising error resilience.

It is worth stressing that MPEG–1, which encodes full frames in the I, P and B modes in a periodic pattern, is not suited, nor designed for real-time communication [49].

Layered source coding

Layered (or, hierarchical) coding means that the signal is separated into components with differing visual importance [6] [19][42]. The layers are formed from the signal components after the energy compaction, and they are separately quantized, entropy coded and framed. The idea is based on the realization that all quanta of rate (ΔR) allocated to a source do not yield the same amount of reduction in distortion, as illustrated in Figure 7.3. However, if all allocations are within the same service class then it is reasonable to assume that the cost increases linearly with the allocated rate (denoted by curve *(a)*). By letting each quantum of rate being a layer and by matching service class to the importance of the layer, it would be possible to reduce the cost of the session, as shown by curve *(b)*.Thus, the idea is that each layer can be transferred in the network with a quality and cost that matches its importance, in order to keep the relation between cost and utility constant.

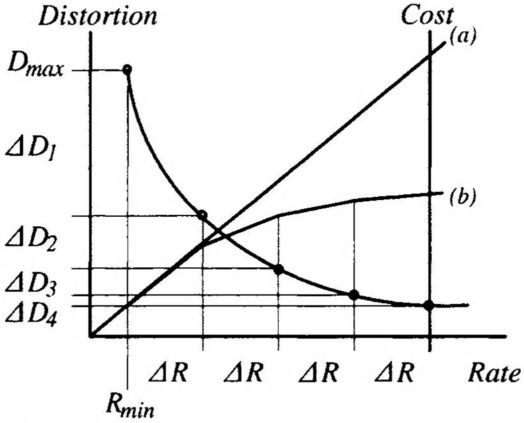


Figure 7.3 Illustration of layering: increases in rate above a needed minimum (R_{min}) are treated as separate layers. The layers do not give equal reduction in distortion.

Vital layers may thus be transferred in a class with guaranteed quality, while a signal layer that enhances the quality could be sent at "best effort". The hope is that the overall transfer is more economical than if the transfer was done over one channel with a service quality determined by the most sensitive part of the information. Layering therefore assumes that a set of connections of differing capacity and quality of service is cheaper than one connection for the aggregate stream. This can be firmly established first when tariff structures for broadband networks are in place, however.

Layered coding is also useful when a specific target bit rate or quality level cannot be stated *a priori* for the transfer. By layering, the sender can provide a range of bit rates and qualities in one and the same encoding of the information, and the particular point in that range can be chosen dynamically. It can, for instance, be beneficially used for stored programs so that rate control can be exercised when the video is being retrieved: the server can add and drop layers to fit a given channel for the transfer. Layered encoding usually require a higher bitrate for a given quality compared to a single–layer encoding. Also, layered transfers require inter-layer synchronization at the decoder.

2.2. INFORMATION FRAMING

Before or after the bit–rate regulation is the information segmentation with application framing. A frame is a segment of data with added control information.

Segments that are formed at the application level typically constitute the loss unit: errors and loss in the network lead to the loss of one or more application segments. Further segmentation occurs at the adaptation layer where the data are segmented into the multiplexing units, ATM cells, which are the smallest loss-units for the network. The application layer segmentation and framing should simplify the handling of information loss that may occur during the transfer. The network framing is needed to detect and possibly correct bit and burst errors as well as cell losses.

The ITU–T recommendation for ATM network adaptation is the H.222.1, which builds on the more generic H.222.0 | ISO/IEC 13818–1 (MPEG–2 Systems), and it is part of the system recommendation H.310 [68]. H.222.1 includes both adaptation layers 1 and 5 (ITU–T recommendations I.363.1 and I.363.5) for constant–rate transfers; there is currently no recommendation for variable-rate transfers. The formats of the protocol framing information for these layers are shown in Figure 7.4.

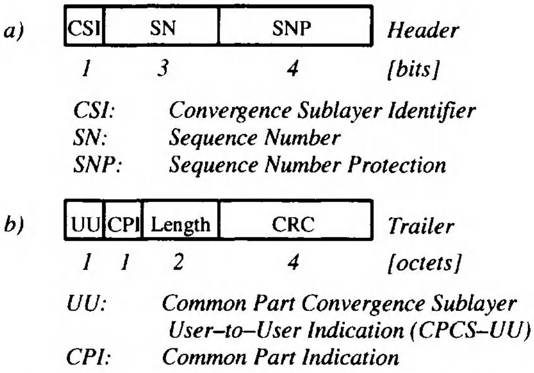

a)
CSI
SN
SNP
Header
1
3
4
[bits]
CSI: Convergence Sublayer Identifier
SN: Sequence Number
SNP: Sequence Number Protection
b)
UU
CPI
Length
CRC
Trailer
1
1
2
4
[octets]
UU: Common Part Convergence Sublayer User–to–User Indication (CPCS–UU)
CPI: Common Part Indication

Figure 7.4 Formats for control information in ATM adaptation layers 1 (a) and 5 (b).

Layer 1 is aimed at constant–rate real–time services and does not support transfer of partially filled cells, as needed for variable bit rate. The sequence number has three bits and it is protected by a four–bit check sum with the possibility of correcting a single bit–error (although such a rare event could be treated as cell loss). Detected cell losses become erasures since all cells are filled to the same level (not necessarily 47 octets when using the, so called, the P–format).

The CS–indication bit may be used to send residual time–stamps, which may be used for synchronization of the sender and receiver clocks in relation to the network clock. The synchronization would be used to remove jitter and to regain the original pacing of the stream. Forward error–correction with Reed–Solomon (124, 128) coding may be used with interleaving to correct four or less erasures. The interleaving creates 128 cells delay and the coding adds 3 percent load.

Layer 5 is frame-based to suit data–communication. Each frame is protected by a 32–bit CRC for error detection. A length-field gives the amount of user–data in the protocol data–unit. The convergence sublayer has to collect a specific amount of octets to send as a frame, if data is delivered in a stream. The longer the frame, the lower is the overhead but the higher the delay at the receiver (where the full frame has to be reassembled for verification by the CRC). There is no way of telling how much user–data has been effected once a bit or burst error, or cell loss has been detected since the length filed cannot be trusted (unless all frames are of the same length).

A light–weight adaptation layer with framing and loss detection for arbitrarily–sized SAR–PDUs has been proposed for variable and constant rate video in [44].

2.3. BIT-RATE REGULATION

The bit-rate regulation is used to adapt the varying output rate of the coder to the channel in the network. It is needed when there are restrictions on the permissible bit rate from the coder. An obvious restriction is that of the access link's capacity to the network. The regulation may flatten the bit–rate variations by buffering and may regulate the compression to avoid overflow. The feedback reaches the quantization of the encoder and enforces a higher step–size with increased round-off error as a consequence. If the quantizer step–size is throttled frequently and heavily it may lead to visible quality fluctuations in the reconstructed signal.

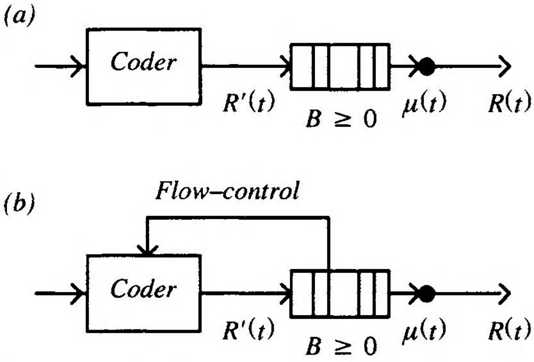


Figure 7.5 The sender is contractually obliged to restrict its traffic by a function $\mu(t)$. There are two cases to consider: without feedback (a), and with feedback (b).

We will assume that a traffic contract is established between the sender and the network. The contract stipulates the allowed behavior of the cell stream admitted into the network (denoted by $\mu(t)$ in Figure 7.5) as well as the service the cells will receive in the network. The following modes of operation will be considered:

- Unregulated coding without or with output buffering (Figure 7.5 *a*).
- Feedback-regulated coding with or without output buffering (Figure 7.5 *b*).

The output rate from the sender is given by $R(t) = \min\{R'(t), \mu(t)\}$. When $R'(t) > \mu(t)$, case *(a)* gives cell loss, and possibly also delay if a smoothing buffer is present, and case *(b)* gives delay and quantization loss. The advantage of case *(b)* over case *(a)* is that the quantization loss can be made less perceptual than the cell loss. This, in turn, means that more loss can be accepted in order to reduce the bit rate.

3. QUALITY REQUIREMENTS

3.1. DELAY TOLERANCE

The first quality parameter we consider is delay. The isochronal sampling of the signal imposes requirements on regular pacing of the signal at the digital to analog conversion, as mentioned before. Delay-variations (jitter) caused by the asynchronous multiplexing and protocol processing are commonly handled by enforcing a delay limit, D_L, in accordance with guidelines given for the total transfer delay (discussed below). All data are then delayed up to this limit (see Figure 7.6). The variations in delay are thus of little concern when the maximum delay in the network is below the tolerable end–to–end limit. When this is not the case, arrivals with delay above the limit are discarded at their reception and excessive delay is hence turned into loss. For a given delay distribution there is consequently a balance between the amount of delay that is accepted and the probability of loss. User preferences in this balance have not been reported. Early arrivals, $D < D_L$, are queued at the receiver and each cell is read out when it has been delayed exactly D_L seconds. The needed buffer size for lossless operation is therefore $D_L - D_{min}$ seconds.

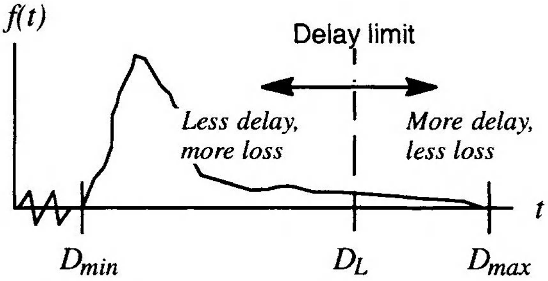


Figure 7.6 The distribution of delays, f(t), with an illustration of the trade–off between delay and loss.

A receiver could be designed to reconstruct the signal piece by piece, in the pace that the data is delivered. This means, however, that the signal is aliased since the sampling intervals are not necessarily maintained. Instead of, for instance, the fixed 40 ms per frame used by European video formats, the video–frame duration would vary. The artefact in this case is a poorer rendition of motion, since smooth movements could appear as a sequence of jolts. There are no subjective test reported on the interval of frame durations that can be deemed acceptable for various applications.

The applications may place delay constraints on the transfer when it is bidirectional and supports an interactive conversation or conference (unidirectional transfers, like broadcast television, do not pose any delay restrictions in general). The recommendations on maximal delay follow those for telephony (ITU–T Rec. G.114: Mean one–way propagation time) [54][55]. There is consequently no or little impact below 150 ms one–way delay, and serious impact above 400 ms. Video may precede or succeed an associated audio stream by up to 80 ms for one–

way sessions without noticeable loss of lip–synchronization [86]. Interactive sessions show less clear limits due to differences caused by the type and content of the conversation [55]. Other types of media synchronizations, such as text annotations for video have looser limits, delay differences within ±240ms are clearly acceptable.

The delay is caused by execution of various protocol functions and by the transfer across the network. The following instances may cause the bulk of end–to–end delays:

- Source coding, rate–regulation and decoding
- Frame segmentation and re–assembly
- Protocol processing
- Channel coding (forward error correction) and decoding
- Wave propagation and transmission
- Queuing

There are two issues to control regarding delay: the variations must be equalized according to the discussion above, and the delay limit, D_L, for a specified probability of loss must be limited for interactive applications. The total delay can only be capped by careful design of all the functions from the information source to its sink. Most articles on delay aspects aim at reducing the queuing delay, but basically any of the instances listed above can yield delay that exceeds the acceptable level for interactive applications. Queuing delay is most effectively reduced by using deterministic rather than statistical multiplexing.

Equalization of delay variations is basically done by buffering data delivered by the network to a predetermined limit before further processing (see Figure 7.6). A common simplification is to equalize queuing delay at the re–assembly point, the adaptation layer in case of ATM [84]. Jitter introduced in the end system is then removed after the decoding to obtain synchronization with the digital to analog converter.

3.2. LOSS AND ERROR TOLERANCE

Most components of the communication system cause information loss: the bandwidth is limited to avoid aliasing in the sampling and the values are distorted when quantized. Source coding, when used, introduces further distortion albeit in controlled amounts and ways. Moreover, the signal will be exposed to bit–errors induced in the electronics and in the optics throughout the system. The probability of bit–error is low, below 10^{-8}, but not negligible. Application frames corrupted by bit errors are often discarded and hence turns errors into losses. Finally, cell loss in the network deletes full stretches of data. The causes are transmission burst–errors, multiplexing overload, and misrouting. The deletion of data can cause unlimited (catastrophic) error propagation if the framing information cannot determine the exact amount of lost data.

The perceptual quality of a reconstructed video signal at the receiver should be adequate for its intended use and whether perceptible loss is caused by source coding, overflow of multiplexing buffers, or transmission errors is irrelevant to a user. Most of the information loss objectively measured would be incurred at the source encoding when such is used. The coding is, however, done with awareness of the signal's information contents and possibly also its application context. The introduced distortion can therefore be made suitably imperceptible.

Acceptable levels for transfer loss and errors are difficult to determine since they depend on human perception of criteria such as the use and cost of the transfer, the duration of the session, the quality of the source signal, and the appearance of the loss and errors in the reconstructed signal [36][80]. Subjective tests can show quite large variations in the tolerances of individual persons taking part in the experiment.

For ATM, the target loss probability for video and audio is often taken to be 10^{-9}. This author has, however, not been able to find a proper justification for it in the research literature (it corresponds to one cell loss per hour for a STM–1 channel with net data rate 150 Mb/s). The network will be operated uneconomically if it turns out to be overly precautious. There are overall few results reported on the perceptual aspects of video transfers and there are consequently few guidelines to follow, especially concerning acceptable loss levels.

In reference [38] there are, however, some results that can be used: The factors that have greatest impact on quality are the number of lost packets, number of pixels in an impaired region and its shape, as well as the ”burstiness” of the loss. For the latter, random packet losses were found to yield greater quality degradation than clustered losses at equal loss ratios. Thus, for a given loss probability one may safely assume uncorrelated loss events which give an upper bound on the quality degradation.

Error recovery is based on limited error propagation, and correction or concealment of the affected portion of the data. Error propagation is restricted by proper framing of the data so that errors and loss can be detected. Correct reception may then be re-initiated at a later point in the data stream and the loss of the intermediate piece of signal is concealed. Cell loss concealment is an active research area and results reported include those in references [20][43][52][53][88][90]. Note that retransmission is usually disregarded as an error control technique for video transfers. First, the delay requirements might not allow it since it adds at least another round–trip delay that is likely to violate end–to–end delay requirements for conversational services. Second, the jitter introduced is much higher than that induced by queuing. Delay equalization is thus further complicated. Third, it complicates multicast and broadcast since the sender can be overwhelmed by retransmission requests (sometimes referred to as “ack implosion”).

4. TRAFFIC CHARACTERIZATION

Video transfers pose requirements on the network–service quality according to the discussion above. In order to provide adequate quality guarantees *a priori* the sender needs to describe the anticipated video traffic for the connection–acceptance control. The description is the basis that the network uses to allocate resources for the connection; the connection is blocked if the resources deemed necessary for providing the requested quality level are not available. The connection–acceptance control could benefit from using measurements of existing connections to supplement the user's descriptor for making the decision. However, the traffic parameters in the set–up message form the only piece of information about the characteristics of the upcoming session.

Another possibility than the *a priori* declaration is that the restriction on the traffic flow into the network is not only given by the initial sender–declared traffic descriptor but that it also is dynamically adjusted according to reports on the network state. There cannot be any clear guarantees stated for the transfer and the service is a "best effort" on behalf of the network. It is, however, clear that the service level can be sufficiently good for specific communication applications although it cannot be promised in advance. For instance, a given restriction on throughput in the network is better enforced in the encoder by coarser quantization than by cell discard in the network nodes. The feedback from the network would provide that possibility.

4.1. TEMPORAL FLUCTUATIONS

It might be useful to mention the factors behind the temporal fluctuations of a video stream. The encoded video bit–rate will depend on the resolution of the source signal and on the actual coding algorithm. These effects can be assessed reasonably well. The uncertainty about the behavior is more caused by the video contents and its many time scales of variation. This is illustrated in Figure 7.7.

The actual program will determine the long term rate which is in the order of minutes to hours. The different scenes within the program give the medium term rates and last in the order of seconds to minutes. It might be more difficult to predict the bit-rate at the scene level than at the program level. In fact, the concept of a scene is not even easy to define unambiguously (like the definition of a burst for data traffic). For a film it is a segment between two cuts, for live transfers it can be a view from one of a set of alternating cameras, or for a single camera it may be determined as a period of no motion, of panning or of zooming motion.

On a yet finer time scale we have the variations that are on the order of a frame time (*T* in Figure 7.7 (b) which is 40 ms for European formats). They stem from interframe and intraframe differences and account for the quickest variations in bit rate. Since many studies (such as [30]) have confirmed that it is always beneficial to smooth the cell sequence over the frame duration, there is no need to consider finer time scales.

The variations at the program level may be quite large, between 15 and 30 Mb/s are the mean rates reported in [33] for a given coder, and between 1 and 18 Mb/s in [30]. Leduc and Delogne show the program mean rate for fifteen television programs encoded at different quality levels with two different algorithms [60]. Their findings show variations of the program mean varies 30 to 50 percent around the ensemble average depending on coding algorithm. These variations decrease as the compression is increased.

For broadcast programs, Leduc and Delogne have found a bimodal distribution of the scene lengths and a bit rate distribution for the scenes that is approximately

Gaussian [60]. The span of mean rates for the scenes is 0.5 Mb/s to 3 Mb/s. The frame to frame variations may be on the order of five times the scene's or the program's mean rate.

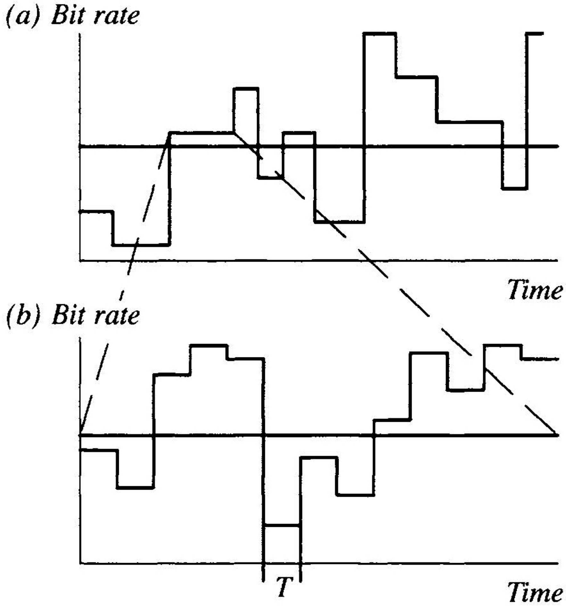


Figure 7.7 (a) The mean bit rate for each scene fluctuates around the program mean, and (b) the bit rate per frame is, in turn, varying around the mean rate of a given scene.

For a given coder we therefore find that the mean rate of a program can deviate from the ensemble average (for instance computed over all previous sessions) by a large factor. Individual scenes within the program may further vary from the program mean by a factor five, and individual frames within a scene may also give a bit rate five times the scene's mean rate. There is little side information that can be added to predict these fluctuations [74].

4.2. SENDER-DECLARED DESCRIPTORS

Two types to traffic descriptors are commonly used in the literature: 1) the time–varying rate from the coder is described by a suitable stochastic model, or 2) its envelope is given some specified bound. The bound could be stochastic or deterministic, although the latter is most common (see [56] for a description of stochastic bounding; it will not be considered further in this survey). One may consider the following points in choosing a specific type of descriptor:

- The source description should capture the traits of the source behavior that influence the multiplexing performance.
- The application (or user) should be able to reliably estimate the parameters of the description before the call has started.
- It should be possible to enforce the description at the source, and for the network to verify it by measurements (so called *policing*).

- The description should be appropriate for fast and accurate connection–acceptance control algorithms.

The specification of the bit stream will be enforced, as shown in Figure 7.5. The bit stream from the coder is fed into a buffer at a rate $R'(t)$ and it is served at some rate $\mu(t)$ so that the output bit rate $R(t)$ meets the specified behavior. Note that the quality of a session could be poorer than expected simply because an inappropriate description has been chosen and enforced by the bit–rate regulation. This could render the connection useless if it cannot be amended by re–negotiation of the connection–parameters.

4.3. DETERMINISTIC BOUNDING

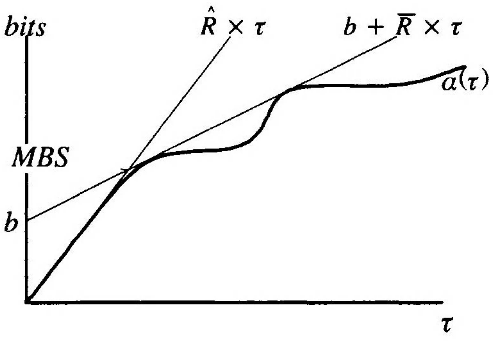


Figure 7.8 An arrival curve bound by two lines.

An *arrival curve* for the traffic from a source into the network is a plot of the maximum number of bits that can be generated during an interval of given length (see [11] [59]). Let $\Re(t)$ denote the number of bits sent by the source from time *0* to time *t*, then the minimum arrival curve is given by $a(\tau) = \sup_{t\geq 0} \Re(\tau + t) - \Re(t)$ (it is called an "empirical envelope" in [89]). The goal for the deterministic bounding is to find an enforceable arrival curve $a^*(\tau)$ that bounds the arrival curve from above as tightly as possible: $a(\tau) \leq a^*(\tau), \forall \tau \geq 0$.

The two most common bounds are the fixed–upper bound $a^*(\tau) = \hat{R} \times \tau$ and the dual leaky–bucket bound $a^*(\tau) = \min\left\{\hat{R} \times \tau,\ b + \bar{R} \times \tau\right\}$ (see Figure 7.8). The parameters $\hat{R}$and $\bar{R}$are usually referred to as peak and sustainable rates (in bits per second), respectively, and *b* is called the burst size (in bits). (The corresponding ATM traffic descriptor is the *maximum burst size* and it is equal to $b \times \hat{R}/\left(\hat{R} - \bar{R}\right)$, as marked in Figure 7.8) The arrival curve is, thus, bounded by one or two lines. One would expect a minimum of three lines in order to bound the arrival curve for the frame, scene and program time scales. More lines may in fact be used to provide a tighter concave hull for the arrival curve (with a bearing on the complexity, naturally) [78] [89]; for *M* lines, the bound becomes $a^*(\tau) = \min_{1\leq i\leq M}\{b_i + R_i \times \tau\}$, where $b_1 = 0,\ b_i < b_j$ and $R_i > R_{j,} \forall i < j$.

4.4. STOCHASTIC MODELING

The bit stream from an encoded video source could be seen as a realization generated by a suitable stochastic model. This has been a popular idea and there is an abundance of models in the literature that have been fitted to recorded traces of encoded movies, television programs, and conference scenes (as examples, see [1][28][32] [39][58][62][64][76]). Although many of the models are amenable for use in connection–admission control, few of them can be enforced and policed over reasonable time periods. Also, most models are fitted to the distribution of bits per encoded frame within a scene and consequently do not capture scene to scene variations.

Of the all video source models in the literature only one has, to the author's knowledge, considered bit–rate regulation. Heeke describes a model–based characterization that can be enforced without noticeably affecting the video quality [31]. The method forces the bit stream to obey a Markov chain model. The admissible rate will be restricted to a few levels with geometrically distributed holding times. Policing is possible by verifying the constant-rate levels and their average holding times.

Another issue is the selection of a model and its parameter before the session has started. It has been thought possible to tabulate this information. The table would provide the descriptor needed when requesting a connection for a specific application. The problem is however moved over to the connection-admission control which should be able to compute quality of service bounds for a mixture of model types and variety of parameter sets. It is also difficult to verify that it is the stated application, which uses the established connection. (The tabulated application with the highest performance to cost ratio might be declared by users who cheat the system for their own benefit.)

Due to the complications of the modeling approach, only deterministic bounding has been specified as traffic descriptor for connections in ATM.

4.5. REGULATION WITHOUT FEEDBACK CONTROL

One of the early argument used to promote asynchronous transfer of video was that the feedback–regulation of the coder would not be needed any longer since the bit rate could be allowed to fluctuate. The benefits would be a less complex encoder, less delay and a constant quality, since the transfer would use whatever capacity is needed for the unrestrained bit stream.

It should, however, be noted that the provision of consistent subjective quality allows bit–rate regulation. In fact, in [69] it is shown that a fixed quantizer does not result in an even quality. Sections of the video signal that are easy to encode should be more coarsely quantized than the average to maintain a uniform quality level. The fixed quantizer approach gives a known lower limit on the quality; the maximum can however be much better. Unless that minimum level correspond to imperceptive distortion, there will be noticeable fluctuations in the quality. Conversely, if the minimum level is adequate then there is an over–allocation of bits to sections of the signal that are coded at (unnecessarily) higher quality. In conclusion, the valid reasons for promoting unregulated encoding are reductions in delay and in complexity.

Rathgeb has studied various bounding techniques and found for the leaky–bucket that the burst size increased rapidly as the sustainable rate was brought down from the peak rate towards the average rate [72]. This is supported by the study of Heyman who shows plots of needed burst size as a function of sustainable rate for a specific video model. The curves are close to vertical for any given loss probability [35]. The conclusion is that a slight increase in allocated bitrate can effectively remove the need for a burst size and, consequently, a single upper bound is as adequate as the leaky–bucket bound.

Given a (single) fixed rate as traffic descriptor, the remaining question is how to choose the specific rate before the session has commenced so that the enforcement loss is acceptable. This issue is the same for smoothed but unregulated coding, for which delay is introduced to lower the declared rate (Figure 7.5 (b)). One solution is to calculate the equivalent capacity of the source and use that as the declared rate.

The equivalent capacity (also called effective bandwidth or bit rate) is a succinct form of summing the resource requirements for a flow given its stochastic behavior and quality expectation. Kelly uses the following form [50]: $a\,(\sigma,\tau) = \frac{1}{\sigma\tau}\log Ee^{\;\sigma\Re(\tau)}$, $0 < \sigma,\tau < \infty$, where $\Re(\tau)$ represent the number of bits sent in an interval τ second long; σ is a quality parameter. The equivalent capacity is bounded by the average and the peak rates of the source, as illustrated in Figure 7.9.

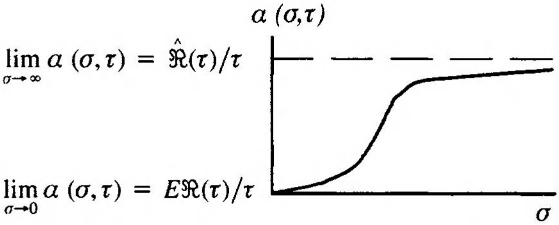


Figure 7.9 The equivalent capacity as a function of σ is bounded by the peak and the average rates of the source.

As mentioned before, stochastic models are seldom useable as traffic descriptors. They can however be used indirectly for calculating the equivalent capacity. The volume of reported work on modeling of unregulated sources is large, as mentioned before. Most models are fitted to video conference sequences and disregard the scene to scene variations (even though it might be present also for this application). In recent year there has been studies suggesting that the bit-rate process exhibits long–range behavior [3] [18]. The implications have been considered recently [25][26][34][77]. The apparent behavior has also be explained as level shifts rather than as long-range dependence [13] [17][22].

In summary, one means of describing unregulated video sources is by fitting a stochastic model that captures the time scales of the program, and to calculate the equivalent capacity for the source. The use of a smoothing buffer can be captured by the calculation and will result in a lower equivalent capacity for a given source model and loss level. Smoothing algorithms, for instance those presented in [7][51][57][71][91], will have to obey the upper limit given by the requested bit rate.

Note that a fixed–rate traffic descriptor does not provide any information to exploit for statistical multiplexing across virtual channels. The needed information for this has to be gotten from on–line measurements of established connections.

4.6. REGULATED ENCODING

The issue of determining a traffic descriptor does not change in principle when the coder can be regulated: One has to choose a type of descriptor and to provide a procedure for estimating the descriptor's parameters.

Leaky bucket descriptors have recently been studied for regulated video and it has been found that it is better fitted for regulated than unregulated coding [29][37]. It is clear that the feedback makes is possible to regulate the bit rate from the coder in order to fit any choice of leaky–bucket parameters. Whether a particular set of parameters is good or not can only be determined by subjectively evaluating the encoding quality. Hsu *et al.* have established that a smoothing buffer of size B together with the leaky–bucket descriptor $\left(\hat{R}, \overline{R}, b\right)$ yieldsthe same quality as a system with a buffer of size $B+b$ and the single upper bound descriptor $\left(\overline{R}, \overline{R}, 0\right)$ [37]. The gain is therefore a lower delay in the former case since the buffer can be b bits smaller without any effect on the quality (it should be noted, however, that the first case might require a higher capacity allocation due to the allowed burstiness).

Consider again the three time scales of a video session: together they range from a few tens of milliseconds up to hours, with a high degree of variability on each scale. It does not appear to be reasonable to expect that a bound can be predicted for the establishment of each new session. The practical way is therefore to fit a small set of bounds to the coding system which will be enforced by the bit rate regulation. The quality impact for each bound would have to be evaluated by means of subjective testing based on test sequences. Thus, a low–end bound would be evaluated by sequences that are representative of applications requiring only a reasonable quality (for instance, an informal conversation). A high–end bound should be enforceable without noticeable impact on the quality for the most demanding test sequences, such as an action film. When in operation, each application is configured to request the most appropriate of the bounds which have been selected by the codec manufacturer.

5. NETWORK SUPPORT

For a network provider the interesting issue is the specific needs for quality guarantees that users may have and the best way of providing them in the network. It is worth noting in this respect that asynchronous time–division multiplexing enables statistical multiplexing but does not mandate it. Capacity allocation could thus be deterministic or statistical, and both types could to be offered in the networks since they are complementary.

There are three general service classes to consider:

- Deterministic multiplexing with quality guarantees.
- Statistical multiplexing with quality guarantees.
- Statistical multiplexing without quality guarantees.

For ATM, in the terminology of the ITU–T Recommendation I.371, these classes are referred to as ATM–layer bearer capabilities. The listed classes are called deterministic bit rate (DBR), statistical bit rate (SBR) (called constant and variable bit rate, respectively, by the ATM Forum), and unspecified bit rate (UBR). A specific instance of the UBR, that includes rate-based flow control, is the available bit rate (ABR) service. We briefly review these classes with respect to video transfer in this section.

5.1. DETERMINISTIC BIT RATE SERVICE

Deterministic multiplexing means that all flows are bounded and that enough capacity and buffers are reserved in the network to assure complete absence of overflow. The bit-rate bound is either fixed for the duration of the session, or it is renegotiated at need, if permitted by the network control [24]. The quality that can be guaranteed is absence of cell loss, and well–bounded delay and delay jitter.

The recommendation for DBR specifies a peak rate as the sole traffic descriptor. This is not an unsuitable descriptor for video, especially if it can be renegotiated at need. The delay bounds that can be computed would not improve if a leaky-bucket descriptor was used. If a rate $\overline{R}$ is allocated in the network and the sender emits cells at a higher rate, $\hat{R} > \overline{R}$, then the maximum delay is the same if the source shapes the rate down to $\overline{R}$, or if one of the nodes along the route does it [59]. The latter case requires more buffering in the network nodes at no gain, which possibly leads to a more expensive system.

The common objection to deterministic multiplexing is that the reserved capacity is poorly used when only loose bounds, such as a fixed rate, can be placed on the connections. A buffering system can, however, be designed so that service classes with statistical multiplexing (*eg,* ABR and UBR) can be offered in addition to the deterministic service. These service classes could then use all unreserved as well as reserved but unused link capacity [45]. Traffic in the statistical classes can thus expend any slack in the reserved capacity.

The issue is of DBR versus SBR therefore one of service costs: given a desired level of perceived quality, will the needed allocation of capacity cost more for DBR service than for SBR service?

5.2. STATISTICAL BIT RATE SERVICE

Statistical multiplexing with quality guarantees is what most researchers have considered the traffic class of choice for video. However, few have evaluated the actual procedure where a connection is characterized solely by a leaky–bucket, and the network would not have any other verifiable information about the traffic on the requested connection.

Rather, most reported works on multiplexing performance for video usually model an aggregate of independent and identically distributed video flows. The model can then be used to evaluate the possible loads that can be sustained for given cell loss probabilities and buffer sizes [9][61][82] [85]. The results from such statistical approach are not relevant as performance predictions for the operational approach that ATM admission control is based on. The attainable statistical multiplexing gain of variable–rate video is far below the values predicted by the statistical approach [73].

The reason is that the bound on the traffic flow could be arbitrarily tight or loose depending on the actual information contents in the signal. Further assumptions about the variations within the bound cannot be justified, and, to be safe, an on–off pattern that is admissible under the given leaky–bucket parameters is often assumed (even though such behavior is not observed for variable bit rate video) [29][73].

The maximum utilization for a given cell loss probability (p_{loss}) decreases with increasing $\hat{R}$ and $\bar{R}$ in relation to the link capacity (C), and with b in relation to the multiplexer buffer size. According to Ch. 5.2 in [75], the allocation for a connection can be approximated by

$$R^* = a\bar{R}\left(1 + 3z\left(1 - \bar{R}/\hat{R}\right)\right) \text{ for } 3z < min(3, \hat{R}/\bar{R}),$$

where

$$a = 1 - log_{10}p_{loss}/50 \text{ and } z = -2\hat{R}\log_{10}p_{loss}/C.$$

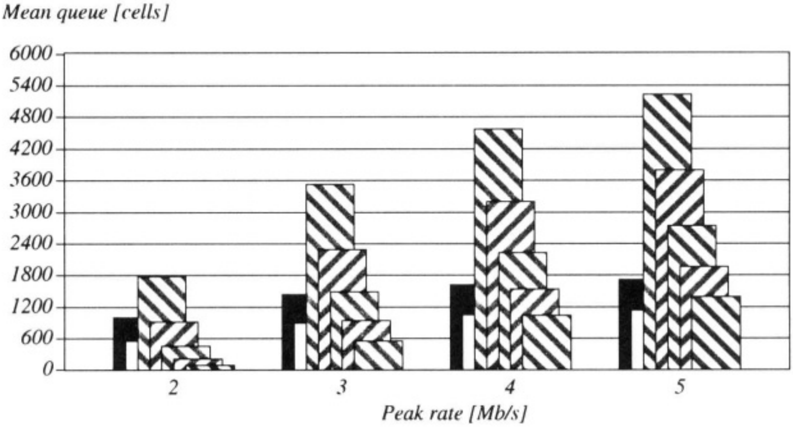
This assumes a fluid-flow model of the traffic and a bufferless multiplexer (thus, b is not appearing in the expression). It follows that the utilization of the allocation is $\bar{R}/R^* < 1$ when $\hat{R} > \bar{R}$. A full utilization means that all connections are allocated their sustainable rates. This should not be mistaken for the actual usage of the link which could be arbitrarily much lower since the sustainable rates could be well above the actual mean rates of the flows. This in turn depends on the accuracy in estimation of the traffic descriptor parameters.

A valid question is what gain can be achieved by SBR over DBR for a given source. Let us consider the following two cases for a rate controlled source: 1) a smoothing buffer of size B_1 with a fixed–rate descriptor $\left(\bar{R}, \bar{R}, 0\right)$ and deterministic multiplexing. An allocation of $R_1^* = \bar{R}$ is hence sufficient for absence of loss and for low delay in the network. Case 2) would be the same source with a smoothing buffer of size B_2 and a $\left(\hat{R}, \bar{R}, b\right)$ leaky–bucket descriptor. With statistical multiplexing, it would require an allocation R_2^*, as computed above. Recall that the encoding quality of the two cases is comparable if $B_1 = B_2 + b$ [37].

In terms of allocation, it is then clear that the deterministic case requires less network capacity, $R_1^* < R_2^*$, yet it provides a higher quality of service than the statistical case. The cost, however, is a higher smoothing delay since $B_1 > B_2$. To compare delays fairly, the allocations could instead be fixed to $R^* = R_1^* = R_2^*$in both cases. The smoothing buffer in case 1) can then be reduced by an amount β since it is emptied at a higher rate while the frequency of feedback should be kept constant. The maximum delay for the first case is therefore $(B_2 + b - \beta)/R^*$ plus some small network contribution due to the asynchronous multiplexing.

The second case has maximum smoothing delay $B_2/\overline{R}$ (the leaky–bucket results in a smaller effective buffer than the previous case but it is serviced at the sustainable rate in the worst case). The network delays are negligible if the allocation is calculated according to the method above, which is for a bufferless system. Comparing the two cases, we find that deterministic multiplexing results in lower total delay when $\beta > b - B_2(R^* - \overline{R})/\overline{R}$.

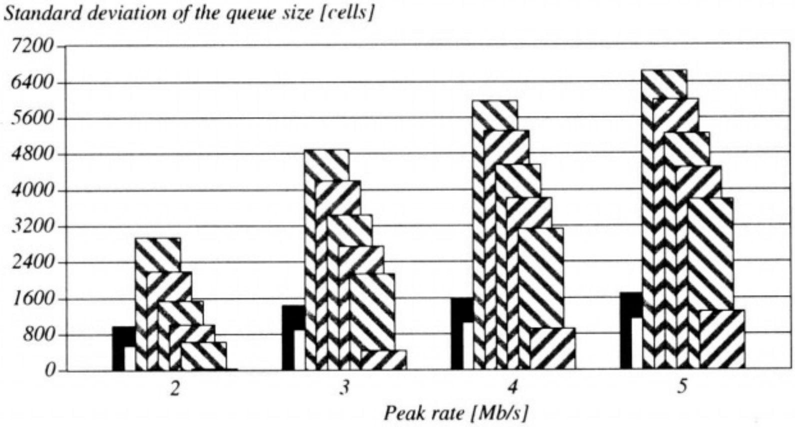
Figure 7.10 The mean queue size seen by an arriving cell as a function of the peak rate $\hat{R}$. The columns from left to right are for DBR with an SBR allocation based on $p_{loss} = 10^{-6}$ (black) and $p_{loss} = 10^{-6}$ (white), and SBR with burst sizes of 1, 2, 3, 4 and 5 Mb.



In [48], DBR and SBR have been compared at equal allocations of capacity in the network. The source model is a two–state Markov chain with a peak rates of 2, 3, 4 and 5 Mb/s; the average rate is 0.8 Mb/s and the average burst size is 1 Mb in all cases. It is thus a model which has the on–off behavior assumed in the formula for allocation of capacity above (the allocation would be more pessimistic for a smoother, more realistic source). Figure 7.10 shows the mean queue lengths as a function of the peak rates. The left-most two solid black and white bars are for DBR with SBR allocation computed for cell–loss probabilities 10^{-6} and 10^{-9}, respectively. The striped bars are for SBR with 1 Mb/s sustainable rate (80 percent utilization), a peak rate equal to that of the source and burst sizes, *b*, of 1,2,3,4 and 5 Mb. The 95–percent confidence interval was within ±4 cells in all cases and the buffer size was large enough to avoid cell loss (simulated time: 5000 seconds per case). Figure 7.11 shows the standard deviation of the queue size (it also includes results for *b* = 10 Mb).

We note for instance that the burst size of the leaky–bucket should be four times the source's average burst size at a peak rate of 4 Mb/s in order for SBR to outperform DBR for a SBR allocation based on 10^{-6} cell loss rate. In general we see that the queue length distribution is much more stretched out for SBR with leaky–bucket service compared to DBR with fixed–rate service. This indicates that there could be many cases for which deterministic multiplexing outperforms statistic multiplexing also in terms of delay [47][48][70].

Figure 7.11: Standard deviation of the queue size. The right–most column is for a burst size of 10 Mb.



One way of circumventing the uncertain multiplexing performance caused by lack of information about source behaviors is to use measurements of the traffic on existing connections. The easiest descriptor for the sender to select and abide to is a single upper bound. By measuring the actual load of the connections on a link it is possible to predict whether the new connection can be established or not. The advantage is that prediction is made for an ensemble of connections and that it may be adjusted for each new connection being accepted [25]. Measurement–based techniques are suggested in [10][14][21][40][79]. It might thus be possible to obtain a reasonable statistical multiplexing gain and yet offer some form of useful quality guarantees without other prior knowledge than the sources' peak rates.

5.3. AVAILABLE AND UNSPECIFIED BIT RATE SERVICES

The quality of "best effort" service is determined by the amount of capacity and the users' demand and behavior. A few ill–behaving users could thus lower the quality for everyone. The notion of "best effort" should therefore include both the network's users and its operator. Misbehavior could in principle be policed although it might be difficult in practice to control a behavior (rather than a bound). The network service for interactive applications is therefore likely to contain some type of quality guarantees, deterministic or statistical. This choice is, however, open and will most likely be determined by economic factors such as system complexity and capacity utilization. Less interactive applications can, however, benefit from the ABR and UBR services.

For ATM, there is a variant defined on the unspecified bit–rate service called available bit rate (ABR). The traffic descriptor for ABR is a single rate which is adjusted according to the load level in the network with consideration of specific needs stated in the set–up message. For instance, a minimum bit rate can be requested which could be chosen according to the bit rate needed for the bare necessity of quality. The cell loss will be minimized for sources which obey congestion notifications from the network [27]. Such messages could naturally be used to regulate the service rate of the buffer at the encoder [5][41]. It should be understood, however, that the enforced limit is chosen by the network without knowledge of the source's varying needs for capacity. The resultant quality may consequently vary noticeably and at time be less than adequate for the communication session.

6. MATCHING OF VIDEO TO NETWORK SERVICES

In order to determine what network services are useful for video, we have to consider the types of video service and the programming. The first determinant is whether the program is recorded or transferred live. In the first case, all parameters of interest to the connection–admission can be computed during the recording. If sent live, the parameters have to be forecasted, or chosen from a pre–determined set of values. The limit in admission control for replayed video is thus the network's acceptance control algorithm: how much information it can include in the computation and how much time it has available for reaching a decision on accepting or rejecting the connection. It should be noted that user–controlled playback of recorded video is not fully predictable due to the possibility for the viewer to pause, skip, rewind and fast forward in the program [67].

There is a second issue to consider, namely the quality requirement. We divide the requirements into generic high, medium and unspecified quality and discuss the pertinent services for each class. The higher the quality, the less underestimation of traffic characteristics can be accepted. And, the longer the time scale a parameter affects, the more will it determine the perceived quality.

High quality

High–quality video is suitable for broadcast television, education, business meetings or similar settings with low tolerance to glitches. High–quality video transfers would most likely be using MPEG–2 coding and the network service of choice is then deterministic multiplexing. The reason is that the high quality and high-bit rate leaves little gain for statistical multiplexing. When the rate is fixed for the duration of the session the sender has to adapt to it; when re–negotiation is allowed, the sender adapts to frequent fluctuations and the network to the low–frequency variations. By dynamic adjustments of the coding mode I, P, and B, MPEG–2 offers a high degree of flexibility in the trade–off between rate control and error resilience.

Connection parameters have to be selected in order to give a high quality throughout the duration of the session. The worst case behavior on the program level has thus to be anticipated and used for the specification. When re–negotiation is allowed, the sender has to be able to anticipate the performance on the scene level and re–negotiate the connection–parameters based on that. Estimation algorithms for the behavior on the program level or scene level are not yet reported to any greater extent in the literature, see [8] [24] for some recent work in the area.

Since this service class is for sessions with stringent quality requirements it has to be understood that possible blocking on the connection–level might not acceptable. After all, a scheduled lecture has to occur and could be booked in advance if the network would allow it. Very little work on advanced reservation has alas been done to date [12][23]. This point is also valid for the next service class.

Medium quality

With relaxed quality requirements, it might be suitable to consider statistical multiplexing with quality guarantees. Re–negotiations might be possible also for SBR service but it should be clear that the connection acceptance is more complex for SBR compared to DBR. The acceptable frequency of changes might therefore be lower, and hence requiring a longer planning horizon for the sender. The usefulness of the re–negotiation option is therefore reduced. Sufficient gain in utilization to compensate for the more complex traffic control compared to a DBR service will most likely only be achieved by measurement–based approaches, as we have indicated above.

Such techniques work better the more flows are aggregated. MPEG–2 video flows with mean bit rates as high as 5 to 10 Mb/s could therefore be too dominant on a link and compromise the workings of the measurement–based performance predictions. It is therefore reasonable to assume that the SBR service class is most suitable for low bit rate video, coded for instance according to ITU–T H.263 [81]. Thus, the SBR service fits well low bit rate applications with reasonable quality expectations [16].

Unspecified quality

There is a common misconception that video must be given quality guarantees in the network. Many video services are one-way and do not have any limits on end–to–end delay that go beyond those of data transfers. The bit rate can therefore be smoothed to near–constant rate to avoid causing load surges and the network nodes could provide ample amounts of buffering. There is also the possibility that the video application could adapt to various degrees of loss. Various video tools are, for instance, used regularly for video conferences over the Internet, often reaching wide audiences via the MBone [15]. The network quality can also be increased by means of forward–error correction [2].

It is, however, clear that the coding schemes developed for a synchronous TDM environment do not handle variations in transfer quality well. Video systems should therefore be developed directly for asynchronous transfers (for an example see [63]). Layered coding in conjunction with concealment can reduce the visibility of losses, and asynchronous decoding can handle some delay jitter on behalf of temporal aliasing.

7. SUMMARY

This paper has summarized some of the known user requirements for video communication and the network services that should ensure them. The coding was explained, with special consideration to MPEG–2 and layered coding, as well as framing and bit rate regulation. In terms of traffic characterization, it is now evident that bounding is the only viable alternative; stochastic models cannot be enforced nor verified. In fact, a single upper limit on the bit rate appears as suitable as a leaky-bucket, and it was shown that deterministic multiplexing may outperform statistic multiplexing both in terms of total delay and multiplexing efficiency. This is in contrast to much of the conventional assumptions about variable bit rate video transfers.

The options to provide appropriate network service for various video applications have been outlined. There are still opportunities for further research in that area to determine suitable traffic parameters for various applications and to find the best matching network service to satisfy the quality expectations.

8 REFERENCES

[1] L. Alparone, *et al.*, “Models for ATM Video Packet Transmission,” European Transaction on Telecommunications, Vol. 3, No. 5, September 1992, pp. 67–73.

[2] E. Ayanoglu, *et al.*, ”Performance Improvement in Broadband Networks Using Forward Error Correction for Lost Packet Recovery”, Journal of High–Speed Networks, Vol. 2, No. 3, 1993, pp. 287–303.

[3] J. Beran, *et al.*, “Long-Range Dependence in Variable–Bit–Rate Video Traffic, IEEE Transactions on Communications, Vol. 43, No. 2/3/4, February/March/ April 1995, pp. 1566–1579.

[4] D. E. Blahut, *et al.*, “Interactive Television,” Proceedings of IEEE, Vol. 83, No. 7, July 1995, pp. 1071–1085.

[5] J–C Bolot, *et al.*, “Scalable Feedback Control for Multicast Video Distribution in the Internet,” ACM Computer Communications Review, Vol. 24, No. 4, October 1994, pp. 58–67.

[6] T. Chiang and D. Anastassiou, “Hierarchical Coding of Digital Television,” IEEE Communications Magazine, Vol. 32, No. 5, May 1994, pp. 38–45.

[7] C–T Chien and A. Wong, “A Self-Governing Rate Buffer Control Strategy for Pseudoconstant Bit Rate Video Coding,” IEEE Transactions on Image Processing, Vol. 2, No. 1, January 1993, pp. 50–59.

[8] S. Chong, *et al.*, “Predictive Dynamic Bandwidth Allocation for Efficient Transport of Real–Time VBR Video over ATM,” IEEE Journal on Selected Areas in Communications, Vol. 13, No. 1, January 1995, pp. 12–23.

[9] D M. Cohen and D. P. Heyman, “Performance Modeling of Video Teleconferencing in ATM Networks,” IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. 6, December 1993, pp. 408–420.

[10] C. Courcoubetis, *et al.*, “Admission Control and Routing in ATM Networks Using Inferences from Measured Buffer Occupancy,” IEEE Transactions on Communications, Vol. 43, No. 2/3/4, February/March/April 1995, pp. 1778–1784.

[11] R. L. Cruz, “A Calculus for Network Delay, Part I: Network Elements in Isolation; Part II: Network Analysis,” IEEE Transactions on Information Theory, Vol. 37, No. 1. January 1991, pp. 114–141.

[12] M. Degermark, *et al.*, “Advance Reservations for Predictive Service in the Internet” ACM–Springer Journal of Multimedia Systems, 1997.

[13] N. G. Duffield, *et al.*, “Predicting Quality of Service for Traffic with Long-Range Fluctuations,” in *Proceedings of IEEE International Conference on Communications,* Seattle, Washington, USA, June 18–22, 1995.

[14] N. G. Duffield, *et al.*, "Entropy of ATM Traffic Streams: A Tool for Estimating QoS Parameters," IEEE Journal on Selected Areas in Communications, Vol. 13, No. 6, August 1995, pp. 981–990.

[15] H. Eriksson, "MBONE: The Multicast Backbone," Communications of the ACM, Vol. 37, No. 8, August 1994, pp. 54 60.

[16] R. S. Fish, *et al.*, "Video as a technology for informal communication," Communication of the ACM, Vol. 36, No. 1, January 1993, pp. 48–61.

[17] M. R. Frater, *et al.*, "A New Statistical Model for Traffic Generated by VBR Coders for Television on the Broadband ISDN," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 4, No. 6, December 1994, pp. 521–526.

[18] M. W. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," ACM Computer Communications Review, Vol. 24, No. 4, October 1994, pp. 269–280.

[19] M. Ghanbari, "Two–layer Coding of Video Signals for VBR Networks," IEEE Journal on Selected Areas in Communications, Vol. 7, No. 5, June 1989, pp. 771–781.

[20] M. Ghanbari and V. Seferidis, "Cell–Loss Concealment in ATM Video Codecs," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. 3, June 1993, pp. 238–247.

[21] R. J. Gibbens, F. P. Kelly, and P. Key "A Decision–Theoretic Approach to Call Admission Control in ATM Networks," IEEE Journal on Selected Areas in Communications, August 1995, pp. 1101–1114.

[22] M. Grasse, M. R. Frater, and J. F. Arnold, "Origins of Long–Range Dependence in Variable Bit Rate Video Traffic," in *Proc. of ITC–15,* Elsevier, June 1997, pp. 1379–1388.

[23] A. Greenberg, R. Srikant, and W. Whitt, "Resource Sharing for Book–Ahead and Instantaneous–Request Calls," in *Proc. of ITC–1*5, Elsevier, June 1997, pp. 539–548.

[24] M. Grossglauser, *et al.*, "RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic," ACM Computer Communications Review, Vol. 25, No.4, October 1995, pp. 219–230.

[25] M. Grossglauser and D. Tse, "Towards a Framework for Robust Measurement-Based Admission Control," ACM Computer Communication Review, Vol. 27, No. 4, October 1997, pp. 237–248.

[26] M. Grossglauser and J–C Bolot, "On the Relevance of Long–Range Dependence in Network Traffic," ACM Computer Communication Review, Vol. 26, No. 4, October 1996, pp. 14–24.

[27] R. Jain, *et al.*, "Source Behavior for ATM ABR Traffic Management: An Explanation," IEEE Communications Magazine, Vol. 34, No. 11, November 1996, pp. 50–57.

[28] R. Grünenfelder, *et al.*, "Characterization of Video Codecs as Autoregressive Moving Average Processes and Related Queueing System Performance," IEEE Journal on Selected Areas in Communications, Vol. 9, No. 3, April 1991, pp 284–293.

[29] M. Hamdi, *et al.*, "Rate Control for VBR Video Coders n Broad-Band Networks," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 1040–1051.

[30] H. Heeke, "Statistical Multiplexing Gain for Variable Bit Rate Video Codecs in ATM Networks," International Journal of Digital and Analog Communication System, Vol. 4, 1991, pp. 261–268.

[31] H. Heeke, "A Traffic–Control Algorithm for ATM Networks," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. 3, June 1993, pp. 182–189.

[32] D. P. Heyman, *et al., "*Statistical Analysis and Simulation of Video Teleconference Traffic in ATM Networks," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 2, No. 1, March 1992, pp. 49–59.

[33] D. P. Heyman and T. V. Lakshman, "Source Models for VBR Broadcast–Video Traffic," IEEE/ACM Transaction on Networking, Vol. 4, No. 1, February 1996, pp. 40–48.

[34] D. P. Heyman and T. V. Lakshman, "What are the implications of Long–Range Dependence for VBR–Video Traffic Engineering?," IEEE/ACM Transactions on Net working, Vol. 4, No. 3, June 1996, pp. 301–317.

[35] D. P. Heyman, "The GBAR Source Model for VBR Videoconferences," IEEE/ACM Transactions on Networking, Vol. 5, No. 4, August 1997, pp. 554–560.

[36] C. J. Hughes, *et al.*, "Modeling and Subjective Assessment of Cell Discard in ATM Video," IEEE Transactions on Image Processing, Vol. 2, No. 2, April 1993, pp. 212–222.

[37] C.Y. Hsu, *et al.* "Joint Selection of Source and Channel Rate for VBR Video Transmission Under ATM Policing Constraints," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 1016–1028.

[38] S. lai and N. Kitawaki, "Effects of Cell Loss on Picture Quality in ATM Networks," Electronics and Communications in Japan, Part 1, Vol. 75, No. 10, pp. 30–41.

[39] B. Jabbari, *et al.*, "Statistical Characterization and Block-Based Modelling of Motion–Adaptive Coded Video," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. 3, June 1993, pp. 199–

[40] S. Jamin, *et al.*, "A Measurement-based Admission Control Algorithm for Integrated Services Packet Networks,," IEEE/ACM Transactions on Networking, February 1997, pp. 56–70.

[41] H. Kanakia, *et al.*, "An Adaptive Congestion Control Scheme for Real–Time Packet Video Transport," ACM Computer Communication Review, Vol. 23, No. 4, October 1993, pp. 20–31.

[42] G. Karlsson and M. Vetterli, "Sub–band coding of video for packet networks," Optical
Engineering, Vol. 27, No. 7, July 1988, pp. 574–586.

[43] G. Karlsson and M. Vetterli, "Packet Video and Its integration Into the Network Architecture," IEEE Journal on Selected Areas in Communications, Vol. 7, No. 5, June 1989, pp. 739–751.

[44] G. Karlsson, "ATM Adaptation for Video," in *Proceedings of Sixth International Workshop on Packet Video*, Portland, OR, September 26–27, 1994, pp. E3.1–5.

[45] G. Karlsson, "Capacity Reservation in ATM Networks," Computer Communications, Vol. 19, No. 3, March, 1996, pp. 180–193.

[46] G. Karlsson, "Asynchronous transfer of video," IEEE Communications Magazine, Vol. 34, No. 8, August 1996, pp. 118–126.

[47] G. Karlsson and G. Djuknic, "On the efficiency of statistical-bitrate service for video," in Performance of Information and Communication Systems (Eds. U. Körner and A. A. Nilsson), Chapman and Hall, 1998, pp. 205–215.

[48] G. Karlsson, "On the quality provisioning for video in ATM networks," in *Proc. 2000 International Zurich Seminar on Broadband Communications,* Zurich, Switzerland, Feb. 15–17, 2000

[49] M. Kawashima, *et al.*, "Adaptation of the MPEG Video–Coding Algorithm to Network Applications," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. 4, August 1993, pp. 261–269.

[50] F. P. Kelly, "Notes on Effective Bandwidth," in *Stochastic Networks: Theory and Applications* (Eds. F. Kelly, S. Zachary, I. Ziedins), Oxford University Press, 1996.

[51] G. Keesman, *et al.*, "Bit–Rate Control for MPEG Encoders," Signal Processing: Image Communication, Vol. 6, 1995, 545–560.

[52] L. H. Kieu and K. N. Ngan, "Cell-Loss Concealment Techniques for Layered Video Codecs in an ATM Network," IEEE Transactions on Image Processing, Vol. 3, No. 5, September 1994, pp. 666–676.

[53] T. Kinoshita, *et al.*, "Variable–Bit–Rate HDTV CODEC with ATM–Cell–Loss Compensation," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. June 1993, pp. 230–237.

[54] N. Kitawaki and K. Itoh, "Pure Delay Effects on Speech Quality in Telecommunications," IEEE Journal on Selected Areas in Communications, Vol. 9, No. 4, May 1991, pp. 586–593.

[55] T. Kurita, *et al.,* "Effects of Transmission Delay in Audiovisual Communcations," Electronics and Communications in Japan, Part 1, Vol. 77, No. 3, 1994, pp. 63–74.

[56] J. Kurose, "On Computing Per–Session Performance Bounds in High–Speed Multi–hop Computer Networks," Performance Evaluation Review, Vol. 20, No. 1, June 1992, pp. 128–139.

[57] S. S. Lam, S. Chow, and D. K. Y. Yau, "A Lossless Smoothing Algorithm for Compressed Video," IEEE/ACM Transactions on Networking, Vol. 4, No. 5, October 1996, pp. 697–708.

[58] A. A. Lazar, *et al.*, "Modeling Video Sources for Real-Time Scheduling," Multimedia Systems, Vol. 1, 1994, pp. 253–266.

[59] J.–Y. Le Boudec, "Application of Network Calculus to Guaranteed Service Networks," IEEE Transactions on Information Theory, Vol. 44, No. 3, May 1998, pp. 1087–1096.

[60] J–P Leduc and P. Delogne, "Statistics for variable bit–rate digital television sources," Signal Processing: Image Communication, Vol. 8, No. 5, July 1996, pp. 443–464.

[61] B. Maglaris, *et al.*, "Performance Models of Statistical Multiplexing in Packet Video Communications," IEEE Transactions on Communications, Vol. COM–36, No. 7, July 1988, pp. 834–844.

[62] N. M. Marafih, *et al.*, "Modelling and Queuing Analysis of Variable Bit-rate Coded Video Sources in ATM Networks," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 4, No. 2, April 1994, pp. 121–128.

[63] S. McCanne and V. Jacobson, "vic: A Flexible Framework for Packet Video," in *Proceedings of ACM Multimedia,* San Francisco, CA, November 5–9, 1995, pp. 511–522.

[64] D. L. McLaren and D. T. Nguyen, "Variable Bit-Rate Source Modelling of ATM–Based Video Services," Signal Processing: Image Communication, Vol. 4, 1992, pp. 233–244.

[65] J. L. Mitchell, *et al.*, *MPEG Video: Compression Standard,* Chapman and Hall, New York, NY, 1997.

[66] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation, Compression and Standards*, Plenum Press, 1995

[67] J–P Nussbaumer, *et al.*, "Networking Requirements for Interactive Video on Demand," IEEE Journal on Selected Areas in Communications, Vol. 13, No. 5, June 1995, pp. 779–787.

[68] S. Okubo, *et al.*, "ITU–T Standardization of Audiovisual Communication Systems in ATM and LAN Envrionments," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 965–982.

[69] A. Ortega, *et al.*, "Rate Constraints for Video Transmission over ATM Networks Based on Joint Source/Network Criteria," Annales des Telecommunications, Vol. 50, No. 7–8, 1995, pp. 603–616.

[70] B. V. Patel and C. C. Bisdikian, "End-Station Performance under Leaky Bucket Traffic Shaping," IEEE Network, September/October 1996, pp. 40–47.

[71] M. R. Pickering and J. F. Arnold, "A Perceptually Efficient VBR Rate Control Algorithm," IEEE Transactions on Image Processing, vol. 3, No. 5, September 1994, pp. 527–531.

[72] E. P. Rathgeb, "Policing of Realistic VBR Video Traffic in an ATM Network," International Journal of Digital and Analog Communication System, Vol. 6, 1993, pp. 213–226.

[73] A. R. Reibman and A. W. Berger, "Traffic Descriptors for VBR Video Teleconferencing over ATM Networks," IEEE/ACM Transactions on Networking, Vol. 3, No. 3, June 1995, pp. 329–339.

[74] R. M. Rodriguez–Dagnino, *et al.*, "Prediction of Bit Rate Sequences of Encoded Video Signals," IEEE Journal on Selected Areas in Communications, Vol. 9, No. 3, April 1991, pp 305–314.

[75] J. Roberts, U. Mocci and J. Virtamo (Eds.), *Broadband Network Teletraffic*, Springer, 1996.

[76] O. Rose and M. R. Frater, "A Comparison of Models for VBR Video Traffic Sources in B-ISDN," in *IFIP Broadband Communications II*, Eds. S. Tohmé and A. Casaca, Elsevier Science (North-Holland), 1994, pp. 275–287.

[77] B. K. Ryu and A. Elwalid, "The Importance of Long–Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities," ACM Compu ter Communication Review, Vol. 26, No. 4, October 1996, pp. 3–14.

[78] D. Saha, *et al.*, "Multirate Scheduling of VBR Video Traffic in ATM Networks," IEEE Journal on Selected Areas in Communications, Vol. 15, No. 6, August 1997, pp. 1132–1147

[79] H. Saito, "Dynamic Resource Allocation in ATM Networks," IEEE Communication Magazine, Vol. 35, No. 5, May 1997, pp. 146–153.

[80] N. B. Seitz, *et al.*, "User-Oriented Measures of Telecommunication Quality," IEEE Communications Magazine, Vol. 32, No. 1, January 1994, pp. 56–66.

[81] R. Schäfer and T. Sikora, "Digital Video Coding Standards and Their Role in Video Communication," Proceedings of IEEE, Vol. 83, No. 6, June 1995, pp. 907–924.

[82] C. Shim, *et al.*, "Modeling and Call Admission Control Algorithm of Variable Bit Rate Video in ATM Networks," IEEE Journal on Selected Areas in Communications, Vol. 12, No. 2, February 1994, pp. 332 – 344

[83] T. Sikora, "MPEG Digital Video–Coding Standards," IEEE Signal Processing Magazine, Vol. 14, No. 5, September 1997, pp. 82–100.

[84] R. P. Singh *et al.*, "Jitter and Clock Recovery for Periodic Traffic in Broadband Packet Networks," IEEE Transactions on Communications, Vol. 42, No. 5, May 1994, pp. 2189–2196.

[85] P. Skelly, *et al.*, "A Histogram–Based Model for Video Traffic Behavior in an ATM Multiplexer," IEEE/ACM Transactions on Networking, Vol. 1, No. 4, August 1993, pp. 446–459.

[86] R. Steinmetz, "Human Perception of Jitter and Media Synchronization," IEEE Journal on Selected Areas in Communications, Vol. 14, No. 1, January 1996, pp. 61–72.

[87] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, 1995

[88] M. Wada, "Selective Recovery of Video Packet Loss Using Error Concealment," IEEE Journal on Selected Areas in Communications, Vol. 7, No. 5, June 1989, pp. 807–814.

[89] D. Wrege, *et al.*, "Deterministic Delay Bounds for VBR Video in Packet–Switching Networks: Fundamental Limits and Practical Trade–offs," IEEE/ACM Transactions on Networking, Vol. 4, No. 3, June 1996, pp. 352–362.

[90] Q–F Zhu, *et al.*, "Coding and Cell–Loss Recovery in DCT–Based Packet video," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 3, No. 3, June 1993, pp. 248 –

[91] J. Zdepski, *et al.*, "Statistically Based Buffer Control Policies for Constant Rate Transmission of Compressed Digital Video," IEEE Transactions on Communications, Vol. 39, No. 6, June 1991, pp. 947–957.

Gunnar Karlsson is professor at the Department of Teleinformatics of the Royal Institute of Technology (KTH) since 1998. He has previously worked for IBM Zurich Research Laboratory and the Swedish Institute of Computer Science (SICS). His Ph.D. is from Columbia University, New York, and the M.Sc. from Chalmers University of Technology in Gothenburg. He has been visiting professor at EPFL in Switzerland, and the Helsinki University of Technology in Finland. His research interests lie within the general field of multimedia networking.

Chapter 8

OPTIMAL RESOURCE MANAGEMENT IN ATM NETWORKS

Based on Virtual Path Bandwidth Control

Michael D. Logothetis
Wire Communications Laboratory, Department of Electrical & Computer Engineering, University of Patras, 265 00 Patras, Greece.
m-logo@wcl.ee.upatras.gr

Abstract: In the beginning an overview of network/traffic control in ATM networks is presented, based on the fact that traffic control is distinguished in two levels, the Call-level and the Cell-level control, according to the distinction of ATM traffic in call and cell components, respectively. Afterwards, the paper concentrates on the Call-level and the impact of Virtual Path Bandwidth **(VPB)** control on ATM network performance. In particular the optimal VPB control is presented, minimizing the worst Call Blocking Probability of all Virtual Paths **(VP)** of the network. A VPB controller solves a large network optimization problem by a rigorous analytical procedure, while can assure network reliability. The procedure for optimal VPB allocation is clarified, step-by-step, in a tutorial application example. In a more realistic example, the optimal VPB control is applied on a model ATM network.

Keywords: ATM, Traffic Control, Virtual Path, Bandwidth Control, Optimization.

1. INTRODUCTION

The role of bandwidth management in quality and network-reliability assurance is upgraded in the expected environment of ATM networks [1]. In the near future, ATM networks will convey traffic of several service-classes with very different requirements in bandwidth (bits per second) and quality of service **(QoS)** per call, while reliable traffic demand forecasting for these services seems to be impossible. Moreover, different traffic streams are mixed and commonly share an end-to-end link. This wide variety of service-classes renders the resource management more difficult but also more important. In order to simplify the study, this paper concerns service-classes with strict QoS requirements, as they are the constant bit rate (CBR) and the variable bit rate (VBR) service-classes.

Two levels of traffic control, the Call level and the Cell level control are present in ATM networks. These correspond to the distinction of traffic in call and cell components, respectively (Fig. 8.1) [2]. This paper is concentrated on VPB control, which is a medium- or long-term Call-level network control. In cooperation with a bandwidth reservation control scheme, it changes the installed bandwidth in the VPs according to the offered traffic so as to improve the global performance of the network, under constraints posed by the transmission links capacities [3,4]. The resultant distribution of the totally installed bandwidth in the network to the VPs is the VPB allocation. VPB allocation can assure network reliability in a high degree. A reliable bandwidth allocation is considered, by enforcing the bandwidth to be distributed at least in two VPs of every switching pair (end-to-end link). To ensure network reliability, however, we need to install an enormous amount of bandwidth in the transmission links, in comparison to an unreliable network, whereas due to traffic variations a lot of bandwidth remains unused. Therefore, the optimal VPB control becomes essential.

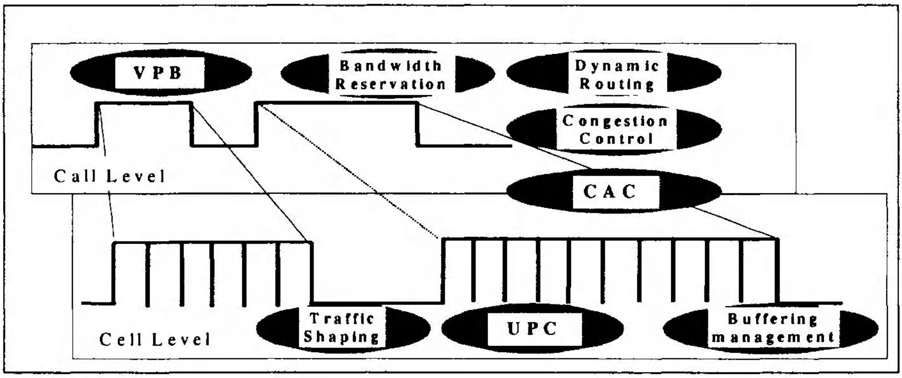


Figure 8.1. Layered structure of traffic control in ATM networks

The optimal VPB allocation is achieved through a network optimization model [5,6]. Many heuristic and efficient algorithms to solve a network optimization problem have been proposed for ATM networks [3,4,7,8,9], whereas path bandwidth management has been considered in synchronous transfer mode **(STM)** networks too [10-14]. All the proposed algorithms, however, lead to sub-optimal or practically optimal results. For a refined network study and for evaluation of the various bandwidth control schemes, it is necessary to apply analytical algorithms, whereby we can obtain accurate (optimal) results even with much consumption of computer memory and CPU-time. To compose the network optimization problem the following are needed to be taken into account: offered traffic, network topology, routing table comprising all VPs, installed bandwidth in transmission links, demand for reliability and optimization criterion. The criterion of minimizing the worst call blocking probability (**CBP**) of the network has been widely adopted. A non-linear-programming problem is formulated, where the objective function is to minimize the worst CBP of the whole network under the following two main constraints:

a) Bandwidth capacity of the transmission links.
b) Reliability constraints.

A rigorous and analytical procedure is presented which leads to the exact (optimal) solution of the network optimization problem [5].

Two application examples are presented in order to clarify and reveal the efficiency of VBP control in resource management. In the first example, the optimal VPB Control is applied on a small network of three nodes, for tutorial purposes. In the second example, the optimal VPB Control is applied on an 8-node ATM network of realistic dimensions.

The organization of this paper is as follows: Section 2 discusses the traffic management in ATM networks from the viewpoint of quality assurance. The two layering controls, Cell-level and Call-level, for traffic management are presented briefly in subsections 2.1 and 2.2, respectively. Section 3 concentrates on the VPB control in more detail. The optimal VPB allocation is presented in section 4. Subsection 4.1 presents an appropriate ATM network architecture for resource/bandwidth management. Network reliability is discussed in subsection 4.2. Subsection 4.3 presents the definition of a network optimization model, in order to obtain the optimal VPB allocation. The solution of the optimization model is presented in subsection 4.4. Section 5 presents two application examples of VPB control in ATM networks: a tutorial example in subsection 5.1 and a more realistic application example, in subsection 5.2, which reveal the performance of VPB management. A conclusion is given in section 6.

2. TRAFFIC MANAGEMENT IN ATM NETWORKS

Traffic management is considered as a series of traffic handling procedures necessary for proper network operation and quality of service assurance. Three main areas of traffic management are distinguished:

a) Network planning.
b) Traffic control.
c) Traffic & QoS measurements.

Network planning is a long-term traffic management and aims at an adequate topological network design, as well as at a proper network dimension (resource allocation) so as to meet the best possible QoS specifications, by taking into account mainly economic factors (available capital investments, etc.). An example of the network-planning subject for an existing network is the planning of the extensions of the bandwidth-capacities of the transmission links of the network.

Traffic control is a rather medium- or short-term control and aims at achieving the best possible QoS for certain (given) network resources. QoS often decreases when an imbalance exists between the network resources and the offered traffic. To improve the QoS, a first action (short-term control) is taken by a traffic control mechanism in order to remove the cause, while a final action (long-term) is taken by the network planning.

Offered/carried traffic measurements and QoS measurements are important, because they are necessary for the traffic control and network planning. Real-time traffic monitoring is necessary in changing the traffic control parameters adaptively so as to improve the flexibility and reliability of the traffic control. Long-term measurements of QoS of the network lead the network planning to timely assignment of the network resources.

The subjects of the traffic control are presented below, according to the layering structure of traffic in ATM networks. As it is illustrated in figure 8.1, the main objectives of the Call-level and the Cell-level traffic controls are the calls and the cells, respectively. The QoS index of the Cell-level traffic management is expressed by the cell-loss probability (cell loss rate, **CLR**) and the cell transfer delay (**CTD**), whereas the QoS index of the Call-level traffic management is expressed by the CBP. The subjects (functions) of the Cell-level traffic controls are buffering management, usage parameter control, traffic shaping, and connection (call) admission control. The Call-level traffic controls include the following functions: call congestion control, bandwidth (trunk) reservation control, VPB control and dynamic routing. Call admission control is also related to the Call-level traffic controls. Concerning VP bandwidth dimensioning (Call- and Cell-level) and buffering dimensioning (Cell-level) are subjects of the network planning.

2.1 CELL-LEVEL TRAFFIC CONTROL

Cell-level traffic control is responsible for the Cell-level QoS assurance and comprises the following specific controls (Fig. 8.1):

– Buffering management (Priority Control)

ATM-cells (traffic streams) with different QoS requirements will be mixed and will commonly share a VP. The most stringent requirements of these cells should be satisfied if they would be handled in the same way. This would lead to excess QoS specifications, which in turn lead to lower traffic throughput. Buffering management control assigns a higher buffer-usage priority to cells with stringent QoS requirements so as to achieve a higher traffic throughput.

– Connection Admission Control (CAC)

When a call set-up request arrives at an ATM network, the ATM switches have to decide whether to establish a virtual channel/path (VC/VP) connection or reject the call request. A connection request is accepted only when sufficient resources are available to establish the connection through the whole network at the required QoS and to maintain the agreed QoS of existing connections.

– Usage and Network Parameter Control (Traffic Policing)

Usage parameter control **(UPC)** is performed at the input port of ATM switches in the user-to-network interface **(UNI),** whereas network parameter control **(NPC)** is performed at network-network interface **(NNI)** to ensure that traffic generated by a user is within the negotiated contract. When violations are detected, UPC or NPC discards all violating cells or sets the cell loss priority **(CLP)** bit to 1 in the header of the ATM-cells.

– Traffic Shaping

For most VBR sources, cells are generated at the peak rate during the active period, while no cells are transmitted during the silent period. Therefore, it is possible to reduce the peak rate by buffering cells before they enter the network so that the departure rate of the queue is less than the peak arrival rates of the cells. This is called traffic shaping and can be done at the source equipment or at the network access point.

2.2 CALL-LEVEL TRAFFIC CONTROL

Call-level traffic control is responsible for the Call-level QoS assurance and comprises the following traffic controls (Fig. 8.1):

– Congestion control

When many call set-up requests arrive (congest) at a specific ATM switch, it is probable that not all of them will be accepted. Nevertheless, all the calls need a processing offered by the ATM switch. Due to this processing of even unsuccessful calls the performance of the switch deteriorates. To avoid this phenomenon, called congestion, the congestion control restricts the number of call set-up requests when the number of arriving calls exceeds the switching capacity of the destination switch.

– Bandwidth (trunk) reservation control

In ATM networks cells of different service-classes, which have different bandwidth requirements per call are integrated and commonly share a VP. Therefore, the CBP of service-classes with higher bandwidth requirements becomes worse than that of service-classes requiring lower bandwidth. To decrease this imbalance of the CBP, the bandwidth reservation control reserves some fraction of the VP bandwidth to benefit the high-speed calls.

– Virtual Channel Routing control

The Virtual Channel Routing control, also known as dynamic routing, monitors the traffic flow in the transmission links of the network and selects the least loaded route to convey a call. The CBP met in the transmission links as well as the end-to-end CBP of the switching pairs are improved.

– Virtual Path Bandwidth control

The VPB control changes the installed bandwidth in the VPs, according to the offered traffic variation in order to eliminate the imbalance between the VP bandwidth and the offered traffic, improving in this way the end-to-end CBP of the switching pairs of the network.

3. VPB CONTROL

VPB, Dynamic Routing and bandwidth reservation controls drastically influence network resources and the global performance of ATM network under constraints posed by the bandwidth capacities of transmission links.

VPB control is illustrated in figure 8.2. It shows a small network of three ATM switches **(ATM-SW)** that are interconnected through one Cross-Connect System **(ATM-XC).** Suppose that this network has been designed perfectly and at the time of installation it satisfies the design Call-level QoS of 1% (end-to-end CBP). As time goes by, however, traffic changes and in one VP the CBP is high, while in other VPs the blocking remains low. To improve this network status, a VPB controller changes the initial bandwidth allocation in the network so as to reduce the maximum CBP of the network.

To rearrange the VP bandwidth dynamically the following types of VPB control schemes have been proposed:

a) Very Short-term control schemes based on the information of the concurrent connections in the VPs [7], with control interval less than 5 min.
b) Short-term control schemes based on the blocking measurements taken during the control interval, which ranges from several minutes to few hours [4].
c) Long-term control schemes based on traffic prediction with control interval ranging from a few hours to a few days [9,15].
d) Medium-term VPB control based on traffic measurements, with control interval ranging from several minutes to few hours [16].

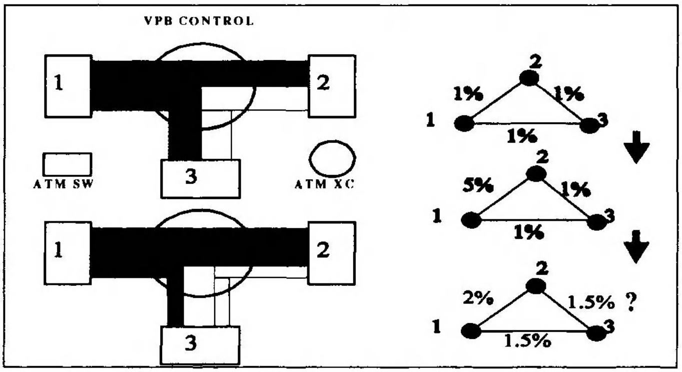


Figure 8.2. Virtual Path Bandwidth Control

The Very Short-term and the Short-term control must be distributed control schemes in order to respond fast to sharp traffic fluctuations and absorb them. To achieve this, they need very simple computations. They could ignore the traffic characteristics of service-classes [7], which is an important advantage in the B-ISDN environment. The Very-Short-term control achieves an optimal network performance. The implementation, however, of this control scheme is very difficult and, therefore, it is only of theoretical value. A large number of control steps are needed especially when the traffic volume is large. The Short-term control schemes are easier implemented but they lack optimality.

On the other hand, the Long-term control is a centralized control where the controller aims at an optimal network performance in the control interval by solving a large network optimization problem. However, the controller is based on the prediction of the offered traffic, which is a time consuming task, though it is not possible to be accurate. Therefore, the importance of the achieved optimality is weakened. The main advantage of the Long-term control schemes is that they can easily be implemented, because VP bandwidth is rearranged only a few times per day, at most.

The Medium-term VPB control scheme reconciles the advantages and disadvantages of the Short-term and Long-term control schemes. The controller must be a centralized one in order to optimize the network performance globally within its control interval. The control interval must be rather short in order to respond satisfactorily to medium-term traffic fluctuations. Short-term traffic fluctuations could be absorbed by the implementation of Dynamic Routing in a further stage [17]. To achieve this Medium-term VPB control, the controller is based on on-line measurements of the offered traffic.

4. OPTIMAL VPB ALLOCATION

4.1 NETWORK ARCHITECTURE

ATM-network architecture is considered in which each ATM-SW is accompanied by an ATM-XC system. The ATM-XCs are interconnected by a ring transmission line and compose the backbone network (Fig. 8.4a) [18]. This ATM-network architecture is similar to an existing STM-network architecture where there are digital cross-connect systems (**DCS**) instead of ATM-XCs. It has the advantage of simplicity and offers higher transmission line utilization [19]. It is worth mentioning that other network architectures could be considered as well, without important changes in the modeling of the optimization problem.

Thanks to the Virtual Path (VP) concept, the traffic management by reallocating the established bandwidth of the paths (VPB management) according to the traffic variations becomes favourable in ATM networks. The concept of VP, whereby two ATM-SWs face only the direct logical (imaginary) link (VP) between them, makes the structure of the backbone network transparent to the ATM-SW pairs. This is due to flexibility of the ATM-XCs to provide the required bandwidth in the end-to-end links of the ATM-SWs. Therefore, from the VPB management point of view, the whole ATM network is equivalent with a meshed network in which only the direct links are used (Fig. 8.4b). The transmission links are assumed bi-directional. A connection between ATM-SWs is established via any available path that has been registered in a table, called Routing Table **(RT).** Under the consideration of this study the route of a path between ATM-SWs passes through ATM-XCs only. This implies that the total amount of the buffer memory in the ATM-SWs should be involved into the constraint part of the optimization procedure, as it is an existing problem. However, it is not taken into consideration in order to reduce the problem complexity.

In the backbone network with a basic structure of Fig. 8.4a, two parts can be distinguished, in order to make the network study easier. The part of the network composed of the ATM-SWs and their direct connection to ATM-XCs, called outer network, and the part of the network composed of the interconnected ATM-XCs, called inner network.

The VPB (centralized) controller is located at an administrative center. It communicates with the ATM-SWs to collect the measurements of carried traffic and blocking during its control interval. Based on these measurements, it calculates the offered traffic. From the offered traffic, the installed bandwidth in the transmission links and the VPs listed in the RT, the VPB controller determines the distribution of the bandwidth to the VPs, by solving a large network optimization model. Then, it updates the data relevant to the VP bandwidth in the ATM-SWs. The realization of the produced VPB allocation is executed by the ATM-SWs simultaneously, after a delay due to the existing call-connections at the time point of bandwidth rearrangement [20]. The ATM-SWs increase, or decrease the number of cells, which have a specific Virtual Path Identifier [2] when the bandwidth of this VP is increased or decreased, accordingly. It is worth mentioning that, no communication between the VPB controller and the ATM-CXs is required.

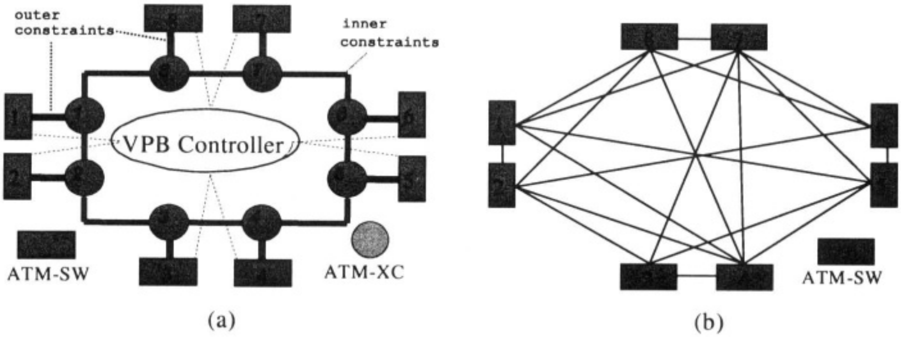


Figure 8.3. (a) ATM network architecture (b) VP connections network

4.2 NETWORK RELIABILITY

Reliable network under the consideration of bandwidth management means that in every switching pair if a transmission link failure occurs, bandwidth still remains. A reliability degree is the amount of the remaining bandwidth and the way it is distributed. Network reliability can be resulted by several schemes of bandwidth distribution to paths. As an example the following two schemes can be considered.

1. In a first scheme, we assume that for every pair of switches p at least two paths (VPs) exist between them and we enforce a certain percentage g_p of the total bandwidth V_p to be allocated to the shortest path. Logical values for g_p are in the range of 50% (most reliable) to 75% (less reliable). Although values of g_p in the range of 75% to 100% are problematic from the reliability point of view, they are permitted. The value g_p=100% means that there is no reliability on bandwidth allocation because the total bandwidth being assigned to each switching pair is allocated to only one path. The certain percentage g_p of the bandwidth, which is allocated to the shortest path, could be the same for all switching pairs (i.e. g_p=g) or could be fixed according to the degree of reliability we want to ensure for each switching pair individually. For instance, in order to guarantee best reliability between the switches A and B, g_p is set 50%, while between the switches A and C g_p might be 75%.

2. In a second scheme, we enforce the bandwidth to be distributed to the paths so that the allocated bandwidth to each path r becomes not less than a specific value q_r. Again, this value could be the same for all the paths (VPs) of the network or could be specialized as in the first scheme, so long as these specific values satisfy the constraints posed by the installed bandwidth in the transmission links. The value $q_r=0$ is also permitted.

The first scheme is preferable because the bandwidth allocation is clearly described.

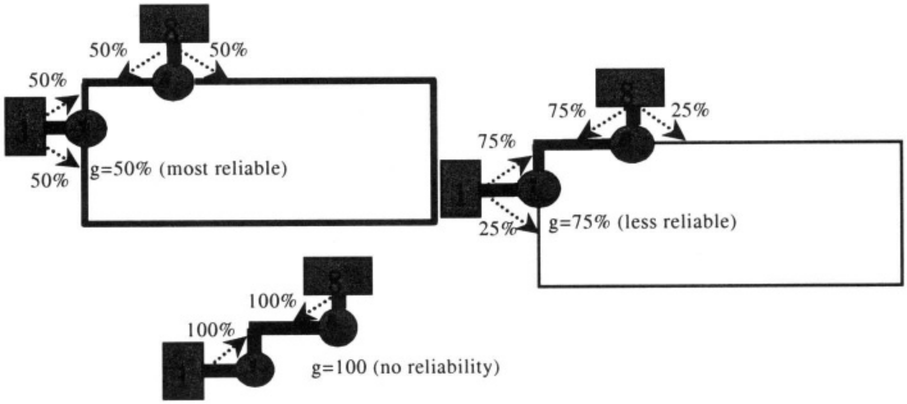


Figure 8.4. Network reliability from VPB management point of view

4.3 DEFINITION OF THE OPTIMIZATION PROBLEM

To set up the optimization problem mathematically the following notations are introduced:

- S Set of ATM-SWs.
- P Set of ATM-SW pairs.
- R Set of all paths between all ATM-SW pairs (listed in Routing Table).
- R* Set of the shortest-paths for all ATM-SW pairs ($R^* \subseteq R$).
- r Set of transmission links (sequence of nodes) defining a route r of a path ($r \in R$).
- R_p Set of available paths assigned to the ATM-SW pair p, ($p \in P$).
- R_s Set of paths where the ATM-SW s is either source or destination node ($s \in S$).
- Cs Installed bandwidth between the ATM-SW s and its accompanied ATM-XC ($s \in S$).

- L Set of bi-directional transmission links of the "inner" network.
- R_l Set of paths, which utilize the transmission link l ($l \in L$).
- C_l Installed bandwidth to transmission link l, ($l \in L$).
- W_r Bandwidth occupied by a path r between ATM-SWs (decision variables), ($r \in R$).
- A_p Traffic offered to ATM-SW pair p, ($p \in P$).
- B_p Call Blocking Probability for ATM-SW pair p.
- V_p Total Virtual Path Bandwidth of the switching pair p.
- g_p Percentage for definition of a reliability demand according to the reliability scheme (i), mentioned in section 4.2, ($p \in P$).
- q_r Bandwidth for definition of a reliability demand according to the reliability scheme (ii), mentioned in section 4.2, ($r \in R$).

The C_s is determined at the design phase of the network as:

$$C_s = \sum_{r \in R_s} W_r \qquad (8.1)$$

The C_l is determined at the design phase of the network as:

$$C_l = \sum_{r \in R_l} W_r \qquad (8.2)$$

The Virtual Path Bandwidth (VPB) of ATM-SW pair p, V_p, is the summation of the bandwidth occupied by all paths established for the ATM-SW pair p:

$$V_p = \sum_{r \in R_p} W_r \qquad (8.3)$$

The optimization problem is formulated as mathematical integer programming problem with the following linear constraints and the non-linear objective function:

– CONSTRAINTS

a) Due to the limited capacities of the ATM-SWs (outer network - outer constraints):

$$\sum_{r \in R_s} W_r \le C_s \quad \text{for all} \quad s \in S \qquad (48.)$$

b) Due to the limited bandwidth of the transmission links (inner network - inner constraints):

$$\sum_{r \in R_l} W_r \leq C_l \quad for \quad all \quad l \in L \tag{8.5}$$

c) Due to demand for reliability (according to the two bandwidth distribution schemes):

$$\begin{aligned} &i)\, W_r = g_p V_p \ \ for\, all\ r \in R^*,\ \ p \in P,\ or \\ &ii)\, W_r = q_r \ \ for\, all\ r \in R \end{aligned} \tag{8.6}$$

d) Concerning the decision variables:

$$W_r = n_r W_{unit} \tag{8.7}$$

where, n_r: non negative integer and W_{unit}: VP bandwidth unit.

Remarks:

1. The term outer and inner constraints for the C_s and C_l, respectively, are introduced due to their different influence on the VPB allocation. If the reason of the worst CBP is the capacity of a transmission link (inner constraint) it is easy to improve the performance of the network, by re-routing traffic through alternative routes and avoiding the congested link. However, if the reason is the capacity of an ATM-SW (outer constraint), this possibility does not exist. So, the worst CBP in case that it is only due to the outer constraints becomes independent of the configuration and the topology of the inner network.
2. In order for the reliability demand to be meaningful, we must assure that between ATM-SWs at least two paths are registered in the RT, that is $R^* \subset R$.
3. Not only the decision variables W_r but every notation which expresses bandwidth (C_s, C_l, $g_p V_p$ and q_r) must be an integer multiple of the W_{unit}.

– OBJECTIVE FUNCTION

$$\max \ of\, B_p = G(V_p, A_p) \Rightarrow \min \tag{8.8}$$

Where G stands for the function giving the CBP from the offered traffic $\mathbf{A_p}$ and the available bandwidth V_p of the ATM-SW pair p.

In the STM environment where only one service-class exists, G is the well-known Erlang B-Formula. Whereas in the environment of ATM networks, where at least two service-classes are assumed sharing a VP equally, the calculation of CBP can be done by using the recurrent formula given in Ref. [21,22]:

$$G(i) = \left\langle \begin{array}{c} 1 \quad for\ i = 0 \\ \frac{1}{i}\sum_{k=1}^{K} a_{c_k} b_{c_k} G(\ i - b_{c_k}\ )\ for\ i = 1,\ldots, V_p \\ 0 \quad otherwise \end{array} \right\rangle \tag{8.9}$$

Where K is the number of service-classes serviced by the ATM network, b_{ck} is the required bandwidth per call of the service-class c_k and a_{ck} is the offered traffic of the service class c_k to the switching pair p, that is, $\mathbf{A_p}$ is a K-size array with elements the a_{ck}'s. $\mathbf{B_p}$ is a K-size array, too.

The Call Blocking Probability B_{pck} of the ATM-SW pair p for the service-class c_k, is defined as:

$$B_{pc_k} = \sum_{j=1}^{b_{c_k}-1} G^{-1}G(V_p - j), \text{ where } G = \sum_{i=1}^{V_p} G(i) \tag{8.10}$$

In the above formula, bandwidth (trunk) reservation schemes are not incorporated. According to the trunk (bandwidth) reservation concept [23], calls of service class c_k are refused for service when less than $t(c_k)$ bandwidth units remain available in the VP. By selecting properly the numbers $t(c_k)$, it is possible to meet the same grade of service among the service classes and so, the worst CBP of the whole network can be improved. To incorporate the Bandwidth Reservation control to VPB control, a good approach is found in Ref. [24]. For calculation of CBP, the following modifications are introduced to the above expressions.

Equation 8.9 must be modified for i = 1,...,V_P as:

$$G(\ i\ ) = \frac{1}{i}\sum_{k=1}^{K} a_{c_k}\ D_{c_k}(i - b_{c_k})G(i - b_{c_k}) \tag{8.11}$$

$$\text{Where } D_{c_k}(i - b_{c_k}) = \left\langle \begin{array}{l} b_{c_k}\ \ for\ i \le V_p - t(c_k) \\ 0\ \ for\ i > V_p - t(c_k) \end{array} \right\rangle \tag{8.12}$$

And equation 8.10, because of the upper limit of the summation, as:

$$B_{pc_k} = \sum_{j=1}^{b_{c_k}+t(c_k)-1} G(V_p - j)G^{-1} \tag{8.13}$$

By the above approximation, the accuracy of CBP calculation is satisfactory especially when the differences of holding-times of the service classes are small [3].

The formulas for calculating CBP in ATM networks are perfectly fixed for CBR service-classes. For VBR service-classes the constant traffic load required by the above formulas may correspond to the sustainable cell rate (**SCR**) or to the notion of effective bit rate (equivalent bandwidth) [25]. However, for the rest ATM service-classes, as they are the available bit rate (**ABR**) and the unspecified bit rate (**UBR**) service classes, the CBP calculation is an open problem, since even the notion of blocking has to be reconsidered [26].

4.4 OPTIMAL SOLUTION

The network optimization problem has been formulated as a non-linear integer-programming problem. Obviously, the main difficulty in its solution consists in the non-linearity of the objective function. Besides, a difficulty arises from the constraints set c (equation 8.6), in the first case of demand for reliability where the right-hand-side values are not constant, as they are in all the other constraints. One more difficulty arises from the demand for integer values for the decision values.

To solve the above model analytically the analytical method proposed and proved by Prof. M. Akimaru [14] is followed in general. In the following, it is described how this method is used to overcome the above difficulties and achieve the optimal solution to the present optimization problem; that is, how the bandwidth allocation that minimizes the network's worst CBP is defined. The following approach transforms the non-linear optimization model to a succession of linear integer programming models, in four steps:

Step 1: Calculate the initial worst and minimum CBP of the network x_{max} and x_{min} respectively, using the function G and based on the initial bandwidth allocation and the traffic demand matrix (it is valid to assume that initially x_{max}=l and x_{min}=0).

Step 2: Define a new improved worst CBP as: $x_{new}=(x_{max}+x_{min})/2$.

Step 3: Find out whether the value x_{new} can stand for the worst CBP or not, by the following way:

- With the aid of a function (let us call it G*) which determines bandwidth from the offered traffic and a given grade of service, calculate the V_p based on A_p and by using as grade-of-service the x_{new} for all $p \in P$.

The bandwidth V_p is calculated through G* so as to be an integer multiple of W_{unit}. Because of the constraint set d (equation 8.7) and the third remark above, the integer multiples of W_{unit} are referred in the following by using brackets, i.e. $[V_p]$ stands for the integer value V_p/W_{unit}, $[W_r]$ for n_r.

- Distribute all V_p's to the VPs (i.e. define W_r) under the constraints posed by the installed bandwidth to the transmission links (constraint sets a, i.e. equation 8.4, and b, i.e. equation 8.5, and according to the reliability scheme (constraint set c, i.e. equation 8.6, i or ii).

  The difficulty arisen from the constraint set c (i) does not exist any more, because the V_p has been defined (g_p is parameter). So, the variables W_r can be defined through the solution of the following set of equations:

  - If W_s stands for the possible free bandwidth between the ATM-SW s and its corresponding ATM-XC, then according to constraint set a,

$$\sum_{r \in R_s} [W_r] + [W_s] = [C_s] \quad for\ all\ \ s \in S \tag{8.14}$$

  - If W_l stands for the possible free bandwidth of transmission link 1, then, according to constraint set b,

$$\sum_{r \in R_l} [W_r] + [W_l] = [C_l] \quad for\ all\ \ l \in L \tag{8.15}$$

  - Constraint set assuring the grade-of-service x_{new}

$$\sum_{r \in R_p} [W_r] = [V_p] \quad for\ all\ \ p \in P \tag{8.16}$$

  - Introducing the variable W_{qr} which expresses the surplus bandwidth for path r over the demanded value of q_r, then for the constraint sets c,

$$\begin{aligned} &i)\ [W_r] = [g_p V_p]\ for\ all\ r \in R^*,\ p \in P,\ \ or \\ &ii)\ [W_r] - [W_{qr}] = [q_r]\ forall\ r \in R \end{aligned} \tag{8.17}$$

  - Furthermore,

$$[W_s] \geq 0,\ [W_l] \geq 0,\ [W_{q_r}] \geq 0,\ [W_r] \geq 0\ (integers) \tag{8.18}$$

To solve this set of equations, it is considered as the constraint part of an optimization problem with a linear objective function, which is artificially introduced:

$$\sum_{r \in R} W_r \Rightarrow max \tag{8.19}$$

Thus, a linear integer-programming problem results, which can be solved by classic integer programming techniques. Due to inconvenience of the commercial software packages that support optimization problems with integer stipulations, an iterative algorithm is implemented in FORTRAN, based on the well-known Simplex method. This is the "primal cutting-plane" algorithm [27]. It guarantees convergence and satisfies throughout (in every iteration) the linear restrictions and the integer stipulations. In addition, the computational technique of Big M method [27] is applied to ensure the equalities in the constraint sets that assure the grade-of-service x_{new} and the first scheme for network reliability.

If the so formulated integer programming model has a feasible solution it means that all V_p's are distributed to the paths and x_{new} can stand for the new worst CBP; then put $x_{max} = x_{new}$. Otherwise put $x_{min} = x_{new}$.

Step 4: Repeat the procedure from the second step until the difference x_{max}-x_{min} becomes equal or less than an error e which expresses the accuracy by which we want to estimate the network's worst CBPs (e=0 is valid).

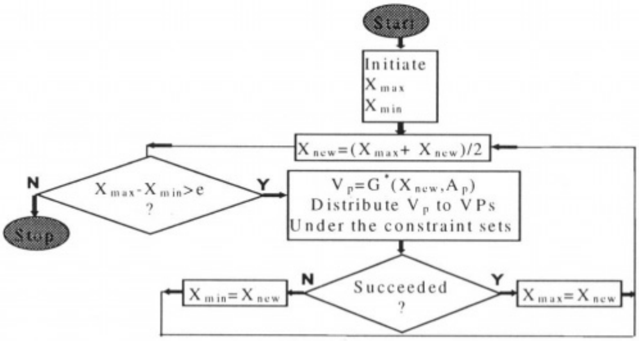


Figure 8.6. Flow-chart of the optimization procedure

Concisely, this algorithm is presented in the flow chart of Fig. 8.6. A remaining problem in solving such a model is the huge computer memory that is required for large networks. For a ring type network of N ATM-SWs (Fig. 8.4a), to set up the optimization procedure, N^2 constraint equations and 2N(N-1) variables are required. Regarding the CPU-time, setting properly the initial values of x_{max} and x_{min}, considerable time could be saved if the worst CBP could be estimated approximately.

5. APPLICATION EXAMPLES

Two application examples of optimal VPB are considered. The first example is for tutorial purposes. The second example presents the efficiency of the optimal VPB Control on the performance of a realistic ATM network.

5.1 TUTORIAL EXAMPLE

The optimal VPB control is applied on the 3-node network of figure 8.2. For simplicity, the network accommodates one service-class, which is the telephone service and it has been designed (dimensioned) so as to satisfy the Call-level QoS of 1% (CBP) for all switching (node) pairs. The designed traffic-load is 37 ERL for each switching pair per traffic-flow direction. The required number of trunks per VP is 49 (VP capacity); it results through the Erlang B-Formula. In practice, however, the trunks are provided in bundles. As an example, if a bundle of trunks consist of 5 trunks, the VP capacity becomes 50 trunks. So, initially, the CBP for all switching pairs is 0.73% (<1%). Considering that 1 trunk corresponds to the bandwidth unit, W_{unit}, and W_{unit}=64 Kbps, the resultant bandwidth-capacity for each VP is 3.2 Mbps. Note that only one VP per traffic-flow direction is established for each switching pair, that is, $V_p=W_r$ (e.g $V_{(1,\ 2)}=W_{(1\ \ 4\ \ 2)}=50$, i.e. there is one Virtual Path between the switches 1 and 2, p=(1,2) and is routed on the path r=(1 4 2)). Also, due to the topology of this network, $C_S=C_l$. The transmission links are fully occupied (used) and have the following capacities:

$$C_1=C_{(1,4)}= V_{(1,2)}+V_{(2,1)}+V_{(1,3)}+V_{(3,1)}=W_{(1\ 4\ 2)}+W_{(2\ 4\ 1)}+W_{(1\ 4\ 3)}+W_{(3\ 4\ 1)}= 200$$
$$C_2=C_{(2,4)}= V_{(2,1)}+V_{(1,2)}+V_{(2,3)}+V_{(3,2)}=W_{(2\ 4\ 1)}+W_{(1\ 4\ 2)}+W_{(2\ 4\ 3)}+W_{(3\ 4\ 2)}= 200$$
$$C_3=C_{(3,4)}= V_{(3,1)}+V_{(1,3)}+V_{(3,2)}+V_{(2,3)}=W_{(3\ 4\ 1)}+W_{(1\ 4\ 3)}+W_{(3\ 4\ 2)}+W_{(2\ 4\ 3)}= 200$$

As time goes by, traffic changes and some switching pairs meet a worse QoS while some other switching pairs meet a better QoS. Tables 8.1a and 1b show the new traffic load and the corresponding end-to-end CBP of the network, respectively. The RT with the VP capacities (in bandwidth units, that is, in trunks) is shown in Table 8. 1c.

A centralized VPB controller will reallocate the VP bandwidth so as to minimize the worse CBP (5.41%) of the network by applying the following optimization procedure (the first repetition is presented in detail). It is assumed that the bandwidth reallocation unit is 320 Kbps (i.e. 5 trunks). The error e is defined to be e=0.00001.

Table 8.1. Before VPB Reallocation: (a) Traffic in Erlang, (b) End-to-end Call Blocking Probabilities (%), (c) Routing Table with VP capacities

	1	2	3
1	0	45	37
2	30	0	30
3	45	37	0

(a)

	1	2	3
1	0	5.41	0.73
2	0.02	0	0.02
3	5.41	0.73	0

(b)

RT(r)	W_r
1 4 2	50
1 4 3	50
2 4 1	50
2 4 3	50
3 4 1	50
3 4 2	50

(c)

1st Repetition

Step 1: x_{max}=0.0541 and x_{min}=0.0002 (CBPs)

Step 2: x_{new}=0.0272

Step 3: For each switching pair p, calculate the bandwidth $[V_p]$ which guarantee the grade-of-service x_{new}. To distribute the $[V_p]$ to VPs the following integer programming model must be solve (the $[V_p]$ appears in the right-hand-side of equations 8.4 - 8.9):

– CONSTRAINTS

Since there is no distinction in outer and inner network the constraint sets a, and b coincide into one set consisted of equations 8.1 to 8.3. Regarding the network reliability it is assumed that g_p=100% (no reliability) for all the switching pairs.

1. $[W_{(1\ 2)}]+[W_{(1\ 4\ 3)}]+[W_{(2\ 4\ 1)}]+[W_{(3\ 4\ 1)}]+[W_1]=200$ for the transmission link (1,4)
2. $[W_{(1\ 4\ 2)}]+[W_{(2\ 4\ 1)}]+[W_{(2\ 4\ 3)}]+[W_{(3\ 4\ 2)}]+[W_2]=200$ for the transmission link (2,4)
3. $[W_{(1\ 4\ 3)}]+[W_{(2\ 4\ 3)}]+[W_{(3\ 4\ 1)}]+[W_{(3\ 4\ 2)}]+[W_3]=200$ for the transmission link (3,4)
4. $[W_{(1\ 4\ 2)}] + [W_{(1,2)}] = 55$ for the switching pair (1,2)
5. $[W_{(1\ 4\ 3)}] + [W_{(1,3)}] = 50$ for the switching pair (1,3)
6. $[W_{(2\ 4\ 1)}] + [W_{(2,1)}] = 40$ for the switching pair (2,1)
7. $[W_{(2\ 4\ 3)}] + [W_{(2,3)}] = 40$ for the switching pair (2,3)
8. $[W_{(3\ 4\ 1)}] + [W_{(3,1)}] = 55$ for the switching pair (3,1)
9. $[W_{(3\ 4\ 2)}] + [W_{(3,2)}] = 50$ for the switching pair (3,2)

– ARTIFICIAL OBJECTIVE FUNCTION

$$[W_{(1\ 4\ 2)}] + [W_{(1\ 4\ 3)}] + [W_{(2\ 4\ 1)}] + [W_{(2\ 4\ 3)}] + [W_{(3\ 4\ 1)}] + [W_{(3\ 4\ 2)}] \rightarrow \text{MAX}$$

Where all the variables must be non-negative integers. To solve this model, the primal cutting-plane algorithm is applied. The variable

$[W_1]$, $[W_2]$ and $[W_3]$ are considered slack variables. The variables $[W_{(1,2)}]$, $[W_{(1,3)}]$, $[W_{(2,1\}}]$, $[W_{(2,3)}]$, $[W_{(3,1))}]$ and $[W_{(3,2)}]$ are considered artificial variables on which the Big-M method must be applied so as to be assigned a zero value (if possible) to these variables. The solution of this model is feasible and therefore:

x_{max} becomes 0.0272 (x_{max}=0.0272) and

x_{min} remains 0.0002 (x_{min}=0.0002)

Step 4: Since the difference x_{max}-x_{min} > e this procedure is repeated from Step 2.

The results of each repetition are presented in Table 8.2.

Table 8.3, presents the network status after the VPB reallocation. Table 8.3a presents the occupied bandwidth in each transmission link. It shows that the transmission link C_1, i.e. (1,4), is fully used while the free bandwidth in transmission links C_2 and C_3 is 15 and 5 bandwidth units. Table 8.3b shows the end-to-end CBPs of the network. Comparing Table 8.1b with Table 8.3b, it results that the maximum CBP of the network has been reduced from 5.41% to 2.03% and that the switching pairs (2,1) and (2,3) have been paid for this improvement. Table 8.3c shows the RT with the final VP capacities.

Table 8.2. Intermediate results of the optimal VPB controller

Repetition	x_{max}	x_{min}	x_{new}	DISTRIBUTION
1	5.41%	0.02%	2.72%	SUCCEEDED
2	2.72%	0.02%	1.37%	NO
3	2.72%	1.37%	2.04%	SUCCEEDED
4	2.04%	1.37%	1.71%	NO
5	2.04%	1.71%	1.87%	NO
6	2.04%	1.87%	1.96%	NO
7	2.04%	1.96%	2.00%	NO
8	2.04%	2.00%	2.02%	NO
9	2.04%	2.02%	2.03%	SUCCEEDED
10	2.03%	2.02%	2.03%	NO
11	2.03%	2.03%	2.03%	NO
12	2.03%	2.03%	2.03%	NO
13	2.03%	2.03%	2.03%	NO

Table 8.3. (a) Used bandwidth of the transmission links,.(b) End-to-end Call Blocking Probabilities (%), (c) Routing Table with VP capacities.

Link	Band-width
C_1	200
C_2	185
C_3	195

(a)

	1	2	3
1	0	2.03	0.73
2	1.44	0	1.44
3	2.03	0.73	0

(b)

W_r	RT (r)
55	1 4 2
50	1 4 3
40	2 4 1
40	2 4 3
55	3 4 1
50	3 4 2

(c)

5.2 REALISTIC APPLICATION EXAMPLE-PERFORMANCE OF VPB CONTROL

The optimization procedure for VPB allocation is applied to a model ATM-network of 8 ATM-SWs (nodes) and 8 ATM-XCs with a ring topology (Fig. 8.4a). Although this topology seems to be a simple one, it is the worse case from the bandwidth management viewpoint. For instance, if the VBP controller tries to allocate bandwidth to the longer path between ATM-SW 1 and 2, affects the performance of all the other switching pairs.

For presentation purposes, the network accommodates two service-classes only. The required bandwidth per call of the 1st service is 64 Kbits/sec and of the 2nd service is 1.536 Mbits/sec. Calls of both service-classes arrive according to a Poisson process with exponentially distributed holding-times of the same average value. As an example, the 1st service-class could correspond to the telephone service while the 2nd to a video service. The VPs are shared equally by the calls of the service-classes, while the bandwidth reservation scheme is applied to them, so that their resultant blocking within the VPs of a switching pair is equalized. This can be achieved by equalizing the required bandwidth per call among the service-classes, i.e. within the VPs of each switching pair (distinguishing the traffic-flow directions) a bandwidth of 1.472 Mbits/sec must be reserved for benefit of the 2nd class. This equalization procedure is in harmony with the optimization criterion of minimizing the worst CBP of all switching pairs.

The network is dimensioned so as to satisfy the Call-level grade of service of 3%. The same traffic load for all ATM-SW pairs is considered. The traffic of the 1st service is 500 ERL and for the 2nd service is 25 ERL, in each flow direction. Bandwidth is allocated to the VPs in units of 1.536 Mbits/sec, which is also the bandwidth rearrangement unit for the VPB management. The same bandwidth unit is assumed in dimensioning the backbone network so as, initially, the transmission links is fully utilized. This is done in order to evaluate readily bandwidth distribution schemes assuring different degrees of network reliability. Bandwidth distribution schemes of the first type only (section 4.2) are examined where the percentage g_p is the same for all the switching pairs p (g_p=g).

So, initially the CBP for every ATM-SW pair p is 2.78% and each V_p is 79.872 Mbits/sec. Regarding reliability, if g=50% the bandwidth of 79.872 Mbits/sec is equally shared between the shortest path and the unique alternative path. If g=100% it means that only the shortest path is used. If g=70% the shortest path has a bandwidth of 56.832 Mbits/sec and the alternative path 23.040 Mbits/sec, etc.

The bandwidth-capacity of a transmission link is calculated as the sum of bandwidth of those paths whose the route passes through this transmission link. One rule that is followed in the formation of RT is to convey the traffic of both flow directions through the same path between two nodes.

Considering bi-directional transmission links, for the backbone network of Fig. 8.4a, every transmission link between the ATM-SWs and the ATM-XCs (outer network) has a bandwidth-capacity of 1118.208 Mbits/sec irrespectively from reliability. In the transmission links between the ATM-XCs (inner network) though, the bandwidth-capacity is highly dependent on the desired degree of reliability (i.e. on g).

Fig. 8.7 presents the total bandwidth increment in the inner network as percentage of the installed bandwidth when g=100% (then the total bandwidth of the inner network is 10.2 Gbits/sec), versus various reliability degrees. It shows that a considerable amount of bandwidth is required, in order to increase reliability.

Fig. 8.8 presents the performance of the optimal VPB management when the offered traffic fluctuates randomly according to the uniform distribution by a maximum of 20% to 100% (in steps of 20%). These results are valid in the examined network for all reliability degrees considered in the design phase of the network, because the resultant worst CBP depends on the installed bandwidth in the outer network. More precisely, when the network is designed with a reliability degree (g=50%, 60%, 70%) the same optimal values of worst CBP are achieved even if, afterwards, we retain the same reliability degree or decrease it (e.g. g=75%). An exception occurs in one case where the designed g is 70% and in order to achieve the best performance it is necessary to reduce the reliability (to g=75%).

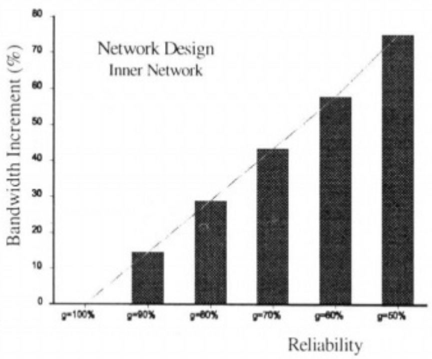


Figure 8.7. Required bandwidth for network reliability.

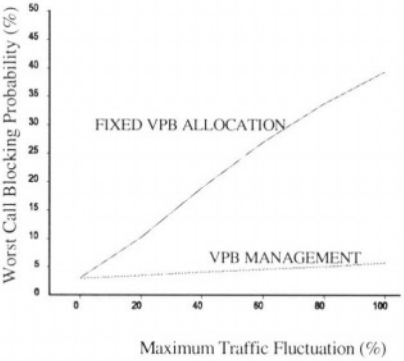


Figure 8.8. Performance of VPB management.

Fig. 8.9 shows the performance of the optimal VPB management in the case where the desired reliability degree is even greater than that of the designed reliability degree. Likewise, the performance of VPB management depends much more on the reliability degree than on the maximum traffic fluctuation (legend of figure 8.9).

Figure 8.10 shows the transmission links utilization when the offered traffic fluctuates and the network is designed for best reliability. The throughput is measured, i.e. the used bandwidth as the percentage of the installed bandwidth (that is, the ratio of the total VP bandwidth to the total transmission capacity). It is shown how much the throughput decreases especially in the inner network, when the desired reliability degree decreases, in comparison to the design reliability under the same traffic load (from g=50% to g=70%).

The required computer memory to manage the ATM-network of Fig. 8.4a, in terms of VAX/VMS, is 670 Kbytes of Peak-Working-Set size and maximum CPU-time, running in MicroVAX-3110 about 10 min.

The complexity of the network optimization procedure as a function of network size is presented in figure 8.11. It shows how the CPU-time and memory increase versus the number of ATM-SWs (ring-type networks, like Fig. 8.4). The measurements have been taken from MicroVAX 3110, when the initial traffic load for all ATM-SW pairs is as in the case of network of figure 8.4 and fluctuates uniformly by a maximum of 100%. No reliability constraints are assumed. It is worth mentioning that in a modern computer systems the CPU-time will decrease about 10 times.

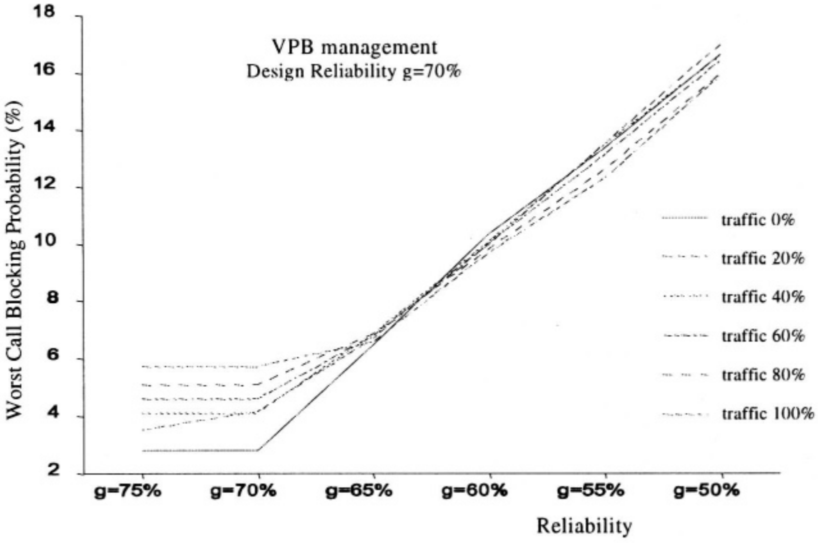


Figure 8.9. VPB management performance when the desired reliability increases.

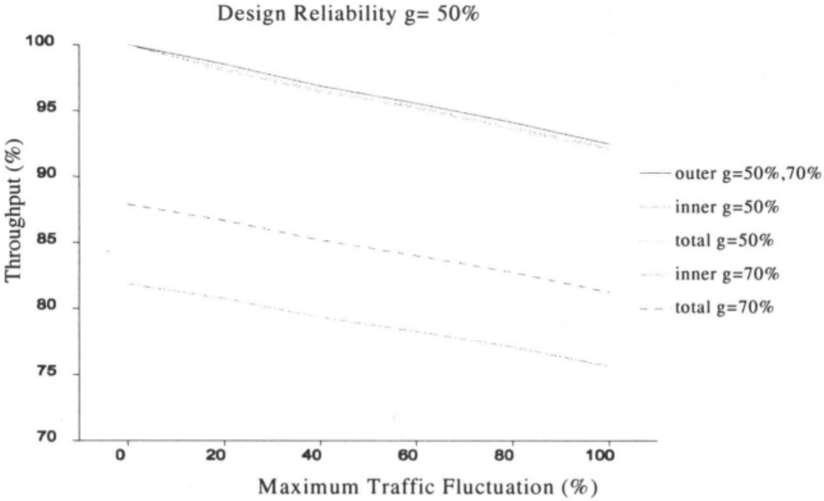


Figure 8.10. Bandwidth utilization of the most rteliable network design

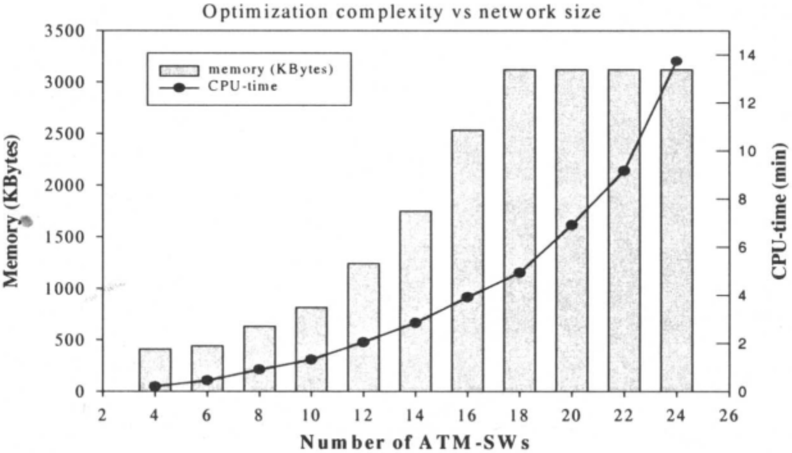


Figure 8.11. Computer memory and CPU-time of the optimization procedure as a function of network size.

6. SUMMARY

Optimal resource management results from traffic management, which has a layered architecture in ATM networks. The Call-level and the Cell-level traffic controls are surveyed. The impact of VPB control on managing the network resources is presented. The paper points at the importance of the optimal VPB control and presents ATM network architecture that is appropriate for VP bandwidth management. The network consists of ATM cross-connect systems, for readily reallocation of the VP bandwidth. Especially to ensure network reliability huge bandwidth is required. Therefore, the necessity of the optimal bandwidth management becomes more essential. The technological progress gives us the possibility of a global network optimization. A rigorous and analytical procedure is presented for solving the formulated non-linear integer programming optimization problem, by transforming it into a sequence of linear integer programming models and applying classic techniques of the operation research. In a tutorial example, the optimal VBP control is applied on a very small network accommodating a single service-class in order for the steps of the optimization procedure to be clarified. As a realistic application example, a model ATM network is considered. In figures, the performance of VPB control and the throughput of the model network are shown when the offered traffic fluctuates and the desired degree of reliability (at the design phase of the network or afterwards) varies.

Acknowledgments

The author is grateful to Professor G. Kokkinakis (University of Patras/Greece) and Mr. S. Shioda (NTT/Japan) for the knowledge they gave to him to write this paper.

References

[1] K. Mase and H. Yamamoto, “Advanced traffic control methods for network management”, IEEE Commun. Mag., Vol. 28, No 10, 1990.

[2] H. Saito, K. Kawashima and K. Sato, “Traffic Control Technologies in ATM Networks” IEICE Trans., Vol. E 74, pp. 761-771, Apr. 1991.

[3] M. Logothetis and S. Shioda, “Centralized Virtual Path Bandwidth Allocation Scheme for ATM networks”, IEICE Trans. Commun., Vol. E75-B, No. 10, Oct. 1992.

[4] S. Shioda, H. Uose, “Virtual Path Bandwidth Control. Method for ATM-Networks: Successive Modification Method”, IEICE Trans., Vol. E 74, pp. 4061-4068, Dec 1991.

[5] M. Logothetis, S. Shioda, G. Kokkinakis, “Optimal Virtual Path Bandwidth Management Assuring Network Reliability”, in Proc. ICC'93, Geneva, 1993.

[6] S. Shioda, H. Uose, “Virtual Path Bandwidth Control for ATM Networks: Batch Modification Method”, IEICE Trans. Vol. J75-B-I, No. 5, May 1992.

[7] S. Ohta, K. Sato and I. Tokizawa: “A dynamically controllable ATM transport network based on the Virtual Path Concept”, Proc. GLOBECOM'88, pp. 1272-1276, 1988.

[8] Gerla M, Monteiro J.A.S and Pazos R., “Topology and Bandwidth Allocation in ATM Nets”, IEEE J. Selec. Areas in Commun., Vol7, No. 8, pp. 1253-1262, Oct. 1989.

[9] J.A.S. Monteiro and M. Gerla, “Topological Reconfiguration of ATM networks”, Proc. GLOBECOM 1990.

[10] G. Gopal, C. Kim and A. Weinrib, “Algorithms for reconfigurable networks”, Proc. ITC-13, 1991.

[11] M. Logothetis “Centralized Path Bandwidth Control through Digital Cross-Connect Systems”, IEICE Technical Report, Vol. 91, No 381, IN91-122, 1991.

[12] M. Logothetis and G. Kokkinakis: “Optimal computer-aided capacity management in digital networks”, Proc. EURINFO 88, Athens, 1988.

[13] K. Mase, M. Imase, “An adaptive Capacity Allocation Scheme in Telephone Networks”, IEEE Trans. on Commun., Vol. COM-32, Feb. 1982.

[14] M. Akimaru, “Variable communication network design”, Proc. ITC-9, 1979.

[15] M. Logothetis, G. Kokkinakis “Network Planning Based on Virtual Bath Bandwidth Management”, International Journal of Communications Systems, No 8, Aug. 1995.

[16] M. Logothetis and S. Shioda, “Medium-Term Centralized Virtual Path Bandwidth Control Based on Traffic Measurements”, IEEE Trans. on Commun. Vol. 43, Oct. 1995.

[17] I.Z. Papanikos, M. Logothetis and G. Kokkinakis, “Virtual Path Bandwidth Control versus Dynamic Routing Control”, in *ATM Networks: Performance Modeling and Evaluation,* Vol.2, (Ed. D. Kouvatsos), Chapman & Hall, London, 1996.

[18] K. Sato, S. Ohta and I. Tokizawa, “Broad-Band ATM Network Architecture Based on Virtual Paths,” IEEE Trans. Commun., Vol. COM-38, pp. 1212-1222, Aug. 1990.

[19] H. Obara, M.Sasagawa and I. Tokizawa, “An ATM Cross-Connect System for Broadband Transport Networks Based on Virtual Path Concept”, Proc. GLOBECOM'90, 1990.

[20] M. Logothetis and G. Kokkinakis, “Influence of Bandwidth Rearrangement Time on Bandwidth Control Schemes”, Proc. 4th International Conference in Commun. & Control, COMCON4, Rhodes/Geece 1993.

[21] J. S. Kaufman, “Blocking in a Shared Resource Environment”, IEEE Trans. Comm., Vol. COM-29, October 1981.

[22] M. Schwartz, B. Kraimeche, “An Analytic Control Model for an Integrated Node”, Proc. INFOCOM 1983.

[23] T. Oda, H. Fukuoka and Yu Watanabe, “Comparison of Traffic Characteristics of GOS Control Methods for a Trunk Group Carrying Multislot Calls”, Electronics and Communications in Japan, Part 1, Vol. 73, No. 7, 1990.

[24] J. W. Roberts, “Teletraffic models for the Telecom 1 Integrated Services Network”, Proc. ITC-10, 1982.

[25] G. de Deciana, G. Kesidis and J. Walrand, “Resource Management in Wide-Area ATM Networks Using Effective Bandwidths”, IEEE J. Select. Areas in Commun., Vol. 13, No 6, pp. 1081-1089, Aug. 1995.

[26] G. Fodor, A. Racz, S. Blaabjerg, “Simulative Analysis of Routing and Link Allocation Strategies in ATM Networks Supporting ABR

Services", IEICE Trans. Commun. Special Issue on ATM Traffic Control and Performance Evaluation, pp. 985-995, Vol. E81-B, No. 5, May, 1998.

[27] H.W. Wagner, *Principles of Operations Research,* Prentice Hall, 1969.

Michael D. Logothetis was born in Stenies, Andros, Greece, in 1959. He received the Dipl.-Eng. and Ph.D. degrees in electrical engineering, both from the University of Patras, Patras/Greece, in 1981 and 1990 respectively. From 1982 to 1990, he was a Teaching and Research Assistant at the laboratory of Wire Communications, University of Patras, and participated in many national research programs and three EEC projects (ESPRIT, LRE), dealing with telecommunication networks, as well as with natural language processing. From 1991 to 1992, he was Research Associate in NTT's Telecommunication Networks Laboratories. From 1992 to 1996, he was a Lecturer in the Department of Electrical & Computer Engineering of the University of Patras and since 1996 he is an Assistant Professor in the same university. His research interests include traffic control, network management, simulation and performance optimization of telecommunication networks. He is a member of the IEEE (Commun. Society - CNOM), IEICE and the Technical Chamber of Greece (TEE).

Chapter 9

ATM MULTICAST ROUTING

Gill Waters
University of Kent at Canterbury
Canterbury, Kent, CT2 7NZ,
England
A.G.Waters@ukc.ac.uk

John Crawford
University of Kent at Canterbury
Canterbury, Kent, CT2 7NZ,
England
J.S.Crawford@ukc.ac.uk

Abstract Several multicast routing heuristics have been proposed to support multimedia services, both interactive and distribution, in high speed networks such as B-ISDN/ATM. Since such services may have large numbers of members and have real-time constraints, the objective of the heuristics is to minimise the multicast tree cost while maintaining a bound on delay. They should also be fast to compute and may need to be suitable for dynamic groups.

We present an introduction to the problem and some key heuristic solutions and compare their performance. We show that the specific efficiency of a heuristic solution depends on the topology of both the network and the multicast, and that it is difficult to predict.

Because of this unpredicatability, we propose the integration of two heuristics with Dijkstra's shortest path tree algorithm to produce a hybrid that consistently generates efficient multicast solutions for all possible multicast groups in any network. The hybrid shows good performance over a wide range of networks, (both flat and hierarchical) and multicast groups, within differing delay bounds. We also discuss how heuristics can be deployed within the PNNI framework and briefly examine other issues related to multicast routing and PNNI.

Keywords: routing, multicast, Steiner tree, Quality of Service, delay constrained tree, algorithms, heuristics, PNNI

1. INTRODUCTION

Many of the new services envisaged for ATM networks involve point to multipoint connections. Distribution services, such as video on demand or continuous information publishing services, are likely to have large numbers of customers. Interactive services such as multimedia conferencing, co-operative working and educational applications can also be well supported by multicasting. ATM offers the integration of data and real-time components such as audio and video. This implies that, for many multicast services on ATM networks, the network must make appropriate Quality of Service (QoS) provision particularly in terms of maintaining agreed bandwidth and minimising delay. Because of the potentially large numbers of users, routing of multicast connections is an important issue.

Multicast routing for ATM should respond to QoS requirements, make efficient use of the network, be fast to compute, stable for dynamic groups and cater for sparse and dense groups. Efficiency is gained by not transmitting replicated cells down any link and by choosing a cost-effective multicast tree.

The ATM Forum's Private Network Node Interface (PNNI) is emerging as the most important technique for organising large interconnected ATM networks, both public and private (The ATM Forum Technical Committee, 1996). Routing for PNNI is based on link-state information as are the techniques we discuss in detail in this paper. The hierarchical nature of PNNI has scalability advantages achieved by constraining the amount of state information stored by switches and reducing the number of signalling messages. On the other hand, because information on delays and bandwidth is aggregated for use outside each peer group, there is a resultant loss in the accuracy of the routes calculated. This aspect and related work on PNNI multicast routing will be discussed later in the paper.

Our discussion concentrates on graph-theoretical heuristics for multicast routing which combine bounded delay with efficient use of the network, for large-scale real-time multicast services. For networks with n nodes, the lowest delay from a source to each of the other nodes can easily be found in $O(n^2)$ time using Dijkstra's algorithm. The paths found in the process form a broadcast tree which can be pruned beyond the receiving group members. Provided all of the destinations are reachable within the delay bound this offers a satisfactory solution. Where the predominant requirement is efficiency, the total cost of a broadcast tree can be found using techniques such as Prim's or Kruskaal's algorithms. However, the equivalent problem for a proper subset of the nodes of the network is known as the Steiner tree problem which is NP-complete, although heuristics are available which give reasonable solutions. Finding a multicast routing tree which is both efficient and delay bound is also an NP-complete problem.

We discuss and evaluate a number of heuristic techniques for finding such multicast trees. Each link in the networks used in the evaluations has two metrics: cost and delay. The cost metric represents a number of possibilities including the monetary cost of using the link, a parameter related to residual available bandwidth or a value proportional to the length of the link. Delay is taken as a constant for the purpose of calculation, since a multicast tree will generally be set up for the duration of a virtual channel. The fixed value includes an expected component for queueing as well as the fixed switching, transmission and propagation delays. QoS queuing and traffic shaping are likely to reduce the variability of queueing delays experienced in the switches.

The problem of arbitrary delay bound low cost multicast routing in networks was first addressed by Kompella, Pasquale and Polyzos in (Kompella et al., 1993). Evaluation of their work and a number of other proposed solutions (Waters and Crawford, 1996), (Salama et al., 1995) shows that on average these heuristics perform well. Our detailed analysis and evaluation of some of these heuristics shows that there is a wide variance in the efficiency of their solutions, especially considering multicast group size relative to the size of the network. We propose a hybrid, combining two heuristics based on Dijkstra's algorithm that produces reasonably consistent and efficient solutions to the multicasting problem, with an acceptable order of time complexity, for all possible multicast groups in any network.

The rest of this paper is organised as follows. In section 2. we define the bounded delay minimum cost multicast routing problem. In section 3. we describe and assess three heuristics and consider them as candidates for integration. Section 4. describes the network model, benchmark algorithms and arbitrary delay bound we use to evaluate both the candidate heuristics and the hybrid. The candidate heuristics are evaluated in Section 5. Section 6. describes the hybrid heuristic, which is evaluated in Section 7. Section 8. discusses multicast routing within PNNI. In the final section of the paper (Section 9.) we mention other aspects of application of the heuristics and identify further research.

2. DELAY BOUND MINIMUM COST MULTICAST ROUTING

The bounded delay minimum cost multicast routing problem can be stated as follows. Given a connected graph $G = \langle V, E \rangle$ where V is the set of its vertices and E the set of its edges, and the two functions: cost $c(i, j)$ of using edge $(i, j) \in E$ and delay $d(i, j)$ along edge $(i, j) \in E$, find the tree $T = \langle V_T, E_T \rangle$, where $T \subseteq G$, joining the vertices s and $M_{k,k=1,n} \in V$ such that $\sum_{(i,j) \in E_T} c(i, j)$ is minimised and $\forall k, k = 1, n; D(s, M_k) \leq \Delta$, the delay bound, where $D(s, M_k) = \sum_{(i,j)} d(i, j)$ for all (i, j) on the path from s to M_k

in T. Note that, if the delay is unimportant, the problem reduces to the Steiner tree problem. The addition of the finite delay bound makes the problem harder, and it is still NP-complete, as any potential Steiner solution can be checked in polynomial time to see if it meets the delay bound.

3. HEURISTICS WITH AN ARBITRARY DELAY BOUND

Several heuristics have been proposed that use arbitrary delay bounds to constrain multicast trees. Kompella, Pasquale, and Polyzos (Kompella et al., 1993) propose a Constrained Steiner Tree (CST_c) heuristic which uses a constrained application of Floyd's algorithm (Floyd, 1962). Widyono (Widyono, 1994) proposed four heuristics based on a constrained application of the Bellman-Ford algorithm (Bertsekas and Gallager, 1987). Zhu, Parsa and Garcia-Luna-Aceves (Zhu et al., 1995) based their technique on a feasible search optimisation method to find the lowest cost tree in the set of all delay bound Steiner trees for the multicast. Evaluation work carried out by Salama, Reeves and Vinitos (Salama et al., 1997) indicate that Constrained Steiner Tree heuristics have good performance, but high time complexity. The proposals for Constrained Shortest Path Trees by Sun and Langendoerfer (Sun and Langendoerfer, 1995), which we abbreviate as *CSPT* and by Waters (Waters and Crawford, 1996), which we abbreviate as *CCET* (Constrained Cheapest Edge Tree), generally have a lower time complexity than Constrained Steiner Trees, but their solutions are not as efficient.

In the following sections, we concentrate on the solutions offered by Kompella (representative of a very efficient, but high time complexity technique) and those of Waters and Sun and Langendoerfer, which, because they are based on variations of Dijkstra's shortest path algorithm and are of similar time-complexity, are good candidates for a hybrid heuristic.

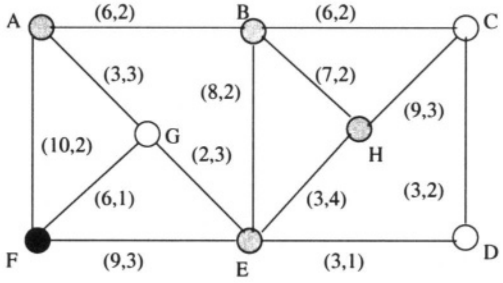


Figure 9.1 The example network

In the worked examples in the following description of these heuristics, we use the network illustrated in Figure 9.1, the edges of which are labelled with (*cost, delay*). The delay bound is set to 7 in all cases. Kompella and Sun use a Δ of 8 since they find paths with delay $< \Delta$; the Waters heuristic uses a Δ of 7 because it finds paths with delay $\leq \Delta$. In each case, the worked example finds the multicast tree connecting source F to the destinations A, B, E and H.

(Note that we consider symmetrical metrics in either direction on a link. In practice, ATM networks may have asymmetric metrics (e.g. bandwidth availability), and the network would be represented as a directed graph.)

3.1 THE CONSTRAINED STEINER TREE (CST_C) HEURISTIC (KOMPELLA, PASQUALE AND POLYZOS)

The CST_C algorithm was first published in (Kompella et al., 1993) and has three main stages (Kompella, 1993).

1. A closure graph (complete graph) of the delay-constrained cheapest paths between all pairs of members of the multicast group is found. The method to do this involves stepping through all the values of delay from 1 to Δ (assuming Δ takes an integer value) and, for each of these delay values, using a similar technique to Floyd's all-pairs shortest path algorithm (see (Floyd, 1962)).

2. A constrained spanning tree of the closure graph is found using a greedy algorithm. Two alternative selection mechanisms are proposed, one based solely on cost, the other on cost and delay. In our evaluation we use the more efficient of these (cost only) which selects edges for the spanning tree using the function :-

$$f_C = \begin{cases} C(v,w) & \text{if } P(v) + D(v,w) < \Delta \\ \infty & \text{otherwise} \end{cases}$$

   where $C(v, w)$ is the cost of a constrained path from node v to node w, $P(v)$ is the delay from the multicast source to node v and $D(v, w)$ is the delay on the path (v, w).

3. The edges of the spanning tree are then mapped back onto their paths in the original graph. Finally any loops are removed by using a shortest paths algorithm on the expanded constrained spanning tree (Kompella, 1993).

   (Note that for very large delay bounds compared to delays within the network, the solutions produced will be similar to those calculated using an approximation of the Steiner Tree Problem, e.g. (Gilbert and Pollack, 1968).)

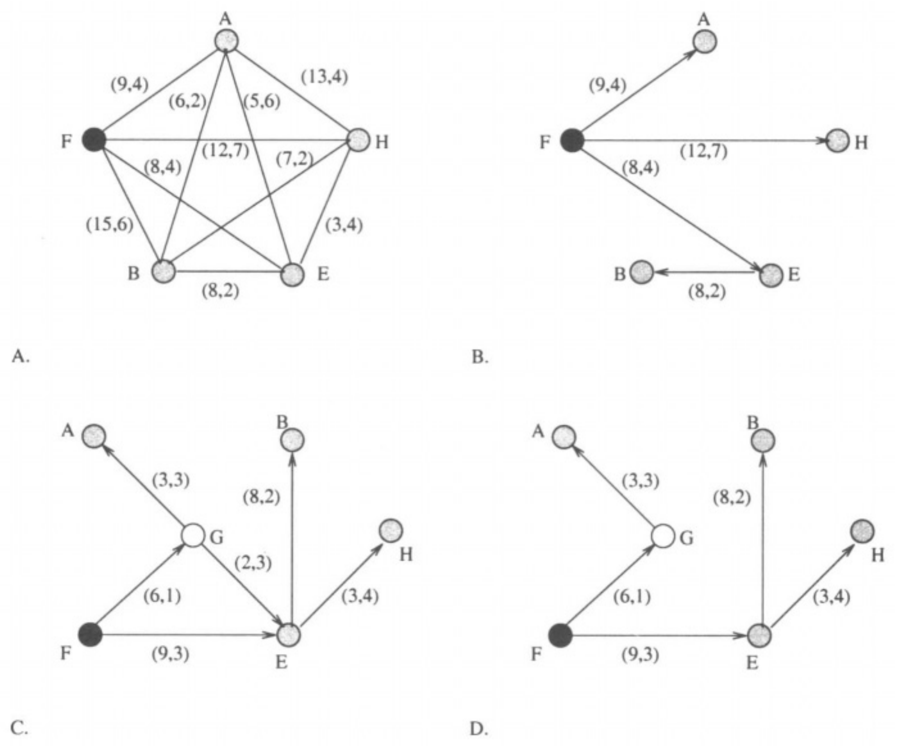


Figure 9.2 The CST_c heuristic

3.1.1 A Worked Example. Applying the first stage of the heuristic to the network in Figure 9.1 produces the constrained closure graph of paths in the multicast group illustrated in Figure 9.2A. Again, all links are labelled (*cost, delay*). Note that this graph need not be a complete graph so long as there are paths between every multicast node and the source.

Figure 9.2B shows the spanning tree obtained from the closure graph using the edge selection function f_C. Expansion of the spanning tree into their original paths results in a graph with a loop (Figure 9.2C.) which when removed produces the solution in Figure 9.2D. This tree has a cost of 29 units and a delay bound of 7.

3.1.2 Discussion of the CST_c Heuristic. The first stage of the heuristic is the most time consuming, giving an overall complexity of $O(\Delta n^3)$, where n is the number of vertices in the graph (Floyd, 1962). The effect of Δ on the time complexity can be reduced by decreasing the granularity of Δ through scaling, although this will compromise the accuracy of the results (Widyono, 1994).

In most cases CST_c calculates multicast solutions that are cheaper than those produced by a Shortest Path Tree algorithm (SPT) based on delay, but it does sometimes generate more expensive solutions. This may happen when a low delay edge is included in the SPT leading to more than one group member, wheras CST_c uses more direct routes to those members at cheaper cost.

For dynamic groups, CST_c may result in a multicast solution with a different topology as each memberjoins and leaves, since the closure graph at the second stage is applied to the current multicast group.

3.2 THE CONSTRAINED CHEAPEST EDGE TREE (*CCET*) HEURISTIC (WATERS)

The CCET heuristic was first published in (Waters, 1994) along with initial evaluations. In (Waters and Crawford, 1996), variations of the heuristic were introduced and comprehensively evaluated. The original heuristic was bound by either the broadcast delay or the multicast delay. Here we use the arbitrary delay, Δ. The CCET heuristic works as follows:

1. Use an extended form of Dijkstra's shortest path algorithm, to find for each $v \in V - \{s\}$ the minimum delay, *dbv*, from s to v. As the algorithm progresses keep a record of all the *dbv* found so far, and build a matrix *Delay* such that *Delay*(v, k_i) is the sum of the delays on edges in a path from s to k_i, whose final edge is (v, k_i), for each k that is adjacent to v.

2. For delay bound Δ, set all elements in *Delay*(v, k) that are greater than Δ to ∞. The matrix *Delay* then represents the edges of a directed graph derived from G which contains many possible solutions to a multicast tree rooted at s which satisfy the delay constraint.

3. Now construct the multicast tree T. Start by setting $T = \langle\{s\}, \emptyset\rangle$.

4. Take $v \in V - V_T$, with the maximum *dbv*, that is less than Δ, and join this to T. Where there is a choice of paths which still offer a solution within the delay bound, choose at each stage the cheapest edge leading to a connection to the tree. Include in E_T all the edges on the path (s, v) not already in E_T and include in V_T all the nodes on the path (s, v) not already in V_T.

5. Repeat step 4 until $V_T = V$, when the broadcast tree will have been built.

6. Prune any unnecessary branches of the tree beyond the multicast recipients.

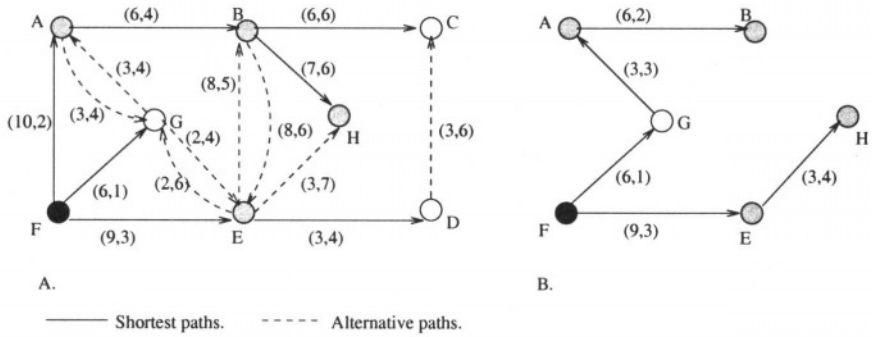


Figure 9.3 The CCET heuristic

3.2.1 A Worked Example. To illustrate the working of the heuristic we start with the graph shown in Fig, 9.1. The bracketed parameters for each link indicate (*cost*, *delay*). The example finds the multicast route from source F to destinations A, B, E and H.

The application of the extended form of Dijkstra's algorithm pruned to the delay bound $\Delta = 7$ results in the directed graph shown in Fig. 9.3A where the parameters shown against each link represent the edge cost and total delay from the source F to reach the node at the end of that link. The multicast tree is then constructed starting with $T = \langle F, \emptyset \rangle$. First H is connected to F using the path HE, EF. Node C is connected via the path CD, DE and then node B is connected via path BA, AG, GF. Finally, the edges CD and DE are pruned to give the multicast tree in Fig. 9.3B, with a cost of 27 units and a final delay bound of 7.

3.2.2 Discussion of the CCET Heuristic. The first stage, determining the directed graph, has the same time complexity as Dijkstra's algorithm, $O(n^2)$. The vertices can be put in delay bound order during the construction of the directed graph. In the second stage, building the multicast tree, requires a depth first search from each leaf node to find a path to the source. As the multicast tree grows, the search space for each leaf to source node path becomes smaller. The time complexity of the depth first search is $O(max(N, |E|)$ (Gibbons, 1989) where N is the number of nodes, and E is the set of edges in the search tree from the leaf node to the source. The number of paths considered in constructing the tree depends on the delay bound and the graph density. "Rogue" paths may be discovered which, although cheap, exceed the delay bound. These must be dicarded and the search recommenced, avoiding loops. In general, as the tree grows, the probability ofjoining the tree at a node closer to the source increases and paths nearer the source usually offer delays well within the bound. Because of these two characteristics, the probability of loops is minimised. The issue of loop removal is discussed in detail in (Crawford and Waters, 1997).

The CCET heuristic selects return paths on the basis of the "cheapest" exits from each node, back towards the source, that do not violate the arbitrary delay bound Δ. In some networks, multicast trees found by the heuristic can be more expensive than might be expected, beause of a trade-off between cheap edges and the alternative paths available within the delay bound. The cost of solutions found using Dijkstra's shortest path algorithm can sometimes be cheaper than those found using the Waters heuristic, depending again on the arrangement of edge costs and delays. Details can be found in (Crawford and Waters, 1997).

The multicast tree constructed by the CCET heuristic is pruned from the broadcast tree for a specific delay and delay bound. This means that in a dynamic environment where the multicast tree grows and shrinks, the broadcast tree need only be recalculated if the topology of the underlying network changes.

3.2.3 Constrained Cheapest Path Tree (CCPT). A variation on the Waters heuristic, proposed by Crawford (Crawford, 1994) uses the cheapest path back to the source rather than the cheapest edge leading to the existing tree as its selection mechanism. The idea is similar to a variation developed independently by Salama (Salama et al., 1995). We have included the CCPT heuristic in the first of our evaluations, but as it generally produces more expensive results than the CSPT heuristic described below, it was omitted from later evaluations.

3.3 THE CONSTRAINED SHORTEST PATH TREE *(CSPT)* HEURISTIC (SUN AND LANGENDOERFER)

This algorithm has three steps.

1. Using Dijkstra's shortest path algorithm compute a lowest *cost* spanning tree to as many destination nodes in the multicast as is possible without any path breaking the arbitrary delay bound, Δ.

2. Use Dijkstra's algorithm to compute a shortest delay path tree to those multicast nodes not reached in the previous step.

3. Combine the lowest cost spanning tree from the first step with the shortest delay path tree from the second step making sure that the delay to any destination node does not break the delay bound, Δ, and that all loops are removed.

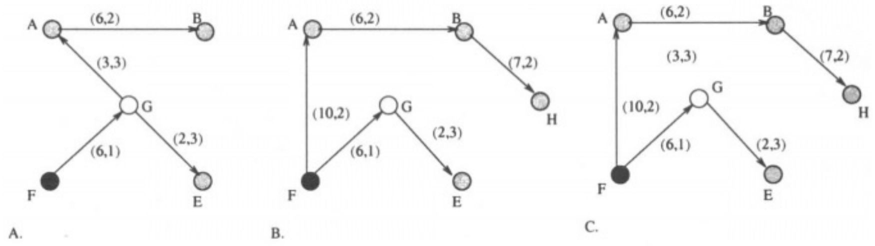


Figure 9.4 The CSPT heuristic

3.3.1 A Worked Example. Applying the first step of the heuristic to the network in Figure 9.1 produces the minimum cost path tree illustrated in Figure 9.4A. Node *H* is not included in this tree because its minimum cost path has a delay of 8, which breaks the delay bound. Figure 9.4B is the shortest delay path tree constructed only as far as node *H*, the multicast node not yet included in the solution. The combination of the minimum cost path tree and the shortest delay path tree will create a loop with nodes *F*,*G* and *A*. For this reason the edge *FA* is selected in preference to edge *GA* to give the final solution in Figure 9.4C. This tree has a cost of 31 units and a delay of 6.

Loop removal in the CSPT heuristic is much simpler than it is with the CST_c heuristic. Because steps 1 and 2 both use Dijkstra's algorithm to compute their trees, a loop occurs. The loop can be avoided by selecting, from the loop's downstream node, the shortest delay path tree branch in preference to the minimum cost path branch. This will increase the tree cost, but prevents violation of the delay bound.

3.3.2 Discussion of the CSPT Heuristic. Each of the first two steps of the heuristic have the time complexity of Dijkstra's algorithm, which is at most $O(n^2)$.

For the majority of multicasts, CSPT also calculates solutions that are cheaper than those produced by Dijkstra's SPT algorithm. As with CCET, there are also some cases where the cost of solutions found using the SPT algorithm can be cheaper than those found using the CSPT heuristic.

As CSPT multicast trees grow, they prone to reconfiguration if the arbitrary delay bound is less than the delay along the cheapest path to the new destination node. To remove loops, the shortest delay path may be substituted for an existing cheapest path in the tree. We propose a minor modification the the CSPT heuristic which eliminates its instability. Instead of calculating a solution for each multicast group, the calculation includes all nodes in the network, as is the case with the CCET heuristic and the multicast tree is pruned from the broadcast tree. We call the modified version the stable CSPT, or sCSPT. The two techniques are compared in (Crawford and Waters, 1997). For smaller multicast groups, the original heuristic produces, on average, more efficient solutions than sCSPT. As the group size increases the performance of the heuristics converges, as expected. The difference between the two techniques is small enough to consider sCSPT as a valid alternative to CSPT in dynamic routing situations.

4. EVALUATION ENVIRONMENT

Two network models are used to generate random networks in the evaluations described in this paper. In most cases, and where not stated explicitly, the network models are single cluster systems such as backbones or autonomous systems. These are generated using Waxman's model (Waxman, 1988) which distributes nodes randomly over a rectangular co-ordinate grid. The Euclidean distance between each pair of nodes is used for the delay metric. Edges are introduced with a probability depending on their length and a scaling factor, introduced by Doar (Doar, 1993) related to the number of nodes in the networks. The cost assigned to each edge is selected at random from the range [1„L] where L is the maximum distance between any two nodes.

We also use a cluster network (connecting a number of clusters via a backbone network) for some of our evaluations, based on the hierarchical model of Doar (Doar, 1993). Further details are given in (Crawford and Waters, 1997).

4.1 BENCHMARK ALGORITHMS AND ARBITRARY DELAY BOUNDS

As an exact solution to the constrained minimum cost tree problem is impractical for large graphs, we use the Minimum Steiner Tree heuristic (*MST*) of Gilbert and Pollack (Gilbert and Pollack, 1968) which approaches a minimum cost for multicast trees, although they are of unbound delay. We also use Dijkstra's *SPT* as a benchmark to evaluate the cost savings made by using the various heuristics.

We chose the network diameter as the arbitrary delay bound for the evaluation of the multicast algorithms. This provides an evaluation "mid-point" between the multicast delay (the tightest bound) and the MST which gives maximum improvement in cost.

5. EVALUATION OF THE CANDIDATE HEURISTICS

For each evaluation, 200 networks of 100 nodes of low edge density were used. Multicast groups were selected for sizes from 5 to 95 nodes, at steps of 5. There were 10 multicast samples for each multicast group size, for each network. (A list of acronyms appears in the Glossary.)

5.1 PERFORMANCE AVERAGES

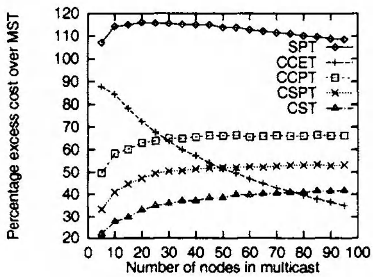


Figure 9.5 Average comparative costs

Figure 9.5 illustrates the percentage excess costs of using the four heuristics described above, relative to the *MST* and *SPT* benchmarks. For the CST_c heuristic we use a granularity of $\Delta/5$ to step through possible delay values (see Section 3.1).

The algorithm of *CSTc* generates multicast solutions that are on average cheaper than the other heuristics although, as the size of the multicast group increases, the *CCET* heuristic's solutions become cheaper than those of CST_c. The performance of the *CCET* heuristic is much better than *CSPT* and *CCPT* for larger multicasts, but is worse for smaller multicasts. The solutions of *CSPT* and *CCPT* are similar because they depend on Dijkstra's *SPT* algorithm for cost and delay. As *CCPT* gives poor perfromance, it is not considered further in this paper. Although *CCET* uses an extension of the *SPT* algorithm to construct its search space, it is not constrained by the algorithm when finding its solution, but it selects cheap edges leading to existing paths in the solution tree. This can result in small multicast solutions being relatively expensive, while large multicast solutions are generally much cheaper.

We have also observed that as the delay bound approaches the *MST* delay, improvements in solution efficiency of the *CCET* heuristic become negligible; maximum efficiency is approached at delay bounds of 3*network diameter or 3*broadcast delay from the source. Restricting the bound to these limits reduces the tree construction time. (See (Crawford and Waters, 1997).)

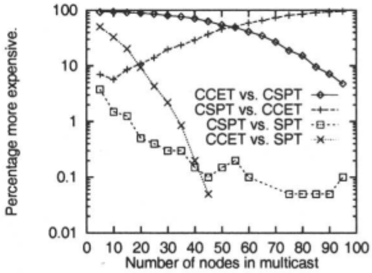


Figure 9.6 Exceptional comparative costs

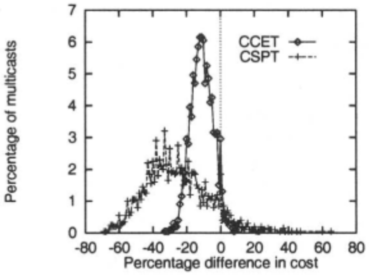


Figure 9.7 Cost distributions

5.2 SPECIFIC MULTICAST COMPARISONS

Although *CSPT* is generally better for smaller groups and *CCET* is more suited to larger multicasts, this is not always the case. Figure 9.6 illustrates a sample of the percentage of times *CCET* solutions are more expensive than those of *CSPT* and vice versa, and when the solutions of both *CSPT* and *CCET* are more expensive than the *SPT*. Despite the expected trend, in nearly 5% of the sample, for groups of 95 nodes, *CCET* was more expensive than *CSPT*. Similarly, in 7% of the sample, for groups of 5 nodes, *CSPT* was more expensive than *CCET*. For smaller multicast groups sizes, both *CSPT* and *CCET* generated some solutions that were more expensive than the *SPT* solutions. For larger multicasts *CSPT* still generates some solutions that are more expensive than *SPT*, while *CCET* does not. Figure 9.7 indicates just how large and varied these differences can be. The graph for *CSPT* plots the percentage cost savings of *CSPT* over *CCET* for small multicasts. While the majority of *CSPT* solutions are up to 69% cheaper, some can be up to 65% more expensive. Similarly, for *CCET* the majority of larger multicasts are up to 33% cheaper than *CSPT*, although some can be as much as 11% more expensive. This behaviour confirms that the solutions each heuristic generates depend on the algorithm, the topology of the network and the topology of the multicast. There is also a wide variance in the cost of solutions between the heuristics for the same size multicasts.

6. HYBRID APPROACH TO MULTICAST ROUTING HEURISTICS

We conclude from our evaluations that none of the heuristics we have considered can provide the "cheapest" multicast solutions in all networks for all sizes of multicast groups. They either take too long to compute or can sometimes generate unacceptable solutions. We propose a combined heuristic, of acceptable time complexity, that will generate solutions that are predominantly cheaper than *SPT*s for all network topologies, for all multicast group sizes.

We discard CST_c because, although it generates good solutions, it has an impractical time complexity and *CCPT* because of its poor overall performance.

The *CCET* and *CSPT* heuristics generate the majority of their most efficient multicast solutions at opposite ends of the multicast group size range, and both base their calculations on trees generated by the *SPT* algorithm. Individually, each is vulnerable to generating some inefficient solutions throughout the full range of multicasts, but rarely will both heuristics generate an inefficient solution for the same network/multicast group pair. We combine the *CCET* and *CSPT* heuristics to obtain a hybrid of acceptable time complexity that produces solutions of significantly improved efficiency over *SPT*s. The hybrid will select the "cheapest" tree provided by each of these heuristics or by the *SPT* as the multicast solution. *SPT* is included as it ocassionally produces cheaper solutions than *CCET* or *CSPT*. The *CCET* function, within the hybrid, must place a maximum limit on the delay bound used, as previously discussed.

The hybrid first calculates the shortest path tree for delay, which is extended for the second stage of the *CCET* heuristic. The *CSPT* heuristic also calculates the *SPT* shortest path tree for cost (possibly concurrently with the delay calculation). Once the trees have been obtained for each method their costs can be easily calculated and the cheapest tree selected as the solution.

The time complexity of the hybrid is dominated by the *CCET* function. The first stage of this function has time complexity of at most $O(n^2)$. The second stage, the construction of the broadcast tree, has a time complexity of $O(max(N, |E|))$, limited in practice by using an acceptable delay bound. The *CSPT* and *SPT* functions have a time complexity of $O(n^2)$.

7. EVALUATION OF THE HYBRID HEURISTIC

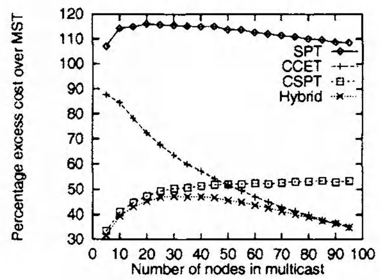


Figure 9.8 Average comparative costs

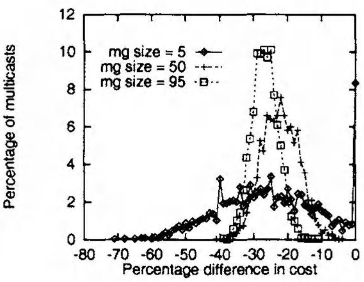
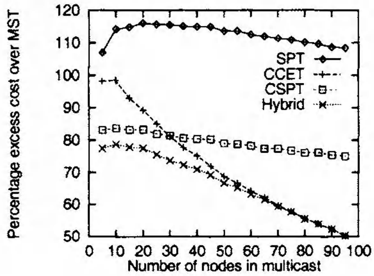


Figure 9.9 Cost distribution

Figure 9.8 illustrates the cost performance of the hybrid heuristic in comparison to *CCET* and *CSPT*. The hybrid outperforms or equals both *CCET* and *CSPT*, as expected. It is interesting to note that for mid-sized multicasts the hybrid is able to provide solutions that are better than either *CSPT* or *CCET* can do separately, since the hybrid is able to choose the most efficient heuristic for each particular multicast. The efficiency of hybrid solutions for small multicasts is still subject to a fairly wide variance as figure 9.9 shows. These graphs plot the cost savings distributions of the hybrid over *SPT* for multicast group sizes of 5, 50 and 95 respectively. The dominance of *CSPT* for small multicast groups and *CCET* for large multicasts is obvious, as is the narrow but sharp intervention of *SPT* when required.



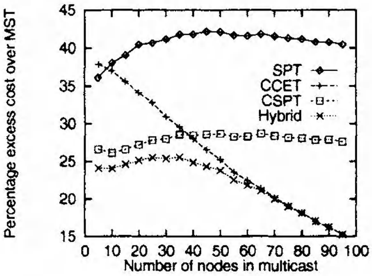


Figure 9.10 Multicast bound; single cluster ***Figure 9.11*** Multicast bound; multi-cluster

Figures 9.10 and 9.11 show the performance of the heuristics at the tightest possible delay bound, the delay to the furthest member of the multicast group. Figure 9.10 is plotted for a single cluster network and Figure 9.11 is a two-level hierarchy with clusters connected by a backbone. Results are similar for both hierarchical and non-hierarchical networks and the improved performance of the hybrid is confirmed. Note that, within this tight delay bound, the CSPT gives much smoother performance across the range of multicast group sizes and it is hard to achieve a very efficient solution for the smaller groups. The hybrid reflects this situation.

7.1 DISTRIBUTION OF PATH DELAYS

We have noticed that the hybrid (and other solutions which attempt to minimise cost) tend to smooth out the distribution of delays perceived by the participants, whereas SPT, although it produces shorter delays, tends to concentrate delays in a smaller band. This smoothing may help to reduce the amount of buffer storage required where it is necessary to store information before playing it back at the same time at all recipients.

8. ROUTING AND PNNI

Current work on routing strategies for large scale ATM networks centres around the ATM Forum's Private Network Node Interface (PNNI). PNNI offers both signalling between nodes and a VC routing protocol. It supports hierarchical routing and Quality of Service with multiple routing metrics and attributes. Within PNNI, sets of nodes are arranged logically into Peer Groups for the purposes of creating a routing hierarchy. Within a Peer Group, information on QoS and reachability is exchanged using flooding. Peer Groups are arranged in a parent-child hierarchy. A Peer Group Leader collects and aggregates this information into data which represents the characteristics of the entire Peer Group. This information is then passed to the parent group enabling it to see the child group as a Logical Node with the aggregated characteristics.

PNNI allows many different algorithms to be used to compute routes. Within a Peer Group, our heuristics can be used to optimise multicast routes. However, when using the Logical Node representation of Peer Groups, two effects are likely. First, we might be more cautious than necessary if an aggregated value for the delay is much higher than the actual delay incurred in reaching the multicast nodes. Secondly, it is likely that the aggregated knowledge will lead to a less efficient route than if we had full knowledge of the network.

A number of other authors have considered aspects of PNNI routing related to optimisation. The clustering of nodes within Peer Groups has been studied by Rougier and Kofman (Rougier and Kofman, 1998), with a view to optimising hierarchical routing. Their technique uses a random geometric approach to obtain a tessellation which partitions the nodes into groups, a process repeated from the top level to successive levels in the hierarchy. Their results show that, after a small number of levels, there is little significant increase in routing table size, for an increase in the number of levels. When VC set-up is done on demand, the higher the hierarchical level, the lower the complexity, so the optimum number of hierarchical levels is a tradeoff between the number of on-demand calls and routing complexity.

Other authors have considered optimisations based on where the multicast copying is done. Tode et al (Tode and Ikeda, 1998) argue that it may not be desirable for all nodes to do multicast copying. They investigate arrangements of copy nodes which still maintain a good geographical distribution. In contrast, Kadirire (Kadirire, 1994) tries to reduce the copying incurred at any specific node whilst also maximising geographic spread such that group joins carry little extra cost as they are likely to be near the existing tree.

Barakat and Rougier (Barakat and Rougier, 1998) discuss optimisation of hierarchical multicast trees in ATM networks. They consider Centre Based shared trees which are now becoming a possibility with recent additions to PNNI capable of supporting many to many connections. By having multiple cores, the problem of concentrating the traffic around the cores is reduced, but the advantages of reduced state information is maintained. In their scheme, each Peer Group has a core; routes then use a combination of shared trees within the Peer Group and links between Peer Groups. This scheme performs best for dense groups, which are not likely to have poorly placed cores.

An alternative solution to core placement is considered by Komandur et al (Koandur and Mosse, 1998). Their scheme is based on routing domains, with a domain at a switch being the highest level Peer Group entered from the incoming link for the connection. The cores are not preconfigured, and one objective is not to overload any one switch. Domain servers help with the task of core placement.

Although the heuristics described in this paper are principally source-based, they might be used to aid in connecting Peer Groups within which shared trees are used. Also, the use of two metrics forms a basis for consideration of multiple QoS parameters for instance by optimising the placement of the core within a Peer Group, which would have application to shared trees.

9. CONCLUSIONS AND FURTHER RESEARCH

We have identified problems of time complexity and performance variability in heuristics that have been proposed to calculate low-cost multicast trees that are bound by an arbitrary delay. By combining appropriate heuristics we propose a hybrid that produces efficient solutions over all multicast group sizes with an acceptable time complexity. The evaluations of the hybrid have included both flat and hierarchical networks over a range of group sizes and using an “average” and a tight delay bound. The hybrid is shown to perform well under all these circumstances.

The hybrid heuristic uses metrics for every link in a network to perform its route calculation and so is amenable for implementation in other link-state routing protocols such as the Internet’s Multicast Open Shortest Path First protocol (Moy, 1994). Further work is needed into QoS routing in the Internet, which is likely to to include ATM segments.

The evaluations of the hybrid have included both flat and hierarchical networks over a range of group sizes and using an “average” and a tight delay bound. The hybrid is shown to perform well under all these circumstances.

For dynamic groups, the hybrid, in common with CST_C (Kompella) will sometimes involve reconfiguration of the multicast tree. Where it is particularly important to have a stable tree, which can be pruned and regrow branches, we suggest the use of the constituent heuristics: CCET (Waters) for large groups relative to the size of the network and the broadcast and prune version of CSPT (Sun) which we propose in Section 3.3.2.

An important result of this work is the integration of several heuristics which are individually unstable into a stable hybrid. Hybrid methods may also have an application in other multicast or load sharing route calculation algorithms.

Further work is needed to evaluate the effect of using the heuristics within a hierarchical network structure.

GLOSSARY

CSTc Constrained Steiner Tree (Kompella, Pasquale and Polyzos)

CSPT Constrained Shortest Path Trees (Sun and Langendoerfer)

CCET Constrained Cheapest Edge Tree (Waters)

CCPT Constrained Cheapest Path Tree (Crawford)

SPT Shortest Path Tree (Dijkstra)

MST Minimum Steiner Tree (Gilbert and Pollack)

Acknowledgments

We would like to acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) for our work on multicast routing (Grant ref. GR/K55837).

References

Barakat, S. and Rougier, J. (1998). Optimization of Hierarchical Multicast Trees in ATM Networks. In *Sixth IFIP Workshop on Performance Modelling and Evaluation ofATM Networks,* pages 44/1–44/10.

Bertsekas, D. and Gallager, R. (1987). *Data Networks.* Prentice-Hall,Inc.

Crawford, J. (1994). Multicast Routing: Evaluation of a New Heuristic. Master's thesis, University of Kent at Canterbury.

Crawford, J. and Waters, A. (1997). Low Cost Quality of Service Multicast Routing in High Speed Networks. Technical Report 13-97, University of Kent at Canterbury.

Doar, J. (1993). Multicast in the Asynchronous Transfer Mode Environment. Technical Report No. 298, University of Cambridge Computing Laboratory.

Floyd, R. (1962). Algorithm 97: Shortest path. *Communications of the ACM,* 5(6):345.

Gibbons, A. (1989). *Algorithmic Graph Theory*. Cambridge University Press.

Gilbert, E. and Pollack, H. (1968). Steiner Minimal Trees. *S1AM Journal on Applied Mathematics*, 16.

Kadirire, J. (1994). Minimising packet copies in multicast routing by exploiting geographic spread. *Computer Communications Review,* 24(3):47–62.

Koandur, S. Doar, M. and Mosse, D. (1998). The Domainserver Hierarchy for Multicast Routing in ATM Netwrorks. In *Sixth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks,* pages 48/1–48/6.

Kompella, V, P. (1993). *Multicast Routing Algorithms for Multimedia Traffic*. PhD thesis, University of California, San Diego, USA.

Kompella, V., Pasquale, J., and Polyzos, G. (1993). Multicast Routing for Multimedia Communications. *IEEE/ACM Transactions on Networking*, 1(3):286–292.

Moy, J. (1994). Multicast Extensions to OSPF. RFC 1584.

Rougier, J. and Kofman, D. (1998). Optimization of Hierarchical Routing Protocols. In *Sixth IFIP Workshop on Performance Modelling and Evaluation ofATM Networks*, pages 43/1–43/10.

Salama, H., Reeves, D., Vinitos, I., and Sheu, T.-L. (1995). Evaluation of Multicast Routing Algorithms for Real-Time Communication on High-Speed Networks. In *Proceedings of the 6th IFIP Conference on High-Performance Networks (HPN'95).*

Salama, H., Reeves, D., and Vinitos, Y. (1997), Evaluation of multicast routing alogorithms for real-time communication on high-speed networks. *IEEE Journal on Selected Areaa in Communications*, 15(3):332–345.

Sun, Q. and Langendoerfer, H. (1995). Efficient Multicast Routing for Delay-Sensitive Applications. In *Second Internatiopnal Workshop on Protocolsfor Multimedia Systems (PROMS'95)*, pages 452–458.

The ATM Forum Technical Committee (1996). *Private Network-Network Interface Soecification, Version 1.0.* The ATM Forum.

Tode, H. Yamauchi, H. and Ikeda, H. (1998). Copy node allocation algorithms for multicast routing in large scale ATM networks. In *Sixth IFIP Workshop on Performance Modelling and Evaluation ofATM Networks,* pages 47/1-47/10.

Waters, A. (1994). A New Heuristic for ATM Multicast Routing. In *2nd IFIP Workshop on Performance Modelling and Evaluation of ATM Networks,* pages 8/1–8/9.

Waters, A. and Crawford, J. (1996). Low-cost ATM Multimedia Routing with Constrained Delays. In *Multimedia Telecommunications and Applications (3rd COST 237 Workshop, Barcelona, Spain*), pages 23–40. Springer.

Waxman, B. (1988). Routing of Multipoint Connections. *IEEE journal on selected areas in communications*, 6(9): 1617–1622.

Widyono, R. (1994). The Design and Evaluation of Routing Algorithms for Real-time Channels. Tr-94-024, University of California at Berkeley and International Computer Science Institute.

Zhu, Q., Parsa, M., and Garcia-Luna-Aceves, J. (1995). A Source-Based Algorithm for Near-Optimum Delay-Constrained Multicasting. In *Proceedings of INFOCOM*, pages 377–385.

10. BIOGRAPHIES

Gill Waters is a Senior Lecturer in Computer Science at the University of Kent at Canterbury, UK. She holds a B.Sc. in Mathematics from Bristol University and a Ph.D. from the University of Essex, where she was a Lecturer from 1984-1994 and has considerable previous experience of software and protocols. Her research concerns distributed applications that use multimedia and/or multicasting and the required network architecture, protocol and system support for these applications. Specific projects include multicast routing, performance modelling, multimedia information retrieval, caching hierarchies, QoS provision on the Internet and design support environments for distributed systems.

John Crawford holds M.Sc. and Ph.D. degrees from the University of Kent. He has a broad industrial background in the development of telecommunication systems, where he has held roles ranging from software engineer to consultant. Since 1994 he has undertaken research and teaching in the Networks and Distributed Systems Group of the Computing Laboratory at the University of Kent, where his main interest concerns multicast routing algorithms, protocols and Quality of Service issues in ATM and IP networks.

Chapter 10

EMBEDDING RESILIENCE IN CORE ATM NETWORKS

Paul Veitch
Advanced Communications Engineering
BT Adastral Park: MLB 3-53e
Martlesham Heath
Ipswich IP5 3RE
England, UK.
paul.veitch@bt.com

Abstract: With the increased deployment of ATM in wide area networks, it is imperative to embed resilience mechanisms in the network elements to mitigate the impact of outages caused by cable breaks and node failures. Although SDH network functionality can be exploited to provide resilient transport of ATM connections, this adds a cost overhead to the overall network design. Furthermore, ATM-layer faults such as ATM switch failures will not be detected by fault-monitoring procedures executed within the SDH layer. It is thus crucial that resilience mechanisms are embedded in the ATM layer. Since user requirements vary from service to service, it is highly likely that different customers will demand variable levels of resilience. For example, mission-critical business-oriented data services will rely on virtually fault-transparent service, whereas residential customers may tolerate breaks in service as long as they do not occur frequently or last a long time. Fortunately, as this article explains, different ATM restoration mechanisms are possible which suit varied customer requirements.

Keywords: ATM, Resilience, Self-healing, Broadband

1. INTRODUCTION

With the extensive deployment of high capacity fibre-optic cables to interconnect telecommunications switching systems capable of handling data at Gbit/s speeds, there is an increasing demand on network planners to incorporate resilience mechanisms into architectural designs. The issue of network resilience has come to the fore in recent years due to a series of highly publicised outages causing widespread service disruption, sizeable revenue losses, and ultimately, loss of customer trust[1]. This article addresses the challenges involved in embedding resilience into wide area asynchronous transfer mode (ATM) networks.

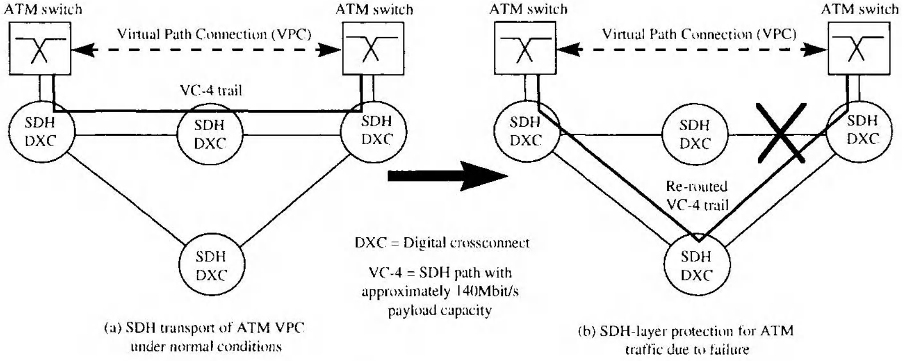
In section 2, the principal drivers for ATM resilience mechanisms will be explained. Section 3 explores the viable options to provide ATM network resilience and considers the impact of each scheme on cost and performance. Section 4 discusses how different restoration schemes may be applied to suit distinct customer requirements where fault-tolerance is concerned. Finally, section 5 concludes the paper.

2. DRIVERS FOR ATM NETWORK RESILIENCE MECHANISMS

Any wide area network is vulnerable to cable breaks and node outages. The prospect of ATM networks being ubiquitously deployed within a Broadband Integrated Services Digital Network (B-ISDN) framework to support a diverse mix of switched and private services generates the concern that such networks will be extremely vulnerable to failures causing huge volumes of information loss. Although ATM services may be run over a Synchronous Digital Hierarchy (SDH) core transport layer with embedded protection capabilities as shown in Figure 10.1, only physical layer faults can be detected and restored with SDH functionality. There must therefore be resilience mechanisms built into the ATM switches to cope with faults originating at the ATM layer, for example switch failure due to routing table corruption. There are further drivers for allowing the ATM layer to be made resilient to *all* faults including those originating from the physical-layer such as cable breaks, namely:

- There are extra costs incurred by having an additional layer of switched transport such as SDH[2].
- Since resilience mechanisms are needed in both layers of a multi-layer architecture comprising SDH and ATM, interactions and escalation must be managed accordingly[3], adding a further level of complexity to the network design.

Hence, given that resilience mechanisms will be an essential feature of wide area ATM networks[4], optional techniques must be considered in terms of cost and performance as detailed in the following section.


ATM switch
ATM switch
Virtual Path Connection (VPC)
VC-4 trail
SDH DXC
SDH DXC
SDH DXC
SDH DXC
ATM switch
ATM switch
Virtual Path Connection (VPC)
SDH DXC
SDH DXC
SDH DXC
Re-routed VC-4 trail
SDH DXC
DXC = Digital crossconnect
VC-4 = SDH path with approximately 140Mbit/s payload capacity

(a) SDH transport of ATM VPC under normal conditions

(b) SDH-layer protection for ATM traffic due to failure

Figure 10.1. SDH transport and protection for ATM traffic

3. ALTERNATIVE TECHNIQUES FOR ATM RESILIENCE

3.1 OVERVIEW OF METHODS

The term “resilience” provides a broad description of various aspects of the design and control of fault-tolerant networks as highlighted in an ITU-T study document on ATM network survivability[5].

- Protection involves the assignment of an alternative route with dedicated bandwidth assignment. When a failure affects the working route, a distributed management protocol realises switch-over.

- Restoration may be performed with a centralised control system (reconfigurable networks), or, it may involve either distributed control or management procedures (self-healing networks). In both cases, resources may be semi-dedicated whereby the alternate route is pre-determined but the bandwidth is assigned “on-demand” following fault-detection, or, both the route and the bandwidth may be assigned in real-time (“on-demand”).

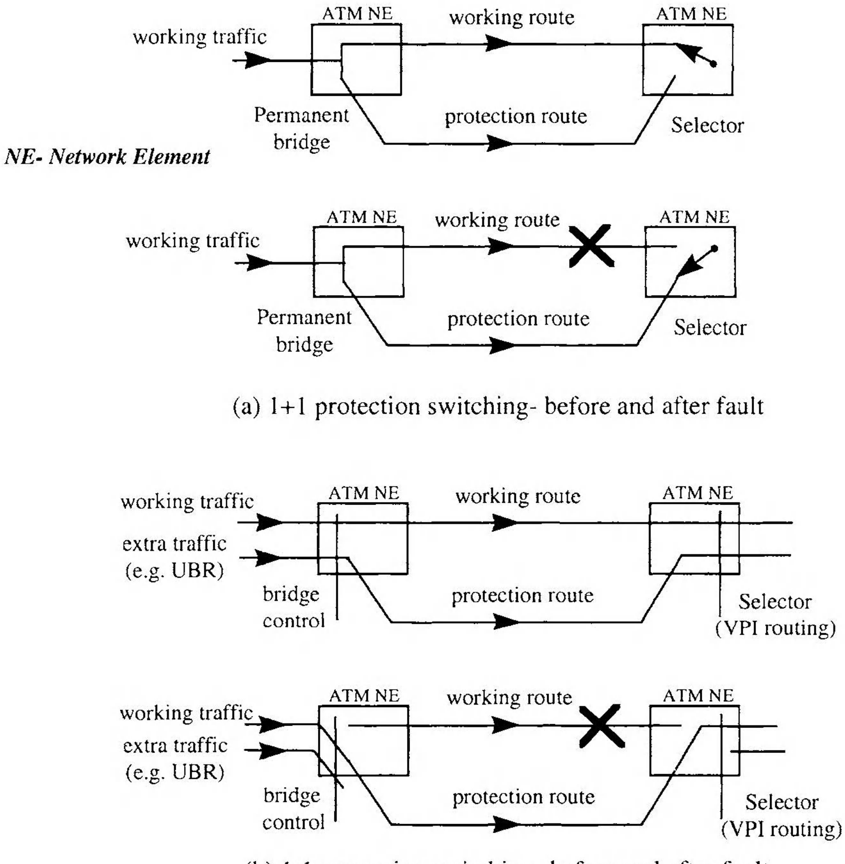
A key differentiator between protection and restoration is that less spare capacity is required with the latter since sharing between failure events is feasible: protection can consume greater than 100% of the working capacity whilst restoration in a well-connected mesh may only require about 50% extra capacity relative to the working demands[6]. Meanwhile, a subtle distinction exists between distributed control and distributed management in that the former relates to connection set-up procedures with control plane signalling cells whilst the latter involves the use of management plane messages in the form of operations and maintenance (OAM) cells.

Each resilience mechanism may operate at virtual path (VP) or virtual channel (VC) level, however, rather than examine every permutation, this article examines a realistic subset of resilience options which have been studied in the literature, and in some cases, have been proposed for standardisation.

3.2 ATM PROTECTION NETWORKS

Protection networks are intended to provide high levels of reliability, and are generally the most expensive resilient architecture since resources (bandwidth and virtual path identifier/virtual channel identifier (VPI/VCI) numbers), are dedicated rather than shared. Consequently, the pre-allocation of routes and resources enables very simple distributed management protocols to be executed in the event of a network impairment. Network connection protection (NCP) may be applied end-to-end or sub-network connection protection (SNCP) may target a segment of a complete connection. To ease control, protected and unprotected segments should align with appropriately designated operations and maintenance (OAM) flows[7].

An ATM VP/VC protection switching protocol has been specified for point-to-point protection architectures operating within both NCP and SNCP domains[5]. The protocol may be executed on a 1 + 1 or a 1:n basis, as shown in Figure 10.2 for the case where n=1. With 1 + 1 protection, the source node of the protected segment is permanently bridged so that traffic occupies the working and protection routes. A selector at the protected segment sink normally chooses the working route, but in the event of a network fault which impairs the working route, switchover to the protection route is instigated, e.g. by VPI/VCI changeover. For unidirectional switchover therefore, actions are required only at the sink node.



(a) 1+1 protection switching- before and after fault

(b) 1:1 protection switching- before and afteı fault

Figure 10.2. 1+1 and 1:1 protection switching concepts

With 1:1 protection, working traffic only fills the working route under normal conditions, and in the event of a failure, the bridge connects the working traffic to the protection route. The selector at the sink works in the same way as before, meaning 1:1 protection switching requires communications in both directions even for unidirectional switching, making it a little slower than 1 + 1 switchover. An advantage of 1:1 protection meanwhile, is the option to transmit "Extra Traffic" on the protection route under normal conditions, suitable say for unspecified bit rate (UBR) services: in the event of switchover, the UBR traffic is discarded. If extra traffic is not used to fill the protection bandwidth, it is reserved and remains idle. Specific details of the information written into OAM cells for 1 + 1 and 1:1 switchover control may be found in reference [5].

ATM protection switching may be supported in ATM crossconnects within a mesh network architecture, or in add-drop multiplexers (ADMs) as part of a self-healing ring. ATM ADMs are conceptually the same as SONET/SDH ADMs[8], except that logical switching of virtual paths is performed in place of synchronous time division multiplexed (TDM) paths. The feasibility of exploiting ATM ADMs in ring architectures was demonstrated in [9].

Although it is very desirable to achieve switchover times which are comparable with SONET/SDH protocols (60 msec including fault detection), the fact that VPs are of variable granularity, and up to 4096 separate connections may be supported on a single link[10], implies a lot of alarm generation and re-routing in the event of a cable break or node failure. This places a significant processing overhead on the ATM network elements. One proposal to mitigate this problem is to assign whole VP connections which follow identical physical routes and have the same source and sink, into virtual path groups (VPGs)[1 1,12]. Identification of which VPs belong to which groups is not accommodated in cell header labels, hence, a logical association between VPs and VPGs is required in routing tables of ATM network elements.

3.3 RECONFIGURABLE NETWORKS

The use of a centralised network management system (NMS) for connection restoration is a relatively simple method of providing resilience, with network elements notifying the NMS of a fault which is then responsible for co-ordinating reconfiguration. It may exploit pre-planned alternate routes or search for them dynamically according to current network state information. Furthermore, it is a simple task to prioritise re-routing of connections to ensure those with critical applications are restored quicker than services which are tolerant of temporary data loss. Nevertheless, there are three stages of processing which cause a generally slow overall response, described with reference to Figure 10.3:

1. Upstream communications between network nodes adjacent to the failure and the NMS;
2. NMS processing required for routing and bandwidth allocation;
3. and finally, downstream communications between the NMS and network nodes, followed by appropriate network element reconfiguration.

Centralised restoration in present-day synchronous transport networks typically takes from a few minutes to tens of minutes [13,14] which may be unsuitable for certain mission-critical services. Moreover, since many centralised network management systems are proprietary, it is difficult to co-ordinate fault management between equipment from different vendors. In terms of restoration times, the resilience mechanism which exhibits intermediate performance between protection networks and reconfigurable networks is "self-healing".

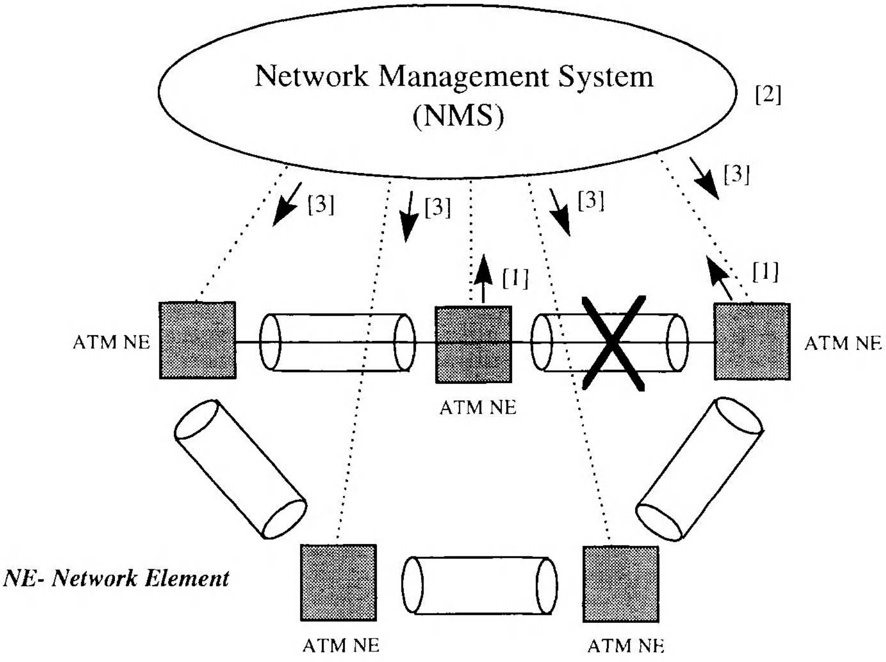


Figure 10.3. Centralised ATM network restoration

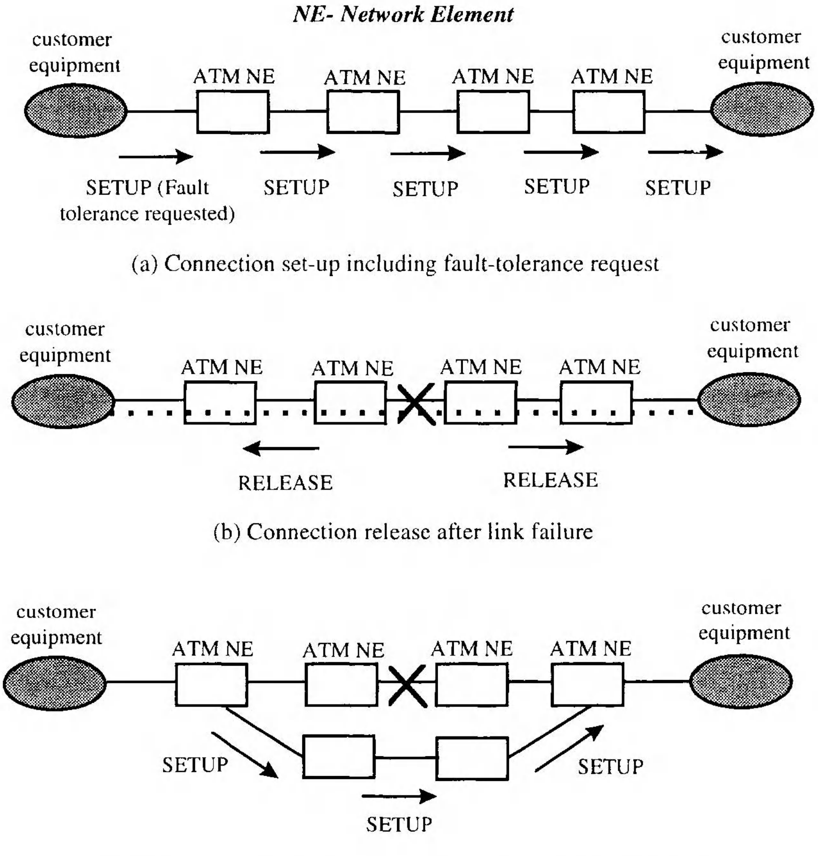
3.4 SELF-HEALING TECHNIQUES FOR ATM NETWORKS

Self-healing employs distributed management or control plane signalling functionality, and involves either on-demand or semi-dedicated resource allocation[4]. Three methods of on-demand self-healing are presently explained followed by an explanation of semi-dedicated backup VPs.

3.4.1 Self-healing with PNNI Routing

The ATM Forum has defined the private network node interface (PNNI) protocol to establish switched VCs (SVCs) between ATM switches within networks which use network service access point (NSAP) type ATM addresses[15]. Since the ATM Forum has specified an NSAP encoding for E.164 addresses, the PNNI routing protocol may also be employed in public networks. The PNNI specification describes how network nodes maintain knowledge about reachability and resources within the network by using a topology state routing protocol involving periodic information exchange between nodes. This lets nodes which receive connection requests determine a route for the signalling packets which will most likely achieve successful call set-up. A “crankback” mechanism also exists to divert connection set-up attempts away from a point of congestion and back to a previous node in the selected set-up path from which a new route will be sought. In the event of a network failure, connections will be cleared down and customers or their customer premises equipment (CPE) will have to instigate re-dial into the network. If the source node that handles the connection request has not yet learned of the topology update following the failure, the crankback mechanism will divert traffic from the failed link, though it is possible that a processing bottleneck will occur in the vicinity of the fault itself.

It is likely that phase 2 PNNI will incorporate *automatic* call re-routing thus relieving customers or their CPEs from having to re-dial. A fault tolerant signalling procedure based on end-to-end re-routing has been proposed to the ATM Forum[16,17] whereby new routes for failed calls will be automatically sought by ingress switches which initially dealt with the associated connection requests. This will be an option supported by a fault-tolerant routing descriptor contained within “SETUP” messages (Figure 10.4(a)). To ensure efficient re-routing in the event of a failure, the source node must have learned about the network failure so it can determine suitable routes for signalling messages which completely avoid the failed element. After receiving a “RELEASE” message therefore (Figure 10.4(b)), a connection recovery timer will be set to allow time for new topology status to be received, after which connection re-establishment will be attempted (Figure 10.4(c)). Due to the possibility of several simultaneous re-dials, there are no guarantees that an alternate route will be found and restoration times may vary from seconds to minutes. A generic re-routing framework is being proposed for PNNI version 2 which will enable a variety of re-route options to be accommodated[18].



(a) Connection set-up including fault-tolerance request

(b) Connection release after link failure

(c) Connection re-establishment after topology updates received

Figure 10.4. Proposed self-healing embedded into PNNI routing

3.4.2 Soft PVCs/PVPs

Usually, permanent VCs/VPs (PVCs/PVPs) will be set up from source to destination using network management procedures. If fault recovery is needed, a proprietary central management system must intervene and orchestrate reconfiguration as shown in Figure 10.3. "Soft" PVCs/PVPs are private circuit connections which, upon ordering from the user to the network management, are actually set up between ATM switches using the distributed control plane messaging as is used for switched connections, such as the PNNI routing protocol. At least one vendor[19] is selling the idea of soft PVCs/PVPs to provide resilient private ATM connections. When a failure occurs, the connection will be cleared as far back as the ingress/egress ATM switches, followed by automatic instigation of re-routing in a similar fashion to that depicted in Figure 10.4.

3.4.3 Distributed Restoration Algorithms (DRAs)

The self-healing schemes described thus far work on a connection basis, whereby detection of a failure is followed by re-instigation of connection establishment procedures. This methodology is similar to a class of distributed restoration algorithm (DRA) called "path DRAs". A DRA is a generic term used to describe protocols which dynamically restore failed transport capacity, and were originally proposed for operation in SONET/SDH networks[20]. The aim of DRAs is to re-route traffic quickly (< 2 seconds) using digital crossconnects switching at high granularity, such that actual user connections such as voiceband calls or data transfer sessions would not be cleared down. In other words, there is a clear distinction between transport protocols operating on paths, and control signalling used for individual circuits. Early research into ATM self-healing focused on the principle of VPs forming a transport layer for VCs, with the target of fast restoration at the VP layer to provide transparency at the VC layer[21-24]. Many papers proposed variations on the general approach of propagating flooded control messages from the nodes adjacent to a failed span as shown in Figure 10.5 to seek out and capture spare capacity for the VPs affected by the failure. The control messages in this approach which is termed "span DRA" thus pertain to any number of impaired VPs.

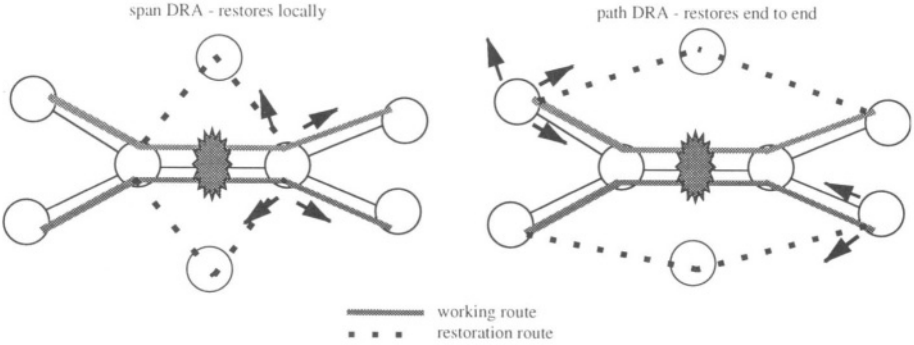


Figure 10.5. Span and path DRA principles

In contrast to the "span DRA" methodology, a path DRA involves tracing failed paths back to their endpoints and instigating flooding (Figure 10.5). In this way, control messages pertain to individual path connections or groups of paths affected by a fault. The path DRAs are more flexible than span DRAs since they can inherently handle multiple span and node failures, while the span DRA requires further extension. This is important because multiple span failures are actually quite common in real networks. Consider Figure 10.6 for example, where a single cable cut results in three logical span failures as a result of line systems bypassing some switches to reduce costs.

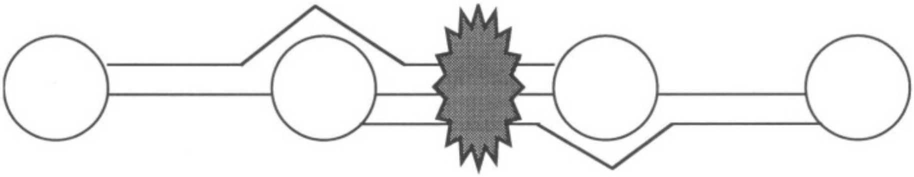


Figure 10.6. Single cable failure resulting in multiple span fail

When automatic re-routing is incorporated into signalling protocols such as PNNI, this method of self-healing and path DRAs are similar. Three apparent distinctions which can be identified, are:

- PNNI works strictly at the virtual channel (VC) or virtual path (VP) level, however DRAs can be easily adapted to work on whole or segmented path groupings, possibly yielding a speed advantage where bulk restoration is concerned.
- PNNI employs source-based routing with crankback[15], whilst the DRA speculatively floods the network on a link-by-link basis.
- PNNI routing holds the advantage of being standardised by the ATM Forum whereas DRAs have not been standardised.

Despite this last point, it is worth noting that a testbed has been developed at BT Labs and Alcatel Telecom to demonstrate the viability of DRAs[25]. The testbed experiments were conducted on a 7-node network to prove the viability of incorporating DRA functionality into ATM switches. The principal restriction with testbed experimentation however, is the size of the network being constructed, which is constrained by costs and implementation overheads. Simulation tools have therefore been exploited as a means to evaluate systems such as large self-healing telecommunications networks, which would be otherwise impractical to construct in the form of a testbed.

The scalability of employing DRAs to restore ATM circuits in a realistically-sized backbone topology was confirmed by employing a simulation tool enabling an object-oriented model of a network to be specified hierarchically[25]. From Figure 10.7, the interconnection of nodes with links is defined at the *network level,* which represents how the ATM network elements are connected together by fibre transmission systems. At the *node level,* the internal architecture of the ATM switch can be defined. It is here that certain abstractions may be made to produce an accurate, yet manageable representation of the structure of the network element. For example, since the performance of a DRA relies principally on the processing and communication of flooding messages, modelling at the node level concentrated on the extraction, processing, generation, and re-insertion characteristics of such messages within a switching node's architecture. The lowest level of the modelling hierarchy is the *process level*, where the actual DRA functionality is embedded into the overall network model.

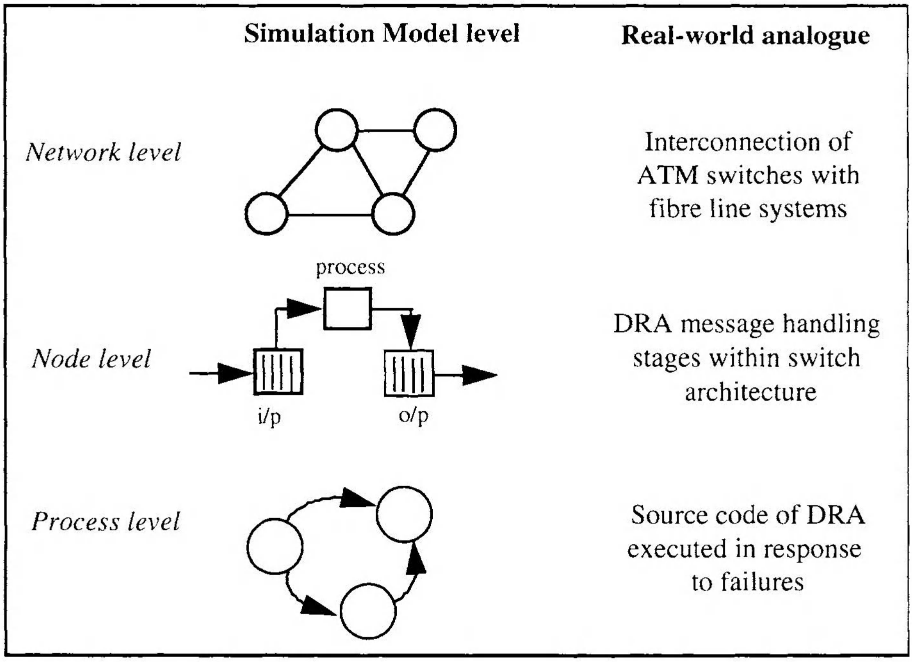

Simulation Model level
Real-world analogue
Network level
Interconnection of ATM switches with fibre line systems
process
Node level
i/p
o/p
DRA message handling stages within switch architecture
Process level
Source code of DRA executed in response to failures

Figure 10.7. Relating the simulation model to actual telecoms network

The scalability of employing DRAs to restore ATM circuits in a realistically-sized backbone topology was confirmed by simulating a path DRA developed at BT on a 32 node network (Figure 10.8) with 340 VPs. The DRA messages were assumed to be 64 bits long which take 10 msec to be processed within nodes, and are transmitted between nodes at 64 kbit/s. The actual VP crossconnection time was assumed to be 20 msec. The results are plotted in Figure 10.9, which shows the percentage of restored VPs against time, averaged over all possible single span failures. It can be seen that > 70% of failed VPs are recovered in under 2 seconds, while 100% restoration of VPs is achieved in around 7 seconds. Such results are very useful in demonstrating what can be achieved if certain node processing times are assumed. Ongoing testbed development has aimed at reducing the message processing overhead in an effort to achieve restoration times of the order of a few seconds[25].

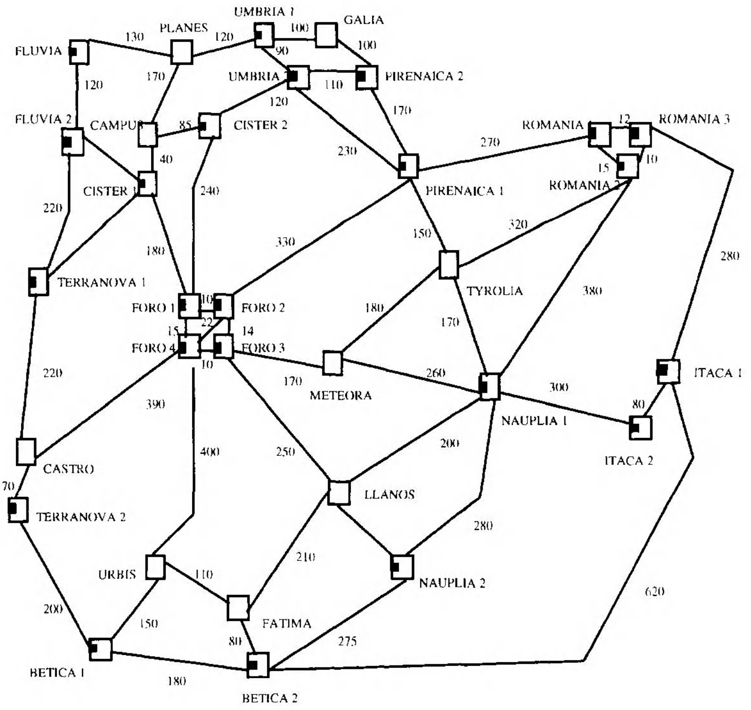


Figure 10.8. 32 node test network for DRA simulations

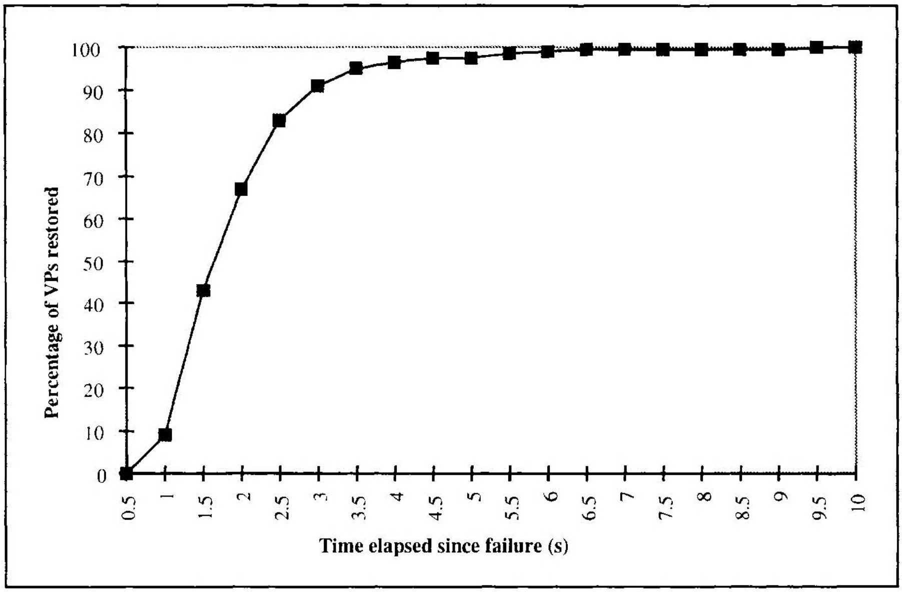


Figure 10.9. Cumulative % restoration resulting from path DRA simulation

3.4.4 Self-healing with Semi-Dedicated Backup VPs

A consistent feature of the self-healing schemes detailed thus far is that both routes and bandwidth are allocated on-demand. A technique with properties falling between protection and "on-demand" self-healing is the semi-dedicated backup VP. Here, backup routes which are disjoint from the working routes are pre-assigned, however spare capacity may be shared for restoration from the most common type of failures like single spans[26-29], as illustrated in Figure 10.10.

It is possible to provision spare capacity on this basis, so that as long as the failure which occurs has been planned for, the backup VP may be activated with sufficient bandwidth to support the re-routed working traffic. Nevertheless, due to the possibility of unexpected multiple failures, confirmation of the available resources on a backup VP is essential. Two separate approaches to capacity confirmation are outlined in [29]:

- To defer switchover from working to protection routes until all links of the backup route have been checked for capacity availability.
- To switch traffic to the protection route *before* checking the availability of capacity on the backup path.

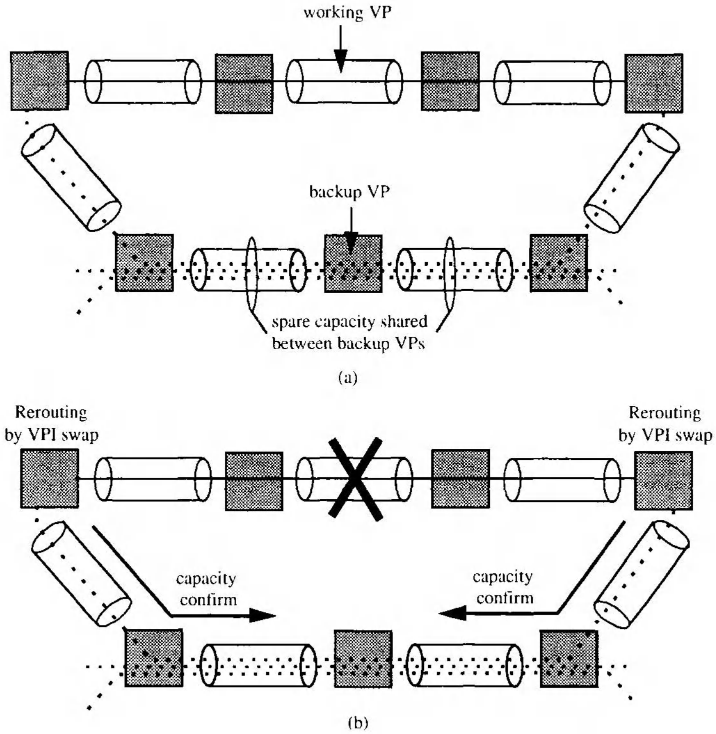


Figure 10.10. Semi-dedicated VP with pre-assigned route and on-demand bandwidth allocation

In the latter case, the cell loss priority (CLP) bit must be set in all diverted cells so that in the event of buffer congestion, switches will drop such cells thus avoiding inadvertent cell loss of other traffic which has not been directly affected by the fault[29]. In either case, if there is insufficient spare bandwidth on at least one link of the backup route, a negative acknowledgement cell must be returned to the re-routing point with any reserved capacity in preceding links relinquished. Different capacity confirmation protocols exist as detailed in [29], and this issue remains a topic for further study by the standards bodies[5].

The performance of semi-dedicated protection VPs in terms of restoration speed ought to be better than the techniques which rely on "on-demand" resource allocation, since no time is spent seeking and establishing actual routes. However, since some link-by-link processing is needed to allocate capacity, the method will be slower than straightforward protection. It has been shown that in a 30 node network with 435 VPs, restoration with semi-dedicated protection can be completed in under 2 seconds[29]. For a much larger number of VPs, restoration will be slower, perhaps taking tens of seconds. However, the VP grouping principle described earlier in the context of protection networks may be applied for semi-dedicated protection to reduce processing overheads and speed up restoration. The main advantage of the semi-dedicated approach is that spare capacity can be shared between other backup VPs, which reducing capital costs. Furthermore, as with 1:1 protection, there is the option to use the spare pool of capacity under normal conditions for low-priority traffic which could tolerate "bumping" in the event of a failure.

4. DISCUSSION

The wide range of services and traffic types on an ATM network means that different services will require certain levels of resilience[31]. A minimum degree of resilience could be provided to *all* services, with additional measures taken to upgrade the resilience of services with more stringent requirements. However, this may not be deemed cost-effective. Instead, resilience can be provided to customers that demand it, hence the matter of which resilience mechanisms to offer customers has to be addressed[4]. Table 1 shows the most suitable resilience mechanisms for different applications.

Dedicated protection could be offered to mission-critical data applications where service-continuity is vital. At the time of writing, ATM-layer protection switching protocols were being developed for standardisation[5], with as yet no switch vendor offering such functionality. In any case, the ability to switch traffic within tens of milliseconds following a major failure like a cable break ultimately depends on the successful administration and operation of VP grouping techniques, the principles of which are still being studied[5,12]. An interim solution is to provide dedicated protection at the SDH layer, though it has been shown that this is inherently more costly than ATM network protection[2].

For customers using non mission-critical applications, it is possible that automatic redial (whether it be for switched VCs/VPs or soft PVCs/PVPs) incurring temporary loss of service, will be tolerated. Restoration times from one customer to another will vary markedly, probably between a few seconds in the best case and minutes in the worst case. The obvious advantage of resilience mechanisms based on control signalling is that they are compliant with standards and as such are supported in off-the-shelf equipment. With distributed restoration algorithms (DRAs) meanwhile, there is currently no standardisation. Centralised restoration provides a reasonable near term solution for non-critical ATM services, and is supported by several ATM switch vendors.

Table 10.1. ATM resilience mechanisms for different applications

Resilience Scheme	***Applications***	***Speed***	***Cost***	***Viability***
Dedicated Protection	Mission critical data, telemedicine	Objective is 60 msec	High (reserved capacity plus VPI/VCIs)	Fast switching depends on effective grouping
Semi-Dedicated	Near mission critical, interruptable	< 10 secs	Medium (shared capacity plus VPI/VCIs)	Depends on effective grouping and capacity allocation
Auto-Redial	Residential, non mission critical data	10s secs -> mins	Low (capacity only)	Based on standard signalling
Centralised, with network management system	Residential, non mission critical data	Several mins	Related to sophistication of NMS	Viable, but limited to proprietary systems

It is debatable whether or not there is a need for a resilience mechanism like semi-dedicated backup VPs with cost and performance characteristics between dedicated protection and on-demand self-healing. If VP grouping is employed and OAM cell processing is fast enough, sub-second restoration should be possible. This would come at lower cost than protection since spare capacity can be shared between backup routes, *and* low priority traffic can be carried over spare capacity under normal operational circumstances. Standards activities for defining an actual bandwidth allocation protocol for backup VPs are in the early stages of progress[32].

5. CONCLUSIONS

Due to the perceived vulnerability of very high speed networks to many different kinds of failure, there is an increasing demand on network planners to incorporate resilience mechanisms into architectural designs. This paper has addressed the challenges involved in embedding resilience into wide area asynchronous transfer mode (ATM) networks. The key conclusions may be summarised as follows:

Whilst ATM-layer protection mechanisms are being developed, SDH-layer restoration may be exploited within a multi-layer framework, such as that defined in [33]. SDH restoration architectures include 1+1 protection with line systems and crossconnects, and Shared Protection Rings (SPRings), as detailed in [34].

Dedicated ATM-layer protection could be offered to mission-critical data applications. Standards activities in protection protocols and VP grouping should encourage vendors to support this functionality in the near future.

Resilience mechanisms based on control-layer signalling protocols like PNNI could ensure a “best-effort” class of restoration for non-critical services. Indeed, such a mechanism could represent a minimal level of resilience supplied to all ATM network users.

There may be scope for providing resilience with performance close to protection, but at much reduced cost by employing semi-dedicated backup VPs. Standards efforts in defining bandwidth allocation protocols have commenced.

As networks based on ATM technology become more widely deployed, real customer resilience requirements should become clearer, whilst ongoing technological developments should ensure the capability to support these requirements. Of increasing interest will be the impact that the extensive deployment of TCP/IP networks based on Gigabit routers has on underlying transport layers such as ATM and SDH. It is vital to explore the implications of such multi-layer architectures on core network resilience.

References

[1] J.C. McDonald. “Public Network Integrity- Avoiding a Crisis in Trust”, IEEE J-SAC, 12(1):5-12, January 1994.

[2] P.A. Veitch, D. Johnson and I. Hawker. “Design of Resilient Core ATM Networks”, in proceedings of IEEE Globecom ’97, Phoenix, AZ, November 1997.

[3] K. Struyve et al. “Design and Evaluation of Multi-Layer Survivability for SDH-Based ATM Networks”, in proceedings of IEEE Globecom ’97, Phoenix, AZ, November 1997.

[4] Paul Veitch and Dave Johnson. “ATM Network Resilience”, IEEE Network, September/October 1997.

[5] J. Anderson (Editor). “ATM Network Survivability Architectures and Mechanisms”, Q.F/13 Report, November 1996.

[6] D. Johnson. “Survivability Strategies for Broadband Networks”, IEEE Globecom ‘96, London, pp 452-456.

[7] ITU-T Rec. I.610. “B-ISDN Operation and Maintenance Principles and Functions”, ITU-T, 1993.

[8] Wu, T-H. “Fiber Network Service Survivability”, Artech House, 1992.

[9] Kajiyama, Y., Tatsuno, H. and Tokura, N. “Virtual Path Recovery Switching and Hitless reversion Switching in 180 km ATM Self-healing Ring”, Electronics Letters, Vol. 30, No. 11.

[10] ITU-T Rec. I.361. “B-ISDN ATM Layer Specification”, ITU-T, 1993.

[11] H. Hadama, R. Kawamura and K-I. Sato. “Virtual Path Restoration Techniques Based on Centralized Control Functions”, Electronics and Communications in Japan, Part 1, Vol 78, No. 3, 1995.

[12] T. Noh. “End-to-End Self-Healing SDH/ATM Networks”, IEEE Globecom ‘96, London, U.K., November 1996, pp 1877-1881.

[13] C-W Chao et al: “FASTAR Platform Gives the Network a Competitive Edge”, AT&T Technical Journal, July/August 94 pp 69-81

[14] K. Yamagishi, N. Sasaki and K. Morino "An Implementation of a TMN-Based
SDH Management System in Japan", IEEE Communications Magazine, March 1995.

[15] A. Alles. “ATM Internetworking”, Cisco Systems, 1995.

[16] D. Kushi and E. M. Spiegel. “Signalling Procedures for Fault Tolerant Connections”, ATM Forum/97-0391R1.

[17] Y. T’Joens et al. “Modified Procedures for Fast Connection Recovery in PNNI Networks”, ATM Forum/97-0671.

[18] H. Masuo et al. “Proposal for a Working Document for Fault Tolerance in PNNI”, ATM Forum/97-0321.

[19] General DataComm Product Specification, “Self-Healing ATM Networks: Using the GDC APEX® ATM Switch to Construct Resilient ATM WANs”, 1996.

[20] W. D. Grover, B.D. Venables, M.H. MacGregor and J.H. Sandham. “Development and Performance Assessment of a Distributed Asynchronous Protocol for Real-Time Network Restoration”, IEEE J-SAC, January 1991, pp 112-125.

[21] R. Kawamura, K-I. Sato and I. Tokizawa. “Self-Healing ATM Network Techniques Utilizing Virtual Paths”, Networks ‘92, Kobe, Japan, May 1992.

[22] H. Fujii and N. Yoshikai. “Restoration Message Transfer Mechanism and Restoration Characteristics of Double-Search Self-Healing ATM Network”, IEEE J-SAC, January 1994, pp 149-157.

[23] M. Azuma et al. "Network Restoration Algorithm for Multimedia Communication Services and its Performance Characteristics", IEICE Transactions in Communications, July 1995, pp 987-994.

[24] L. Nederlof, H. Vanderstraeten and P. Vankwikelberge. "A New Distributed Restoration Algorithm to Protect ATM Meshed Networks Against Link and Node Failures", ISS '95, Berlin, pp 398-402.

[25] L. Nederlof, L. Van Hauwermeiren, P. A. Veitch, C. O'Shea, D. Johnson and P. Gaynord. "Demonstration of Distributed Restoration in an ATM Network", in proceedings of ISS '97, Toronto, September 1997.

[26] R. Kawamura, K-I. Sato and I. Tokizawa. "Self-Healing ATM Networks Based on Virtual Path Concept", IEEE J-SAC, January 1994, pp 120-127.

[27] C.K. Jones and R.R. Henry. "A Fast ATM Rerouting Algorithm for Networks with Unreliable Links", IEEE ICC '94, New Orleans, pp 91-95.

[28] R. Cohen and A. Segall. "Connection Management and Rerouting in ATM Networks", IEEE Infocom '94, Toronto, pp 184-191.

[29] P. A. Veitch, I. Hawker and D.G. Smith. "Administration of Restorable Virtual Path Mesh Networks", IEEE Communications Magazine, December 1996, pp 96-101.

[30] T. Chen, S. Liu, D. Wang, V.K. Samalam, M.J. Procanik & D. Kavouspour. "Monitoring and Control of ATM Networks Using Special Cells", IEEE Network, September/October 1996, pp 28-38.

[31] T. Yahara and R. Kawamura. "Virtual Path Self-healing Scheme Based on Multi-Reliability ATM Network Concept", in proceedings of IEEE Globecom '97, Phoenix, November 1997.

[32] H. Ohta. "Proposed Semi-Dedicated VP Automatic Protection Switching Method", ITU-T study document, September 1997.

[33] ITU-T Rec. G.803. "Architectures of Transport Networks Based on the Synchronous Digital Hierarchy (SDH)", 1993.

[34] P.A. Veitch, P.R. Richards, P.J. McCartney & D. Johnson. "Alternative Transport Architectures for Core ATM Networks", BT Technology Journal, July 1998.

Chapter 11

IP SWITCHING OVER ATM NETWORKS

Andreas Skliros

SOFOSNET Ltd,
3 G. Labraki Str, Aspropyrgos, 19300, Greece,
E-mail: askliros@hol.gr

Abstract: The enormous growth in the Internet is presenting a major challenge to today's Internet Service Providers. It is critical for an ISP to keep pace with the latest technologies and network architectures to insure its ability to deliver the required quality of service at a reasonable price while remaining a profitable commercial venture. Although IP is the most commonly used networking protocol, the increase of the switching capacity of routers cannot meet the explosion of Internet traffic. On the other hand, ATM promises both high transmission speed and QoS guarantees. IP switching is a set of protocols which can be used to combine the flexibility of IP software with the speed of ATM hardware. It is a cost-effective solution which can tackle the problems of IP congestion and poor QoS for multimedia applications.

Keywords: ATM, IP, IP Switching, MPOA, LANE, GSMP, IFMP

1. INTRODUCTION

The Internet Protocol (IP) has become the *de facto* standard network-layer protocol due to its ability to scale from the desktop to the global Internet, and due to the unprecedented growth of Internet and corporate intranets over the last few years. However, today's IP networks are rapidly running out of steam. With the advent of faster workstations, client-server computing and bandwidth-hungry applications, network managers and users are increasingly experiencing traffic congestion problems on their networks. Such problems can take the form of highly variable network response times, higher network failure rates and the inability to support delay-sensitive applications.

ATM is receiving a tremendous amount of attention as a switching technology promising scalability, dramatically increased throughput, and support for multiple types of network traffic through quality-of-service (QoS) guarantees. Although, ATM is a high-speed, scalable, multiservice technology that is the cornerstone of tomorrow's router-less networks, it is also a networking technology so different from current networking architectures that there is no clear migration path to it.

The success of ATM as a future networking technology, however, hinges on its ability to effectively support existing network traffic, a task made difficult by ATM's connection-oriented architecture which creates the need for an additional set of very complex, untested multi-layer protocols. Many of these protocols duplicate the functionality of the well-established TCP/IP protocol suite, and the learning curve associated with these complex new protocols dramatically increases the cost of ownership of ATM devices for network managers.

This tutorial describes in more detail the problems in today's IP networks and presents the IP Switching solution. Section 2 describes the unprecedented growth of IP traffic along with its resulting problems while section 3 reviews various approaches for integrating IP and ATM. Section 4 presents the IP switching functionality and section 5 summarises its advantages.

2. THE GROWTH OF INTERNET (IP)

Originally designed for use on the ARPANET, the Internet Protocol has evolved into the dominant network-layer protocol in use today. All major operating systems now include an implementation of IP, enabling IP and its companion transport-layer protocol, the Transmission Control Protocol (TCP), to be used universally across virtually all hardware platforms. The fundamental driver enabling IP to "win" the networking protocol war is its tremendous scalability. Unlike other internetworking protocols, IP has successfully been implemented in networks comprised of only a few users to enterprise-size networks, and even the global Internet.

The Internet has doubled in size every year since 1988 and, as of July 1996, reached an estimated 12.9 million hosts on over 135,000 interconnected TCP/IP networks. In only a few years, users have created more than 280,000 different multimedia "sites" of information, entertainment and advertising via the World Wide Web, and these sites are accessed by the now ubiquitous Web browser.

While IP is a robust protocol the traditional IP packet-forwarding device on which the Internet is based, the *IP router,* is beginning to show signs of inadequacy. Routers are expensive, complex and of limited throughput when compared to emerging switching technology. Today's routers are roughly *four to five times* as fast as routers five years ago, while transport rates and switching capacity have increased at *much faster rates* over this same time period.

The Figure 11.1 below shows this disparity by comparing the relative increase in router performance versus the growth in traffic on the Internet. (The number of networks connected is used here as a proxy for the amount of traffic on the Internet.)

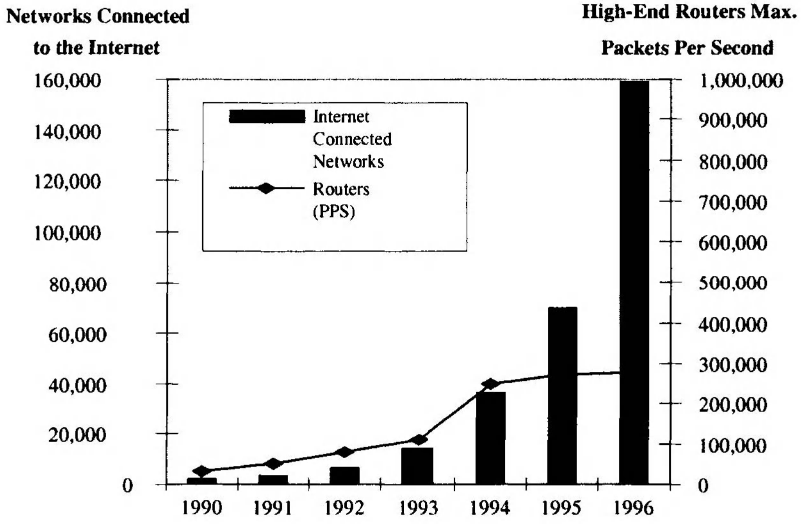


Figure 11.1. Comparing Router Performance Increases vs. Internet Growth

To support the increased traffic demand of the Internet and large enterprise-wide networks, IP needs to go faster and cost less. Additionally, to support the emerging demand for real-time and multimedia applications, IP also needs to support QoS selection.

3. INTEGRATING IP AND ATM

The global Internet and the Internet Protocol (IP) on which it is based have witnessed unprecedented growth and acceptance, with IP emerging as the dominant network-layer protocol. On the other hand, ATM is perceived as the proper WAN solution of the future which can offer high speeds and QoS guarantees. The idea is simple; since both IP and ATM have such significant advantages why not integrate them. The ideal solution would be to achieve the seemingly incompatible goals of seamlessly integrating emerging high-speed ATM switching technology with existing IP networks while avoiding router bottlenecks, increased network management complexity and large, flat networks.

Major networking standards bodies have reacted to these trends by developing a number of new networking architectures. ATM Forum and the Internet Engineering Task Force (IETF) attempt to develop specifications linking existing LAN environments with switched ATM networks, including the LAN emulation (LANE) specification, the Classical IP (CIP) over ATM specification, the Next Hop Resolution Protocol (NHRP) and the Multi-Protocol Over ATM (MPOA) specification.

However, integrating ATM switches into existing IP networks requires the resolution of a technological incongruity. The heart of the problem is to make use of the unparalleled speed and capacity of a connection-oriented ATM switch fabric without sacrificing the scalability and flexibility that come from the connectionless nature of IP. Solving this problem requires either discarding the connectionless nature of IP and allowing ATM to operate as a multi-layer protocol, or discarding the connection-oriented aspects of ATM and allowing connectionless IP functions directly on top of ATM switching hardware.

Currently, the only approaches to integrate ATM into existing IP-based networks have been single-subnet solutions such as Classical IP over ATM (CIP) or LAN Emulation (LANE).

These approaches enable an ATM switch to emulate the functionality of an Ethernet (or other Layer 2) segment. Fundamentally, while LANE and CIP are relatively simple in concept and require no modifications to IP, neither scales well to larger networks because all communication between emulated LANs or logical IP subnets must proceed via routers, most often via so-called “one-armed” routers (routers with a single ATM interface). These routers become a significant throughput bottleneck, especially when a relatively low percentage of network traffic remains within a single subnet, as is increasingly becoming the case with the deployment of centralised server farms, corporate intranets and other applications that are distributed across the entire enterprise. LANE and CIP are simply not acceptable solutions for alleviating backbone congestion, and even as single subnet solutions, these approaches introduce considerable complexity into the network compared to Ethernet switching.

The IETF’s Internetworking Over Non-Broadcast Multi-Access (NBMA) Working Group is attempting to address the issue of communication between different logical subnets within a NBMA network, such as ATM or frame relay. The problem consists of locating the exit point on the cloud nearest to a given destination and obtaining the ATM address for that exit point. The signalling protocol in the control software of the switch (Q.2931 for ATM) may then be used to establish a connection across the cloud to the exit point.

The Next Hop Resolution Protocol (NHRP) has been proposed as a routing protocol to perform this function. NHRP and the Routing Over Large Clouds (ROLC) architecture have been criticised because of their complexity and their inability to scale to very large networks. The Multi-Protocol Over ATM (MPOA) group is addressing the same issues within the ATM Forum and is encountering similar problems in the areas of scalability and manageability. Both of these solutions have not been completed.

Recognising these trends and the need for a measure of simplicity in the internetworking environment, IP Switching solution proposed by *Ipsilon Networks.* It allows networks to efficiently route IP traffic using fast ATM switching technology, dynamically shifting between store-and-forward IP routing and cut-through IP Switching in order to optimise traffic throughput. IP Switching approach allows the integration of complete IP routing functionality directly with ATM switching hardware. The resulting IP Switch combines the simplicity, scalability and robustness of IP with the speed, capacity and multiservice traffic capabilities of ATM.

IP Switching appears to be a better (and more timely) solution which adds IP routing intelligence directly to ATM switches without sacrificing the performance of these switches. IP is sufficient as a Layer 3 protocol, having proven its ability to scale to networks as large as the global Internet. It is also a robust, technology-independent protocol with implementations available for virtually every operating system. In contrast, ATM requires the adoption of an alternative set of new and untested protocols, many of which duplicate the functionality of TCP/IP. Initial implementations of some of these protocols have been problematic, as evidenced by the unacceptably long SVC (switched virtual circuit) connection set-up times currently plaguing the ATM signalling and routing protocols.

4. THE IP SWITCHING ARCHITECTURE

In IP Switching architecture, the underlying switching fabric can be ATM, frame relay, or even LAN switching fabrics such as gigabit Ethernet. However, IP Switching implementations have focused on ATM because of the compelling price/performance, multicast and quality-of-service characteristics of ATM switches. For the purposes of simplicity and clarity in this tutorial paper, we will continue to refer to ATM as the underlying switching fabric in an IP Switch.

The IP Switching approach integrates fast switching technology with IP routing to enable network managers to construct large IP networks without sacrificing the scalability and functionality of IP routing or the performance of the high-speed switches.

The IP Switching solution is based on two key ideas:

1. *Add the complete IP routing functionality directly* ***"on top"*** *of ATM switching hardware by using the IFMP and GSMP protocols to communicate and control the ATM switch, and*
2. *IP packets can be classified as belonging to a* ***flow*** *of similar packets based on certain common characteristics.*

Combining these two points, IP Switching marries IP routing functionality and high-speed switching performance to create a new class of networking device known as *IP Switch.* An IP Switch can dynamically shift between forwarding packets via standard hop-by-hop connectionless IP routing and forwarding packets via the high-throughput ATM switching hardware, depending on the flow classification of the traffic.

The Figure 11.2 below shows a conceptual diagram of an IP Switch. Note that IP Switching replaces the connection-oriented signalling specifications of ATM (SSCOP, Q.2931, etc.) and any new bridging or routing specifications (LANE, PNNI, MPOA, NHRP, etc.) with the IP routing protocols (RIP, OSPF, BGP, etc.) that have become the *de facto* standards for internetworking.

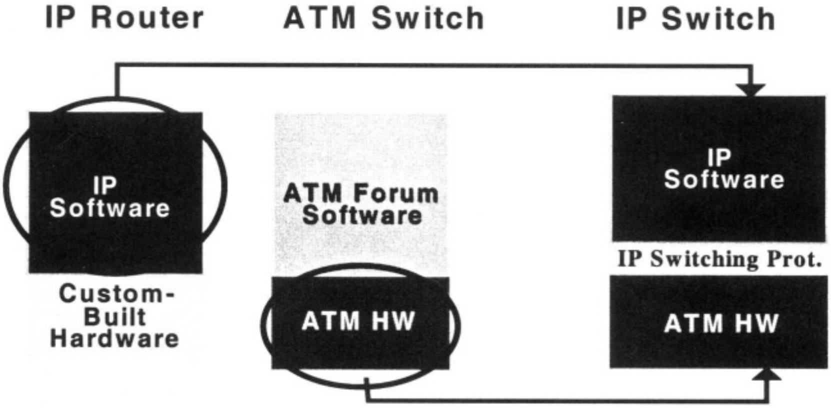


Figure 11.2. Components of an IP Switch

4.1 IP SWITCHING OPERATION

The fundamental idea of IP Switching is to leverage the performance of routers by requiring that it forwards only a small fraction of the traffic and by off-loading the majority of that traffic to the ATM switch.

In order a router to take advantage of the high performance of its associated ATM switch, the IP Switching software must be able to decide when to switch the packet (in cells) directly in the switching hardware without the burden of software processing.

The IP Switching software makes this decision by classifying IP packets as part of either a long-lived or short-lived flow, where ***a flow of IP packets*** *is simply a sequence of packets sentfrom a particular source to a particular destination that may share certain other characteristics such as protocol, TCP/UDP port number, etc.*

In the IP Switching architecture, long duration flows, or flows likely to "last" a long time (such as a file transfer or World Wide Web image download), are ***"cut through"*** in the switching hardware, while short duration flows (such as DNS queries) are forwarded in the standard hop-by-hop manner of a traditional router through the IP Switch Controller.

Phase1

The Figure 11.3 below illustrates the operation of an IP Switch. For the purpose of the example, we assume a simple traffic flow from the upstream node to an IP Switch and on to a downstream node.

The upstream node could be any of a number of devices including another IP Switch, a router, a gateway or a directly attached host or server with IP Switching functionality.

In default operation, IP packets are forwarded hop-by-hop in a connectionless manner using a default VPI/VCI from the upstream node to the IP Switch and on to the downstream node. Within the IP Switch, cells are received over an ATM switch port, sent up to the IP Switch Controller and reassembled into IP packets to be mapped against routing tables and forwarded by the IP Switching routing software in the same manner as a traditional IP router.

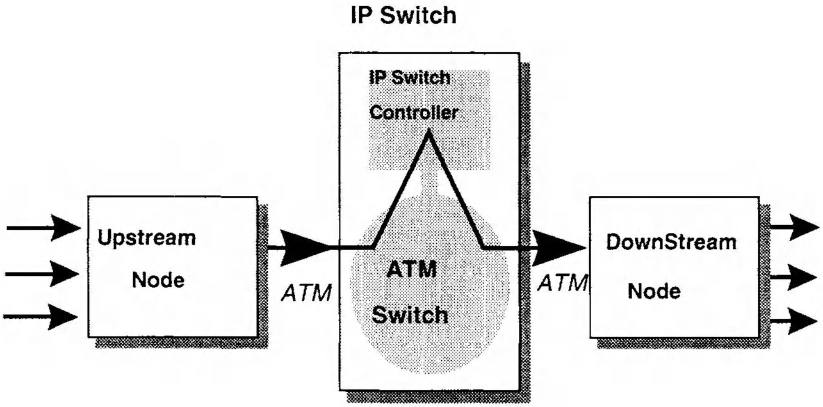


Figure 11.3. Phase 1 of IP Switching

Phase 2

The IP Switching software also performs a flow classification and makes a decision as to whether future IP packets matching the flow classification (i.e., belonging to that flow) can benefit from being switched in the ATM hardware, bypassing the IP Switch Controller. If the IP Switching software decides that a particular flow is a candidate for switching, it sends a redirect message to the upstre am node requesting that future IP packets belonging to that particular flow (as identified by the unique IP header information related to that flow) be sent over the ATM link with a specific VPI/VCI. This initial redirection is depicted below.

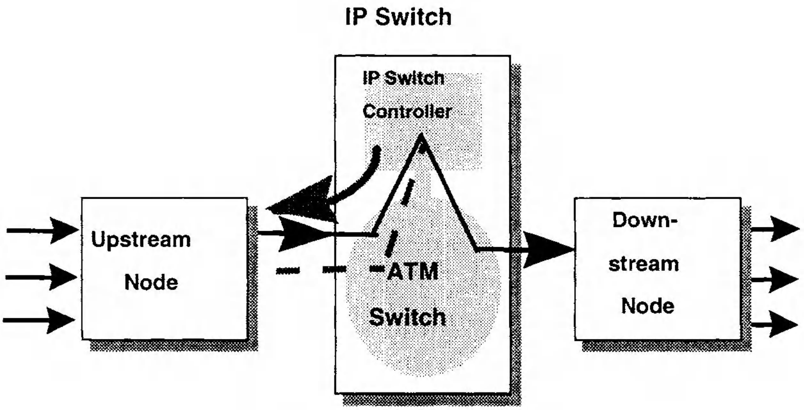


Figure 11.4. Phase 2 of IP Switching

Phase 3

In the same manner, the downstream node may also issue a redirect message for the same flow after performing the flow classification process. In this case, it sends a redirect message to its upstream neighbour, the IP Switch.

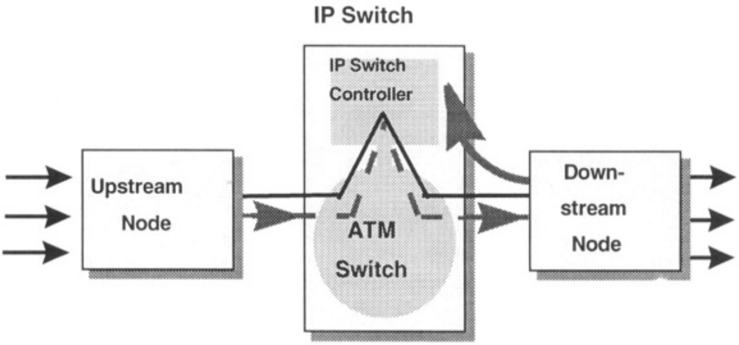


Figure 11.5. Phase 3 of IP Switching

As explained in the previous steps, the packets belonging to the particular flow have now been assigned to unique VCs both upstream and downstream of the IP Switch.

Phase 4

Subsequent traffic belonging to this flow can now be switched completely in the attached ATM switch, thereby off-loading the DP Switch Controller from having to route or process any additional packets belonging to that flow. This process is illustrated below.

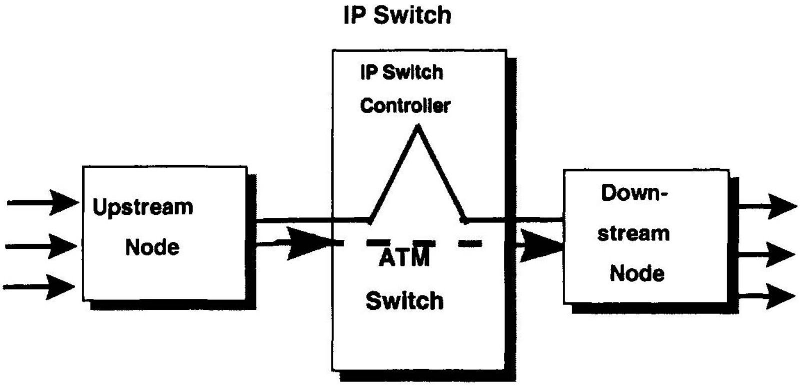


Figure 11.6. Phase 4 of IP Switching

As more and more IP traffic flows are dynamically "pushed down" to the ATM switching fabric, the overall packet throughput of the IP Switch approaches that of the ATM switch. In a well-designed ATM switch, that throughput should approach the combined wire speed of all ATM ports on the switch. Based on a series of traffic traces taken from the core of the Internet and using IP Switching flow classification algorithms, it can be estimated that approximately 90% of traffic (measured in bytes) would be classified as suitable for switching in hardware.

4.2 IP SWITCHING PROTOCOLS

IP Switching is based on two publicly available protocols to support its operation.

The first protocol, known as the *General Switch Management Protocol (GSMP)*, when implemented on an ATM switch, enables the IP Switching software running on the IP Switch Controller to communicate with and control the attached, vendor-independent switch.

The second protocol, known as the *IP Switching Flow Management Protocol (IFMP)* , enables communication between neighbouring devices allowing these devices to issue and respond to redirect messages. Published specifications of the IP Switching IP Switch protocols have been issued by the IETF as informational RFCs.

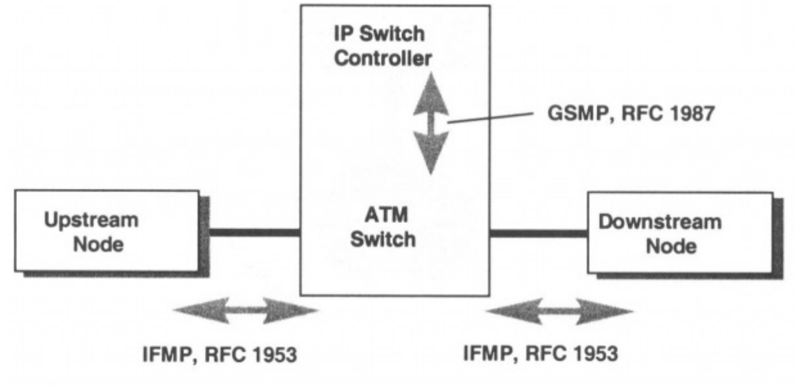


Figure 11.7. IP Switching Protocols

The IP Switching solution and protocols (RFC1953, RFC1954 and RFC 1987) represents a much simpler alternative than the LANE and MPOA methods proposed by the ATM Forum (as shown in the Figure 11.8 below).

The software required to implement the IP Switching solution in hosts and packet-forwarding devices is dramatically smaller than the software required to implement LANE or MPOA. The size of the software (number of lines of software code for implementing the protocols) serves as a proxy for the relative complexity of the competing architectures and gives the network manager a good idea of the additional training, education and knowledge required to successfully implement and manage the network.

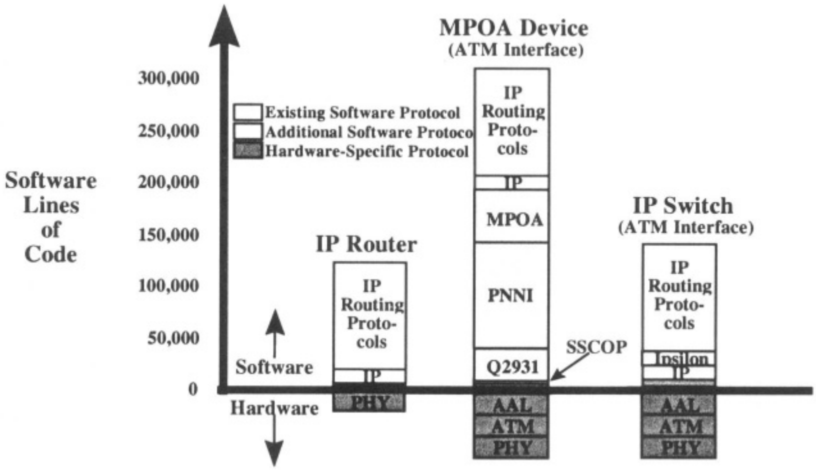


Figure 11.8. Protocol Stack Comparison

4.3 IP SWITCHING QUALITY OF SERVICE

Today's routers have very little ability to manage the quality of service offered to network users. Work on IP protocols such as RSVP to manage quality of service factors such as network delay holds promise, is still in progress. Routers were never designed with quality of service in mind, they were designed to provide connectivity. Switches, on the other hand, offer many more mechanisms to manage quality of service.

Most Frame Relay and ATM switches provide extensive mechanisms to control QoS through sophisticated traffic management. Most switches can fairly supply bandwidth to each user based on a predetermined traffic contract and protect each user from each other user. Routers, on the other hand, have little ability to protect one user's traffic from another user's traffic.

In considering QoS, it is helpful to frame the discussion in terms of two events:

- *a user or application requesting a certain class of service and*
- *the fulfilment of that request by the underlying network.*

Since IP Switches are managed and controlled in the same manner as IP routers, they are able to utilise any method that a traditional router would use to respond to user requests for quality of service, including the proposed RSVP specification. However, fulfilling QoS requests is much easier for IP Switching than for traditional routers since an IP Switch takes advantage of the queuing features inherent in its ATM switching hardware. Although there are implementation differences from one ATM switch to the next (which can result in notable performance differences), such queuing is essentially similar among ATM switches and basically involves buffer manipulation to allow for the prioritisation of certain streams of cells.

The IP Switching software is able to map QoS requests directly into the queuing capabilities of an ATM switching fabric. In contrast, traditional routers can only do an average job of fulfilling QoS requests through software-intensive techniques such as weighted fair queuing, but even then, this will result in a significant decrease in the throughput of these routers by consuming scarce processor and memory resources.

While waiting for RSVP to become a more widely adopted standard supported by a large number of host applications, IP Switching uses the combination of the IP Switching flow classification software and the underlying queuing features of the ATM hardware to offer local policy-based QoS today. That is, IP Switching enables network administrators to prioritise which applications (based on TCP or UDP port numbers) or which users (based on IP addresses) receive the highest and lowest QoS within an IP Switched network.

The ATM Forum is currently struggling to find a solution for mapping RSVP QoS requests into the queuing features of ATM switches. The reason for the difficulty in developing a solution is fairly straightforward – RSVP

QoS requests are receiver-initiated (initiated by the recipient of the data and sent back toward the sender), while ATM connection-oriented call set-ups are sender-initiated (initiated by the sender of the data prior to sending any data.) Resolving this fundamental architectural discrepancy is proving to be very difficult. The ATM Forum has also encountered difficulty in specifying an API for applications to take advantage of native ATM QoS features. Currently, no such standard API exists, so there is no method of taking advantage of ATM QoS other than by developing a proprietary API.

5. CONCLUSIONS

IP has effectively "won" the networking protocol war, but the increasing traffic demands of the Internet and many corporate networks require that IP go faster and support quality of service. These problems can be solved if the throughput increase of routers meets the increase of IP growth and if the IP protocol is modified substantially to address the QoS issues.

Both these solutions require a significant effort. Since ATM can effectively cope with both these IP problems, an integration of IP and ATM based on the IP Switching approach is the most cost-effective solution. Networks based on IP Switching offer several benefits and enhancements over traditional IP router-based networks, emerging switch-based networks and ATM networks incorporating specifications such as LAN emulation, classical IP over ATM or MPOA. The IP switching benefits are summarised below.

- IP Switches solve the backbone congestion problem by integrating high-speed switching technology into existing IP-based networks.
- IP Switches can shift dynamically between store-and-forward IP routing and cut-through IP switching to optimise traffic throughput.
- IP Switches scale to much higher IP packet throughput than conventional routers by using a switch fabric, rather than a shared bus, as a backplane.
- IP Switches offers a 10 to 1 price/performance advantage over conventional alternatives by exploiting industry-wide advances in ATM switching hardware.
- IP Switches can support multiple levels of QoS based on type of application and/or IP source and destination address

References

[1] IETF RFC 1577, M. Laubach, "Classical IP and ARP over ATM,", January 1994.
[2] ATM Forum, "LAN emulation over ATM," Version 1.0, January 1995.
[3] International Data Corporation, *Computer Networking Architectures*, 1995

[4] Network Wizards, *Internet Domain Survey,* July 1996.
[5] J. Barksdale, "The Revolution in Communications and Commerce," *ComNet Keynote Session*, January 31, 1996.
[6] IETF RFC 1953, P. Newman, W. L. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon and G. Minshall, "Ipsilon Flow Management Protocol Specification for IPv4, Version 1.0", , May 1996.
[7] IETF RFC 1954, P. Newman, W. L. Edwards, R. Hinden, E. Hoffman, F. Ching Liaw, T. Lyon and G. Minshall, "Transmission of Flow Labelled IPv4 on ATM Data Links, Ipsilon Version 1.0", , May 1996.
[8] P. Newman, T. Lyon and G. Minshall, "Flow-Labelled IP: Connectionless ATM Under IP", Networld + Interop, Las Vegas, April 1996.
[9] Ipsilon Networks, "An Introduction to IP Switching", Technical White Paper, 1996.
[10] RFC 1987 P. Newman, W. Edwards, R. Hinden, E.Hoffman, et.al,"Ipsilon's General Switch Management Protocol Specification", Ver.1.1, Aug. '96

Dr Andreas SKLIROS received his B.Sc. in Economic Sciences from the University of Athens, Greece and his Ph.D. in Performance Modelling of Computer Communication Networks, Univ. Bradford, UK (1991). Following, he joined BT Labs working as a consultant in the area of ATM Network Performance. His main responsibilities included reviewing and contributing to ATM Standards (ITU, ATM-Forum) and performance studies related with ATM traffic management & control functions (CAC, policing) and QoS of ATM services. He also worked for Telematics International, a packet switch manufacturer, and he was responsible for the design of traffic management and control functions of a new ATM switch, involved with ATM standards, academic institutions, EU projects, ATM testbeds and trial networks. He continued with ECI Telecom as a marketing consultant in the areas of IP Switching, IP Telephony and new IP standards. He is currently the managing director of SOFOSNET and he is interested in IP telephony, and IP multimedia applications in the fields of Electronic Commerce and Teleducation. He has also published several papers for ATM and IP networks.

Chapter 12

AN APPROACH FOR TRAFFIC MANAGEMENT OVER G.983 ATM-BASED PASSIVE OPTICAL NETWORKS

Maurice Gagnaire
ENST InfRes
46, rue Barrault, 75634 Paris
France
gagnaire@enst.fr

Sašo Stojanovski
ENST InfRes
46, rue Barrault, 75634 Paris
France
sassos@enst.fr

Abstract A new generation of access networks is necessary for the provision of broadband services. ATM Passive Optical Networks (APON) are considered as a promising alternative among other numerous technologies based either on copper pairs, coaxial cables or wireless infrastructures. An APON is a point-to-multipoint broadcast system in the downstream direction and a multipoint-to-point shared medium in the upstream direction. A Medium Access Control (MAC) protocol has to be used for the upstream traffic in order to arbitrate concurrent access. The aim of this paper is first to describe the APON system architecture and the physical layer frame format as standardized by ITU-T in the G.983 recommendation. We then propose a possible approach for traffic management over such systems considering three aspects: MAC protocols, service disciplines and buffer management. Two types of traffic are considered, stream and elastic. Stream traffic refers to flows generated by real-time applications such as voice or video, whereas elastic traffic refers to TCP/IP flows. In the last part of our paper, we evaluate the performance of this approach by means of computer simulations.

Keywords: ATM, GFR, MAC protocols, service disciplines, buffer management.

1. INTRODUCTION

The APON systems are designed for a distance of 10km and a splitting ratio of 641, as shown in Figure 12.1. The first parameter covers more than 98% of today's narrowband local loops ([1]), whereas the second is conditioned by the optical power budget. The two extremities of an APON system are the Optical Line Termination (OLT) and the Optical Network Units (ONU). Both a symmetrical 155 Mbit/s and an asymmetrical 622/155 Mbit/s interface are defined in the ITU-T G.983 recommendation. Copper wires may be used to connect more than one customer to the same ONU in the Fibre-To-The-Kerb (FTTK) configuration. An APON system may be seen as a traffic concentrating device which reduces the need for an additional concentration stage at the central office.

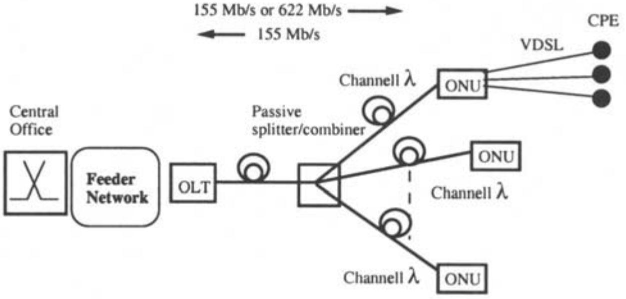


Figure 12.1 APON access system.

A continuous flow of ATM cells is generated by the OLT in the downstream direction. In the upstream direction, ATM cells are encapsulated in APON packets at the ONUs. An ONU filters only information by which it is concerned, according to the VPI/VCI field in the ATM cells' header. Downstream information is encrypted in order to offer privacy and security to the customers. The ONUs are allowed to send an ATM cell only after receiving an explicit permit from the OLT. The transmission being performed in bursty mode, the OLT has to synchronise with every single upstream transmission. The distance between an ONU and the OLT varying from one ONU to another, power ranging and clock ranging are carried out at the OLT. In order to avoid collisions between upstream APON packets, distance equalisation is performed via a ranging procedure. Dynamic bandwidth allocation is implemented via a request/permit mechanism. The ONUs are periodically polled by the OLT to send their bandwidth requests. Based on that information the OLT issues permits. The request/permit mechanism introduces an inherent access delay lower-bounded by the roundtrip delay and is usually a multiple thereof. The roundtrip delay equals 0.1ms in an APON system. In addition to MAC layer functions, the ONUs are supposed to do VPI/VCI translation, buffer management and cell scheduling. In order to perform dynamic bandwidth allocation, both the OLT and the ONUs must have knowledge of the traffic contract for each established ATM connection. This can be done by intercepting the signalling messages in the ATM control plane, or via the Broadband Bearer Connection Control (B-BCC) protocol.

The paper is organized as follows. In section 2, we describe the G.983 physical layer frame format. In section 3 and 4, we discuss several possible approaches for service disciplines and buffer management. In section 5, we evaluate these approaches by means of computer simulations.

2. G.983 PROVISIONS FOR MAC PROTOCOLS

The APON frame format is defined in the ITU-T recommendation G.983. It is illustrated in Figure 12.2. The downstream frame (*downframe*) is composed of 56 ATM cells. Among these, two cells are dedicated for carrying permits to the ONUs. These two cells known as Physical Layer OAM cells (PLOAM) contain 27 and 26 permits respectively. Each permit is identified by a single octet. In fact, six bits suffice to address 64 ONUs. The remaining two bits are used to identify slots for ranging and polling purposes, and possibly for permit "colours".

The upstream frame (*upframe*) is composed of 53 upstream slots (*upslots*). Each upslot is 56-octets long and is composed of the standard ATM cell (53 octets), preceded by a 3-octet physical layer preamble. There is no provision for piggybacked requests. The ONUs are periodically polled via special upslots known as Divided Slots (DivS). The DivS is signalled by a special grant in the downstream PLOAM cells. Like the rest of the upslots, the DivS is also 56-octet long and is divided in several minislots. Each minislot is used to poll a single ONU. The G.983 recommendations does not specify the size and the exact use of these minislots. The only detail that is standardised is that each minislot must start with the same 3-octet physical layer preamble. An example of Divided Slot is shown in Figure 12.3. It is composed of eight minislots, which means that 8 ONUs can be polled via a single DivS.

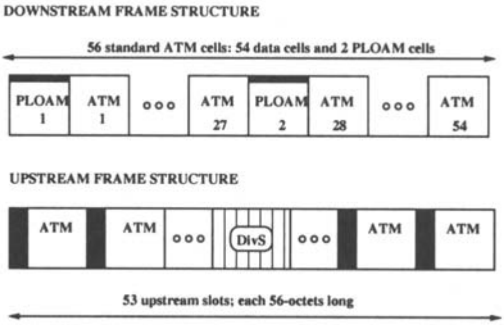

DOWNSTREAM FRAME STRUCTURE
56 standard ATM cells: 54 data cells and 2 PLOAM cells
PLOAM 1
ATM 1
ooo
ATM 27
PLOAM 2
ATM 28
ooo
ATM 54
UPSTREAM FRAME STRUCTURE
ATM
ATM
ooo
DivS
ooo
ATM
ATM
53 upstream slots; each 56-octets long

Figure 12.2 APON frame structure.

Each minislot is 7-octets long and consists of: a physical layer preamble (3 octets), a MAC information field (3 octets) and a CRC protection (1 octet). The DivS polling rate is programmable. Higher polling rate reduces the useful upstream capacity, but decreases the access delay for Stream traffic. Typically, one DivS slot is sent every 32 upslots. With this polling frequency, in an APON system with 64 ONUs each ONU is polled every 256 upslots (0.74ms). The MAC information field can be used in various manners. In the following we consider that any traffic can be categorized as either stream or elastic traffic (see [6]). Stream traffic refers to flows with rate-envelop constraints whereas elastic trafic is unconstrained. The QoS requirements for stream and elastic traffic are expressed in terms of delay and throughput respectively.

We propose two definitions for the minislot format, one for the independent shaping approach (minislot A) and one for the integrated shaping approach (minislot B). In this article, we consider the format shown in Figure 12.3. The MAC information field consists of three separate fields: st-MTC (MAC Transfer Capability), el-MTC and Rsrvd. The former two carry bandwidth requests for the corresponding MAC Transfer Capability, and the last octet is reserved for future use. We also assume that the permits are coloured i.e. they indicate the MTC for which they have been generated.

The traffic generated by the end-users is likely to be distorted upon arrival at the OLT. Therefore, the ATM cells received by the OLT are temporarily stored and thereafter retransmitted at instants that allow all the cells to be compliant to the connection's traffic contract in terms

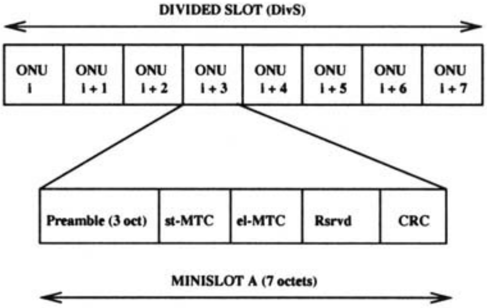


Figure 12.3 Divided slot and minislots.

of CDV requirements. Shaping is typically applied only to CBR and rt-VBR connections.

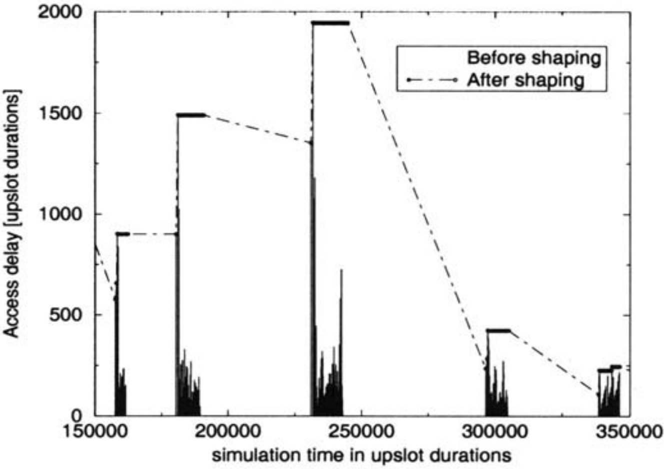


Figure 12.4 Example of distorted rt-VBR traffic flow which is re-shaped in the standalone shaper.

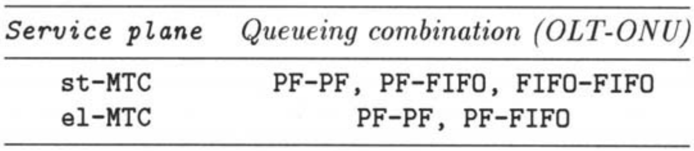
The use of a standalone shaper between the OLT and the ATM switch increases the average access delay but has practically no impact on the maximum experienced access delay. Figure 12.4 illustrates this phenomenon. The vertical bars in solid line correspond to the delay experienced by individual cells within five consecutive bursts of a rt-VBR flow traversing an APON system. At the input of the traffic shaper, the pattern is apparently highly distorted. After applying the shaper, the total delay (APON + shaper) is shown as a dashed-line envelop. As seen from the figure, the CDV within each burst is practically eliminated without any impact on the maximum access delay. The validity of this observation has been formally proven in [2].

3. SERVICE DISCIPLINES

In this section we discuss the use of service disciplines at both the OLT and the ONU. The service disciplines are applied either to aggregated bandwidth requests (at the OLT) or directly to ATM cells (at the ONU). We distinguish between FIFO and per-flow (PF) queueing. In the latter case the term “flow” designates either a single ATM connection (at the ONU) or an aggregated ONU flow (at the OLT). Numerous service disciplines have been proposed in the literature. In the following, we consider the WF^{2}Q+ [5] discipline which is known to be a very good approximation of the fluid Generalised Processor Sharing (GPS discipline [3].

Before applying a particular FIFO or per-flow algorithm, the flows belonging to different traffic categories are typically separated into two different service “planes” with different priority levels (see Figure 12.5). A separate service plane is defined for each MAC service (st-MTC,el-MTC) and a Static Priority (SP) scheduler is applied between them. Stream traffic has absolute priority. Elastic traffic is served only if there is no outstanding Stream bandwidth requests at the OLT. The SP sched. uler exists only at the OLT, provided that the permits are coloured i.e. provided that they indicate the MTC for which they have been issued. When the ONU receives a coloured permit, it merely executes a FIFO or per-flow queueing discipline within the indicated MTC plane in order to determine the flow to be served.

Table 12.1 Possible queueing combinations



Service plane	*Queueing combination (OLT-ONU)*
st-MTC	PF-PF, PF-FIFO, FIFO-FIFO
el-MTC	PF-PF, PF-FIFO

Several queueing combinations are possible within each service plane. The reasonable queueing combinations are listed in Table 12.1. For instance, per-flow queueing may be used at both multiplexing stages (OLT

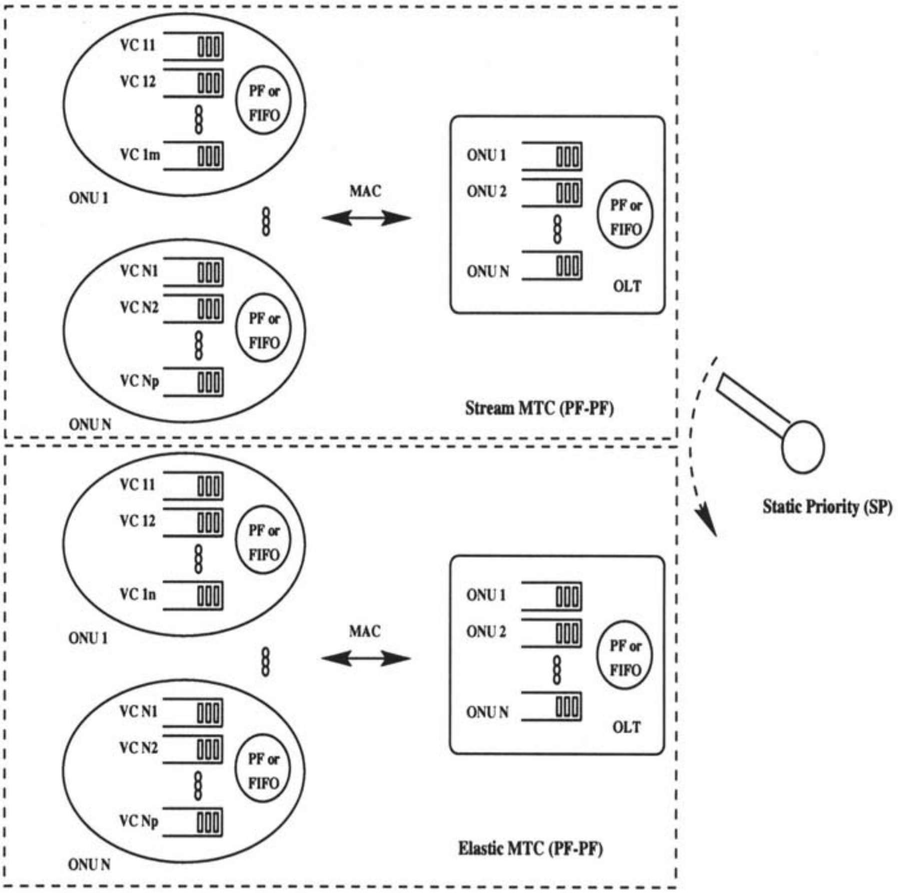


Figure 12.5 Two service planes: st-MTC and el-MTC. Several combinations of per-flow (PF) and FIFO queueing are applicable. A Static Priority (SP) is used between the two planes.

and ONU) in either service plane. This approach is referred to as PF-PF. A simplified version with per-flow queueing at the OLT and FIFO queueing at the ONU is also possible in both planes. In the el-MTC plane the FIFO queueing approach at the ONU must be complemented by active buffer management. Finally, it is possible to do FIFO queueing at both stages (FIFO-FIFO approach), but only in the st-MTC plane, given that the Elastic traffic is unbounded by its nature.

4. BUFFER MANAGEMENT

Roberts in [6] distinguishes two types of multiplexing in packet networks: rate envelop multiplexing (REM), and rate sharing. **Rate Envelop Multiplexing** is a multiplexing approach in which one tends to limit the probability for the aggregate arrival rate of all active connections going beyond a predefined envelop at any instant. This is done via admission control procedures at connection establishment. A new connection is simply rejected if some multiplexer along the path estimates that by allowing the connection to be established, the aggregated flow will sometimes increase beyond the available capacity. REM does not entirely prevent temporary rate overloads. However, if such overloads do occur, they have a small duration (this is also known as "cell scale congestion") and are absorbed in a small buffer (typically about one hundred of cells). That is why REM is also referred to as "bufferless multiplexing". REM is naturally applicable to traffic which has intrinsic rate characteristics, such as the Stream traffic. Under normal network conditions the multiplexer's buffer with REM remains lightly loaded at any instant and consequently, there is no need for buffer management. REM often results in poor network utilisation in case the submitted traffic is bursty. Furthermore, the concept of "rate envelop' is not applicable to elastic traffic. In order to increase the link utilisation when carrying bursty or elastic traffic, a different multiplexing approach is used. Roberts designates it as **rate sharing.** In rate sharing the traffic aggregate is allowed to increase beyond the system capacity from time to time. The temporary bursts (or "burst scale congestion") are being absorbed in buffers. This approach requires that the multiplexers be provisioned for substantial buffering (e.g. on the order of several thousands of cells). The rate sharing obviously increases the queueing delay and is therefore unable to provide delay guarantees. On the other hand, rate sharing is particularly adapted to providing *throughput* guarantees. A simple way to do this is to apply to the buffered cells a service discipline which is known to provide such guarantees (e.g. WF^2Q+).

Rate sharing with Elastic traffic needs further attention. It is well established (see [9]) that fair per-flow service disciplines are sufficient for providing throughput guarantees in situations with infinite buffer or in case of per-flow reservations of buffer space. In practice, per-flow buffer reservations are highly improbable due to scalability problems. When the buffer space is limited and the incoming traffic is unbounded, it may happen that the entire buffer space be monopolised by few greedy flows. So, even if fair queueing is used, some flows may not even be able to join the queue and wait there for fair service. Consequently, when handling

Elastic traffic in shared buffer space, the queueing discipline must be complemented by an active **buffer management scheme.** The latter argument is all the more justified when the traffic carried accross the Elastic flows is responsive to implicit feedback (e.g. TCP traffic).

In the following text we first recall the TCP congestion control mechanisms and then present several known buffer management schemes for both IP and ATM networks.

4.1 SPECIFICITIES OF TCP CONGESTION CONTROL

The TCP protocol provides congestion control at layer 4 in end-to-end manner. The Tahoe version of TCP uses the *congestion window* (*cwnd*) mechanism implemented at the sender's side. In this version, congestion control is organized in two phases, the *Slow Start* and the *Congestion Avoidance*. In case of packet loss, the sender sets its congestion control parameters (the congestion window and the *Slow Start threshold* (*ssthresh*) to such values that the obtained thoughput is strongly reduced.

The next version known as TCP Reno enables the sender to quickly recover from single segment losses. The congestion window is reduced to only half of its size, instead of being reduced to one segment as in TCP Tahoe. Unexpectedly, TCP Reno performs poorly in cases of multiple segment loss. Indeed, for every segment lost from a single TCP window, the TCP sender passes through a separate Fast Recovery phase. Since every Fast Recovery phase results in halving the *cwnd*, the net result of successive Fast Recoveries is an exponential *cwnd* decrease. This problem was referred to as *TCP Reno bug*, and several bug fixes have been proposed (see [8]). The bug-fixed version of TCP Reno is referred to as TCP NewReno. What TCP NewReno tries to achieve in case of multiple segment loss is to keep the TCP sender inside the Fast Recovery phase until the last of the series of lost segments is retransmitted and acknowledged. TCP SACK ([7]), the latest TCP version, uses the NewReno modification for congestion control and, in addition, has a selective acknowledgement scheme for error recovery.

4.2 BUFFER MANAGEMENT IN ATM NETWORKS

Fair buffer management schemes are those which try to allocate buffer space to the currently active connections according to some criterion. For services that have the notion of minimum bandwidth guarantee (GFR, ABR), each connection is typically allocated a weight which is proportional to its bandwidth guarantee. These weights are then used by the fair buffer management schemes for allocation of buffer space. We consider that the **Guaranteed Frame Rate** (GFR) service will typically be used for carrying TCP/IP traffic accross ATM networks. This is the only ATM service for which the conformance definition was specified in terms of frames rather than cells. There are two conformance definitions for the GFR service: GFR.l and GFR.2. The difference between the two is that GFR.2 allows the network to tag the excess traffic using the F-GCRA(T, f) algorithm, where $T = 1/MCR$ and $f \geq (MBS - 1) * (1/MCR - 1/PCR)$. The buffer management schemes typically discard tagged frames (CLP=1) with higher probability than untagged frames (CLP=0).

We consider here one well-known fair buffer management scheme: Weighted Fair Buffer Allocation (WFBA) defined in [4]. Figure 12.6 shows the generic algorithm for WFBA. There are three regions $R1$, $R2$, and $R3$, delimited by two global thresholds LBO and HBO, standing for Low and High Buffer Occupancy, respectively. A buffer allocation weight W_i is associated to each VC_i, proportional to its Minimum Cell Rate (MCR). All frames are accepted in region $R1$ ($X < LBO$), whereas classical Early Packet Discard (EPD) is performed in region $R3$ ($X > HBO$). In region $R2$ tagged frames (CLP=1) are systematically dropped[2]. The differences between the two algorithms appear with the treatment of untagged frames (CLP=0) in region $R2$. In region $R2$ an admission criterion is applied to untagged frames. This criterion is a function of the global and individual buffer occupancies, as well as the allocated weights. The frames which do not pass the admission criterion are dropped in a deterministic manner.

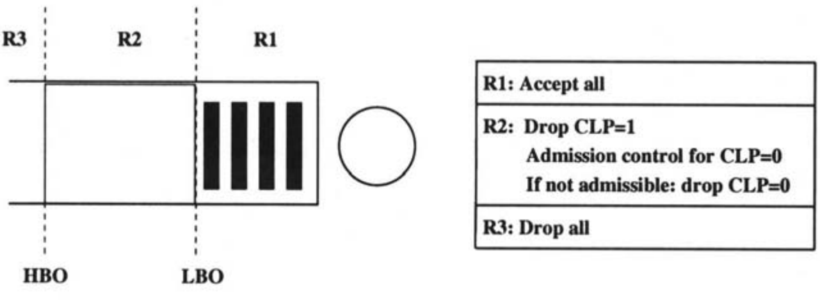


Figure 12.6 The WFBA algorithm.

The admission criterion for WFBA is given by:

$$X_i < \frac{W_i}{\sum_{j \in B(t)} W_j} \cdot \frac{HBO - LBO}{X - LBO} \cdot X \qquad (12.1)$$

where X_i is the buffer occupancy for VC_i and $B(t)$ is the set of backlogged connections at time t. We will primarily be interested for providing support for the GFR service over APON systems. The GFR traffic contract defines, among other things, the MBS and MFS parameters. Note that WFBA is unaware of these two parameters.

5. PERFORMANCE EVALUATION

In this section we evaluate the system performances of an APON access system via computer simulations. Three types of traffic scenarios are considered: scenario with Stream traffic only (Section 5.1), scenario with Elastic traffic only (Section 5.2) and scenario with both types of traffic (Section 5.3).

5.1 STREAM TRAFFIC ONLY

The following traffic flows are considered:

- constant bitrate flow (CBR flow) defined by its Peak Cell Rate (PCR);
- variable bitrate flow (rt-VBR flow) defined by its PCR, Sustainable Cell Rate (SCR) and Maximum Burst Size (MBS).

The numerical values of the above mentioned parameters are: PCR = 5400 cell/s, SCR = PCR / 5 = 1080 cell/s, and MBS = 100 cells. Our rt-VBR flows are modelled as worst-case ON-OFF traffic. We mean by worst-case ON-OFF model that during each burst inter-arrival, the observed rate is equal to SCR. Thus, the burst size (i.e. the number of cells in an ON period) is a random variable whose probability density function (pdf) is defined on the [1, MBS] interval. Once the random burst of size X is generated, the duration of the ON period is determined as $T_{ON} = (X - 1) \bullet \frac{1}{PCR}$, whereas the subsequent OFF period is determined as: $T_{OFF} = X \bullet \frac{1}{SCR} - T_{ON}$. An arbitrary pdf is used for the burst size distribution. Such a traffic model is compliant with the $GCRA(\frac{1}{SCR}, IBT)$ algorithm.

With per-flow queueing (e.g. WF^{2}Q+) at either the OLT or the ONU, the allocated bandwidth to a connection is equal to its PCR and Equivalent Bandwidth (EqBW) for CBR and rt-VBR flows, respectively. We use the following formula for EqBW computation, taken from [6]:

$$EqBW = \begin{cases} a \cdot SCR \cdot (1 + 3 \cdot z \cdot (1 - \frac{SCR}{PCR}) \\ \text{if } 3 \cdot z \leq min(3, \frac{PCR}{SCR}) \end{cases} \tag{12.2}$$

$$EqBW = \begin{cases} a \cdot SCR \cdot (1 + 3 \cdot z^2 \cdot (1 - \frac{SCR}{PCR})) \\ \text{if } (z > 1) \text{ and } (3 \cdot z^2 \leq \frac{PCR}{SCR}) \end{cases} \tag{12.3}$$

$$EqBW = \begin{cases} a \cdot PCR \\ \text{otherwise.} \end{cases} \tag{12.4}$$

where:

$$a = 1 - \frac{log_{10}(P_{loss})}{50} \tag{12.5}$$

$$z = -2 \cdot \frac{PCR}{C} \cdot log_{10}(P_{loss}). \tag{12.6}$$

In the above formula, the "loss" probability, P_{loss}, is actually the probability for the rt-VBR aggregate going beyond its allocated rate envelop. We set P_{loss} equal to 10^{-5}. Given the above formula we calculate that for a rt-VBR flow with PCR = 5400 cell/s and SCR = 1080 cell/s (note that the MBS parameter is not relevant for this formula) the equivalent bandwidth is equal to 2012 cell/s.

When describing the system load in scenarios containing rt-VBR traffic, we make the disctinction between *load* and *actual load*, the former meaning "loaded with equivalent bandwidth" and the latter meaning "loaded with SCR". For instance, in a rt-VBR traffic scenario with global load of 0.90, the actual load is 0.48 (= 0.90• $\frac{SCR}{EqBW}$).

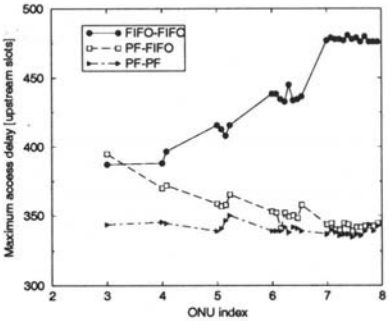
We next describe a traffic scenario consisting of 28 CBR and 75 rt-VBR connections which is used for our computer simulations. We refer to it as CBR28-VBR75 scenario for obvious reasons. The CBR connections are distributed accross 5 consecutive ONUs (ONU_3 to ONU_7), whereas the rt-VBR flows are distributed over 8 consecutive ONUs (ONU_0 to ONU_7). The number of CBR and rt-VBR flows at each ONU is shown in Table 12.2. The global CBR load is 45% of the system capacity. The global rt-VBR load also equals 45% of the system capacity, but the *actual* rt-VBR load equals only 24%.

Figures 12.7 and 12.8 compare the maximum recorded delays for CBR and rt-VBR flows, respectively, under the FIFO-FIFO, PF-FIFO and

Table 12.2 CBR28-VBR75: number of CBR and rt-VBR flows per ONU.

ONU index	0	1	2	3	4	5	6	7
CBR	0	0	0	1	2	4	8	13
rt-VBR	9	9	9	9	9	10	10	10

PF-PF schemes. The former two queueing frameworks (FIFO-FIFO, PF-FIFO) result in the same delay bounds for CBR and rt-VBR flows stemming from the same ONU. This is logical since both CBR and rt-VBR flows share the same FIFO queue at the ONU. It is interesting to note that connections traversing higher-indexed ONUs experience higher delay bounds. By means of repeated simulations we have concluded that this bias towards higher-indexed ONUs is due to the minislot position in the Divided Slot. The ONU whose minislot is on the last position in the Divided Slot yields the highest delays, (squares in Figure 12.7 and 12.8).



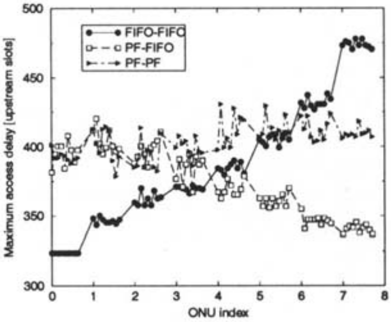


Figure 12.7 CBR28-VBR75: maximum access delays for CBR connections.

Figure 12.8 CBR28-VBR75: maximum access delays for rt-VBR connections.

With the PF-FIFO scheme, the maximum access delay is inversely proportional to the *aggregated* ONU flow. For instance, the ONU_7 is the ONU with largest bandwidth reservation and, therefore, the connections stemming from it experience the lowest delays.

Finally, under the PF-PF scheme, the maximum experienced delay depends mainly on the individual per-VC reservations[3]. All CBR connections experience roughly the same delay bounds. The same is true for all rt-VBR connections. However, the latter systematically experience higher delays than the CBR connections. This is logical since each rt-VBR connection is allocated 2012 cell/s (equivalent bandwidth), whereas each CBR connection is allocated 5400 cell/s (PCR).

5.2 ELASTIC TRAFFIC ONLY

In this section we consider a heterogeneous scenario referred to as GFR32. We consider 32 GFR VCs carrying 10 TCP-NewReno connections each. The total number of TCP connections equals 320. Only 8 ONUs (out of 64) are active and each one is traversed by four GFR VCs. The VCs do not have equal MCR reservations, neither RoundTrip Times (RTT). The parameters which describe the scenario are given in Table 12.3. Table 12.3 contains the TCP-specific parameters, such as: version, Maximum Segment Size or timer granularity. Note that we use a TCP window size of 256 koctets which is larger than the default value of 64 koctets in order to avoid throughput limitation by the TCP flow control. For similar reasons the TCP timer granularity is also smaller than the one which is found in today's implementations (typically 100 ms or 500 ms). Table 12.3 also shows the ONU buffer threshold settings (LBO and HBO) which are used for buffer management, as well as the way the MCR and RTT are distributed accross the 32 VCs and eight active ONUs. The per-VC parameters (MCR and RTT) for scenario GFR32 are also illustrated in Figures 12.9 and 12.10. In these figures, as well as in all subsequent figures, the abscissa values identify the active ONUs.

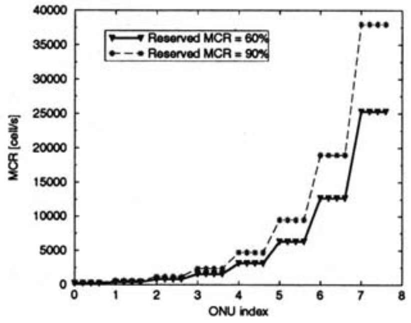


Figure 12.9 (a) GFR reservation MCR_i.

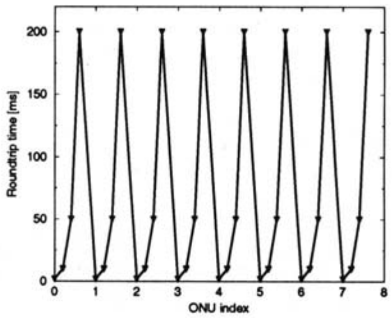


Figure 12.10 (b) TCP roundtrip time

We use the NewReno version of TCP which is described in [8]. This is a bug-fixed version of TCP Reno which improves the performance

Table 12.3 Parameters for scenario GFR32

Active ONUs	8 (out of 64)
GFR connections per ONU	4
TCP connections per VC	10
Total MCR reservation	either 60% or 90% of the system's capacity
TCP window size	256 koctets
TCP Maximum Segment Size (MSS)	1460 octets
Maximum Frame Size (MFS)	32 cells
ONU buffer size	2000 cells
LBO (HBO) threshold	900 (1800) cells
TCP version	NewReno
TCP timer granularity	10 ms
MCR for VCs at the same ONU	relate to each other as 1:1:1:1
Aggregated MCR of the eight active ONUs	relate to each other as 1:2:4:8:16:32:64:128
Roundtrip time	2:10:50:200 ms for every four consecutive VCs; the same pattern is repeated at each ONU

of the latter in case of multiple segments dropped from a window of data. All 320 TCP sources are persistent i.e. they have always data for transmission.

Figure 12.11 illustrates the traffic management functions inside a particular ONU. We assume that each TCP source (40 TCP sources per ONU) is connected to the ONU via a separate physical link with length of 2 km and 51.84 Mbit/s capacity. Of course, this is not very realistic since the ONU will hardly be equipped with 40 physical interfaces. This choice was done to avoid any possibility of congested access links interfering with the APON traffic management. We assume that the transfer between the TCP sources and the ONU takes place in packet mode, the segmentation being done at the ONU's entry. After the segmentation into ATM cells, the TCP connections are multiplexed into GFR VCs, ten TCP connections per VC. The ATM cells are subject to traffic management functions: tagging (F-GCRA), buffer management (WFBA) and queueing (FIFO or per-flow queueing).

By combining the WFBA buffer management scheme which was described in Section 4. with either FIFO or per-flow (PF) queueing, we obtain the following schemes: WFBA+FIFO and WFBA+PF. We investigate their performances by means of simulations.

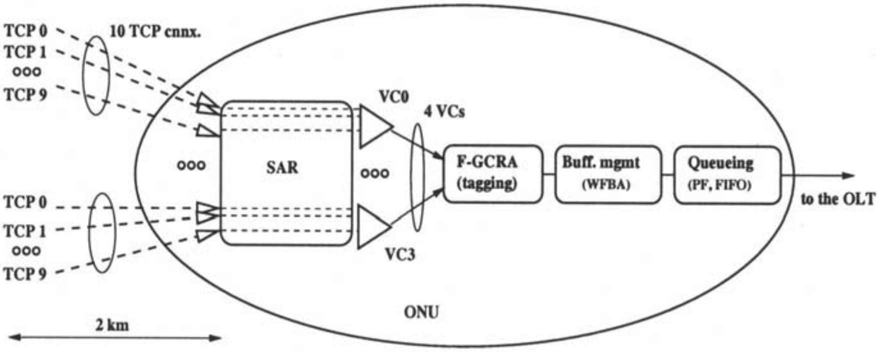


Figure 12.11 Traffic management functions at the ONU.

The simulations reported in this section correspond to 30 seconds of simulated time. The results are expressed via the normalised goodput received by each VC, defined as: $R = \frac{Goodput_i}{MCR_i}$. The value $R = 1$ means that VC_i has realised goodput which is exactly equal to its reservation MCR_i, without receiving any excess bandwidth. Similarly, $R = 2$ means that VC_i has realised goodput which is equal to twice its reservation and $R = 0$ means that VC_i has not realised any goodput at all.

Ideally, the R ratio should be equal to $1.66 = \frac{100}{60}$ and $0.11 = \frac{100}{90}$ for a global MCR reservation of 60% and 90%, respectively. Note, however, that this ratio can never be achieved because some bandwidth is necessarily wasted on TCP retransmissions. We find out that slightly more than 1% of the total carried traffic is wasted on retransmissions, which is a remarkable result. This wasted bandwidth is roughly invariant accross all simulations.

Figure 12.12 illustrate the normalised goodput for scenario GFR32, for schemes relying on FIFO queueing. The global GFR reservation (i.e. the sum of per-connection MCR_i) equals either 60% or 90% of the system capacity. Also shown in the figures are three horizontal lines. The first one, $y = 1.0$, is the lowest value for the normalised goodput at which the bandwidth guarantee is still met. The other two lines ($y = 1.66$ and $y = 1.11$) correspond to the ideal value for the normalised goodput, for which every VC gets a share of the available (non-reserved) bandwidth in proportion to its MCR.

Figure 12.12 shows that even at 60% reservation in scenario GFR32 there are several connections that do not meet the guarantee. This is explained by the fact that the last two ONUs contain VCs whose bandwidth-delay product is greater than the ONU buffer size. Moreover, the non-reserved bandwidth is distributed unfairly since lower-rate

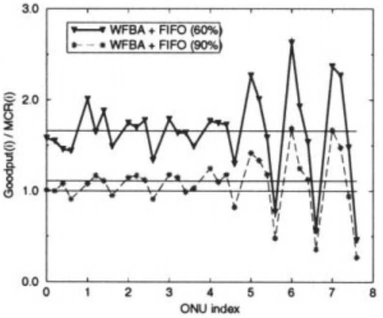


Figure 12.12 (a) GFR32: TCP goodput with FIFO queueing.

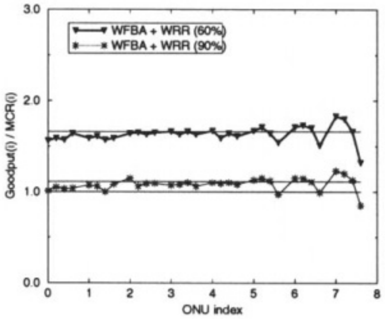


Figure 12.13 (b) GFR32: TCP goodput with PF queueing.

connections realise higher normalised goodput than higher-rate connections.

Figure 12.13 shows the normalised goodput when PF queueing is used at the ONU. Almost all VCs attain their guarantee or succeed to make it within 15% of the MCR guarantee, even for the extreme case of VC_{31} at 90% load. Moreover, the free bandwidth is distributed fairly (the normalised goodput curves are almost flat).

5.3 MIXED TRAFFIC SCENARIO

In this section we consider a scenario consisting of both Stream and Elastic traffic. The Elastic traffic is represented by a down-scaled version of the GFR32 scenario, so that the global GFR MCR reservation equals 45% of the system capacity. The Stream traffic is represented by 140 rt-VBR flows with the following parameters: PCR = 5400 cell/s, SCR = 1080 cell/s and MBS = 100 cells. The number of rt-VBR flows per ONU is given in Table 12.4. The global *actual* rt-VBR load equals 45%, as well.

Table 12.4 Number of rt-VBR flows per ONU for the VBR140 subscenario.

Flow type	ONU_0	ONU_1	ONU_2	ONU_3	ONU_4	ONU_5	ONU_6	ONU_7
rt-VBR	17	17	17	17	18	18	18	18

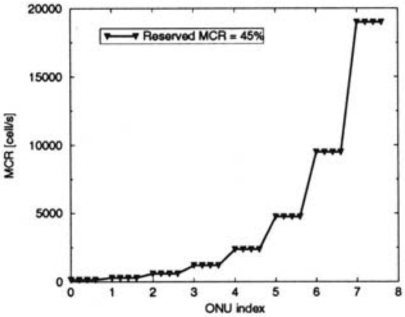


Figure 12.14 GFR32: GFR reservations for 45% load

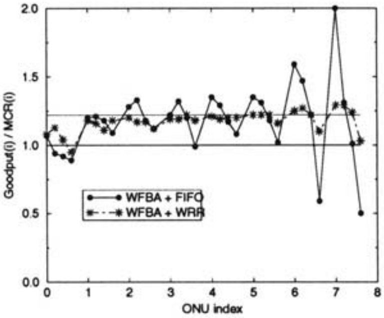


Figure 12.15 Comparison of FIFO and PF queueing at the ONU in terms of normalised TCP goodput.

Given the strict inter-MTC priority, the presence of Elastic traffic has no negative impact on the Stream traffic. Therefore, our interest in this subsection is focused only on the former. We consider the PF-PF or PF-FIFO queueing frameworks in the el-MTC plane, whereas any queueing framework may be used in the st-MTC plane. Buffer space of 2000 cells for Elastic traffic and 200 cells for Stream traffic is reserved at each ONU, and the Weighted Fair Buffer Allocation scheme (WFBA) is used for fair buffer management.

Figure 12.15 shows the normalised TCP goodput for GFR VCs under both PF and FIFO queueing at the ONU. As seen from the figure, FIFO queueing combined with WFBA buffer management fails to provide guarantees for two high-rate flows. Note that this is nothing new since the same observation was made earlier (cf. Figure 12.12). On the contrary, with PF queueing at the ONU it is possible to provide MCR guarantees to individual flows.

The fact that guarantees can be provided under the PF-PF scheme is a remarkable result since in this scenario we have an example of efficient overallocation. Indeed, in this mixed scenario the bandwidth reserved for rt-VBR traffic amounts for 83.76% ($83.76 = 45 \bullet \frac{EqBW}{SCR}$) of the total system capacity. This value is the one perceived by the rt-VBR CAC algorithm at the OLT, although the actual rt-VBR load is 45%. The fact that the unused part of the bandwidth allocated to the rt-VBR traffic could be shared as free bandwidth by the Elastic flows is obvious. What is not so obvious here is that the unused rt-VBR capacity can be "re-sold" as *guaranteed* capacity to the GFR flows. Hence, although the overall amount of allocated capacity (83.76% to Stream traffic plus 45% to Elastic traffic) *exceeds* the system capacity, QoS guarantees are provided to both traffic categories.

6. CONCLUSIONS

In the first part of this article we have described the characteristics of APON access systems of which the physical layer has been standardized (G.983 recommendation). Many aspects of the APON MAC protocol remain open to discussion. In this article, we have presented and compared possible approaches for ATM and TCP-IP traffic management on such access networks. Our proposals are based on the use of a standalone traffic shaper between the OLT and the ATM switch. We consider two MAC Transfer Capabilities for elastic traffic and for stream traffic respectively, a static priority being given to this latter. Various possible combinations between different service disciplines both at the OLT and at the ONUs have been presented. We have then briefly recalled the specificities of TCP congestion control before considering the Guaranteed Frame Rate service for carrying TCP/IP traffic accross an APON. Three types of scenarios have been investigated through computer simulations. The first scenario considering stream traffic only (CBR, rt-VBR) shows that applying a FIFO discipline at the ONUs and a FIFO queueing or a per-flow queueing at the OLT give roughly the same delay bounds between CBR and rt-VBR connections. In the case of per-flow queueing both at the ONUs and at the OLT, CBR connections get the same delay bound, just like rt-VBR connections. The second scenario was concerning elastic traffic only, i.e TCP NewReno connections over GFR virtual connections. The ATM cells issued by segmentation of TCP segments in an ONU are first treated by a Frame-GCRA controller, then submitted to a fair buffer management scheme (WFBA) before being inserted either in a common FIFO queue or in per-flow queues. Our simulation results show that only per-flow queueing at the ONUs is able to guarantee the required goodput to GFR connections, even at high loads. In addition, free bandwidth is fairly shared among these connections. Our last scenario was concerning mixed stream and elastic traffic. Again, per-flow queueing at the ONUs looks much more efficient than FIFO queueing for guaranteeing the reserved goodput to GFR connections.

Acknowledgments

The authors wish to thank Rudy Hoebeke from Alcatel Corporate Research Centre for his encouragement and support.

Notes

1. In fact, the ITU-T G.983 recommendation specifies a maximum distance span of 20km. However, in practice the distance span of an APON system with 64 ONUs will typically be limited to 10km.

2. In its original version [4] WFBA does not distinguish between tagged and untagged frames. However, this extension is straightforward: tagged frames in both schemes are admitted into the buffer only if the global buffer occupancy X is lower than LBO.

3. In fact, it can be shown that the delay bound in a hierarchical server depends both on the individual bandwidth reservation and on the higher-layer aggregated bandwidth reservation. Yet, the influence of the latter is smaller and becomes apparent only under high discrepancies in the aggregated rates.

References

[1] J. Angelopoulos and E. Fragoulopoulos and I. Van de Voorde and I. Venieris and P. Vetter, "Efficient Control of ATM Traffic Accessing Broadband Core Networks via SuperPONs", SPIE Journal, Vol.2357, 34–43, 1996.

[2] L. Georgiadis and R. Guérin and K. N. SivarajanLacey, "Efficient Network QoS Provisioning Based on per Node Traffic Shaping", IEEE/ACM Transactions on Networking, Vol.4, No.4, 1996.

[3] Abhay K. Parekh and Robert G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case", IEEE/ACM Transactions on Networking, Vol.1, No.3, 1993.

[4] Juha Heinanen and Kalevi Kilkki, "A Fair buffer allocation scheme", unpublished manuscript, 1995.

[5] J. C. R. Bennet and H. Zhang, "Hierarchical Packet Fair Queueing Algorithms", IEEE/ACM Transactions on Networking, Vol.5, No.5, 1997.

[6] J. Roberts and U. Mocci and J. Virtamo, "Broadband Network Teletraffic", Final Report of Action, COST 242, Springer Verlag, 1996.

[7] M. Mathis and J. Mahdavi and S. Floyd and A. Romanow, "TCP Selective Acknowledgement Options", IETF RFC 2018, April 2018.

[8] Tom Henderson and Sally Floyd, "The NewReno Modification to TCP's Fast Recovery Algorithm", IETF draft-ietf-tcpimpl-newreno, November 1998.

[9] B. Suter and T. V. Lakshman and D. Stiliadis and A. Choudhury, "Efficient Active Queue Management for Internet Routers", Proceedings of the Networld+Interop Engineers Conference, May 1998.

Maurice Gagnaire is an Associate Professor at the Ecole Nationale Supérieure des Télécommunications (ENST) in Paris-France. He is graduated from the Institut National des Télécommunications in Evry-France. He received the Diplôme d'Etudes Approfondies from the University of Paris-6, the PhD degree from the ENST (1992) and the Habilitation from the University of Versailles-France (1999). He is in the program committee of various IEEE and IFIP international conferences. His research activities are focused on the design and performance evaluation of medium access control protocols (all-optical IP backbones, Fiber-In-The-Loop and Wireless-In-The-Loop access networks). He is co-author of a book on High Speed Networks and author of a book on new broadband access networks.

Sašo Stojanovski Saso Stojanovski, received his B.S. and M.S. degrees in telecommunication engineering at the Faculty for Electrical Engineering in Skopje, Macedonia in 1989 and 1995, respectively. He has been working on telecommunications software development for Nikola Tesla, Zagreb, Croatia and AT&T Barphone, Saumur, France. From 1990 to 1996 he has been with the Telecommunications Department at the Faculty for Electrical Engineering in Skopje, where he worked as a teaching assistant. In December 1996 he was enrolled in a Ph.D. programme at the Ecole Nationale Supérieure des Télécommunications in Paris, France. His current research interests include network architecture and traffic management in ATM and TCP/IP networks.

Chapter 13

WIRELESS ATM: AN INTRODUCTION AND PERFORMANCE ISSUES

Renato Lo Cigno
Dip. di Elettronica - Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
locigno@polito.it

Abstract

This paper presents an overview of the main characteristics of Wireless ATM (W-ATM) Networks, underlining the main differences between W-ATM and other integrated wireless networks. The main architectures for W-ATM that recently appeared in literature are discussed, and pros and cons of possible solutions are investigated. The material covered by the paper is subdivided between topics related to the radio access and topics related to the network management and architecture.

Performance issues relevant to the different discussed topics are identified, and the main areas where research and investigation is needed in order to develop high performance commercial networks are discussed.

MAC protocols for micro- and picocellular networks are discussed in somewhat higher detail, as well as handover procedures suitable for implementation in ATM.

Keywords: Wireless-ATM, Cellular networks, Third Generation Mobile Networks

1 BACKGROUND

In recent years the research and development in telecommunication networks has followed mainly two distinct and separated trends: the provision of broadband integrated services in wired networks and the provision of enhanced mobility services in wireless networks. On the one hand the efforts of the technical and scientific community has been focussed on multimedia services

* This work was supported by Telecom Italia Research Center (CSELT), by the Italian National Research Council (CNR) and by the Italian Ministry for University and Scientific Research (MURST).

in wired networks, offering to the end user a significant amount of transmission capacity at low cost, with high, guaranteed and application oriented Quality of Service (QoS) and easy access to the network resources; ATM networks have been the preferred playground for research on these topics, offering a standard seamless platform for the provision of integrated multimedia services. On the other hand research in wireless networks has been focusing mainly on problems relating to the mobility of terminals, without bothering about wideband transmission or application oriented QoS.

Indeed, the transmission conditions in wired networks are good enough to allow technicians to concentrate on the development of new appealing services and on the efficient exploitation of the available capacity. The transmission conditions on the channel between a mobile terminal and its base station are instead very poor. The channel characteristics change very rapidly in time due to shadowing phenomena (big objects such as buildings or trucks blocking the radio path) and multipath fading (the disruption of the radio signal due to negative interference of reflected and refracted radio waves). Researchers were thus forced to concentrate upon basic issues such as the provision of a reliable transmission channel, disregarding more sophisticated issues.

The Broadband Integrated Services Digital Network (B-ISDN) is evolving from prototypes to commercial deployments. At the same time the second generation of mobile communications systems (e.g., GSM in Europe; IS-54 and IS-95 in the U.S.) is expanding very fast on the market, and the third generation will come very soon. As a consequence the research community interests start to concentrate on the possibility of integrating mobile terminals directly within the B-ISDN, providing the mobile users with the full spectrum of multimedia services typical of the B-ISDN.

Several alternatives can be considered for the provision of integrated multimedia services to mobile users, starting from the Universal Mobile Telecommunication System (UMTS), that aim at the integration of heterogeneous telecommunication networks within a sophisticated and homogeneous interworking framework, to the extension of Internet to mobile applications (Mobile IP). If the target is the integration of mobiles within B-ISDN, however, the natural choice is the adoption of the Asynchronous Transfer Mode (ATM), in the same way as in wired B-ISDN, in wireless networks too. This approach is usually termed Wireless ATM (W-ATM). Evidence of this trend can be found in several directions.

First of all, the number of international research programs that are being carried out (or have just finished) on this topic is very high.

Second, the interest shown on this topic both by governmental standardization bodies and by industrial interest groups. The ATM Forum has recently established a Working Group on Wireless ATM, whose main focus is on the requirements for W-ATM as well as on system and architectural aspects of the problem. Official standardization bodies like ITU-T, ANSI and ETSI are also paying attention to the evolution of wireless networks, with special attention paid to the radio or ‘air’ interface and to the possible integration with B-ISDN: once again W-ATM.

Last but not least, the attention given to the subject by scientific publishers that recently devoted special issues to wireless ATM [1, 2, 3, 4, 5].

Wireless ATM is a new topic and its exact definition is still not completely agreed upon. Section 2 summarizes the general architecture and the key points that define W-ATM.

The design of wireless ATM networks raises a number of challenges, that can be grouped in two broad categories. The first one comprises all problems related to the radio access. Research topics in this category range from modulation and coding for the provision of high user data rates (from 2 Mbit/s up) required by multimedia services over the radio interface, to the Medium Access Control (MAC) protocols that must be used where the radio channel is not rigidly subdivided between connections, but is accessed asynchronously only when data is to be transmitted. Section 3 is dedicated to the discussion of such topics, with particular attention to MAC protocol issues.

The second category comprises those problems related to the management of the ATM network when mobiles are allowed to roam freely through the network. In particular the integration of mobility within B-ISDN implies the dynamic re-establishment of the ATM Virtual Circuits (VCs) within the short time span of the mobile terminal handover from one cell to another. The problem of providing handover procedures integrated within the ATM network will be addressed in Section 4; many other topics, such as mobile location, tariffing and the like, fall within this category, but will not be addressed in this paper.

2 KEY FEATURES OF WIRELESS ATM

Before proceeding further in the analysis of problems and possible solutions for W-ATM, it is useful to spend a few words on identifying the key features of W-ATM and to focus on what makes it different from other proposals for integrated mobile networks.

A W-ATM network comprises Mobile Terminals (MTs), Base Stations (BSs), ATM switches and concentrators (ATM nodes), and Fixed Terminals (FT). Fig. 13.1 gives a simplified representation of a generic W-ATM network. The elementary characteristics of the entities of the network are as follows:

Mobile Terminals are the end points of connections whose peculiar characteristic is the access to the network through a radio link that enables them to roam through the network. The point where the connection from a MT enters the network is by definition always a Base Station. Mobile

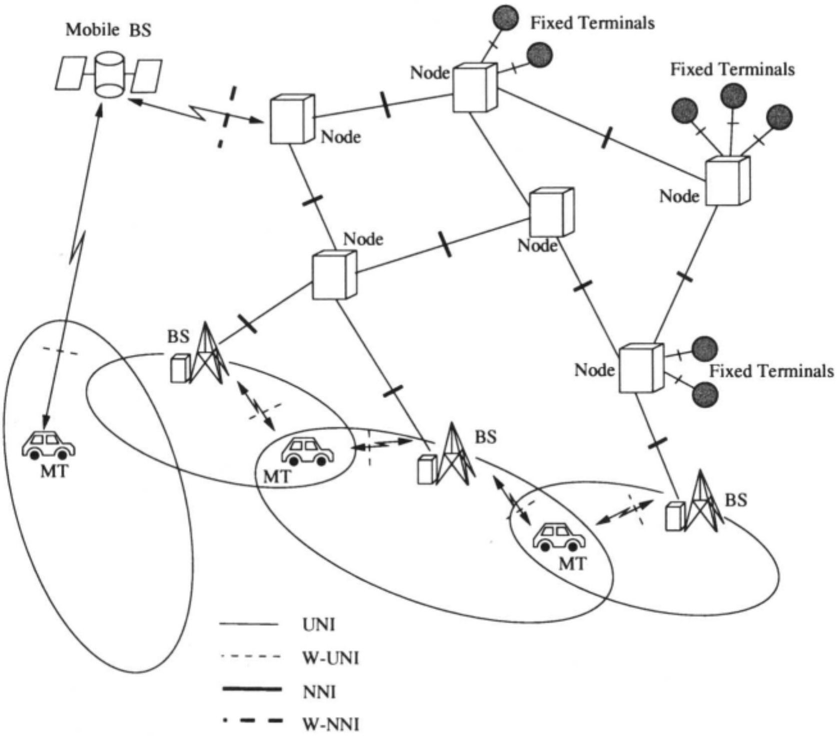


Figure 13.1 W-ATM scenario

Terminals can have multiple connections with different remote hosts, as for any B-ISDN terminal.

Base Stations are the interface between the wired and the wireless parts of the network. A BS can have switching capabilities or not, depending on the implementation and the network architecture. It is an ATM concentrator collecting many different connections from MTs and forwarding them on to the network. A BS can also be mobile itself (e.g., its location being on ships, aircrafts or satellites). A BS always has a connection towards the fixed part of the network, however, if the BS is mobile, this connection changes over time and must be through a radio channel.

ATM nodes are the basic infrastructure of the B-ISDN core. Beside all traditional ATM capabilities, in W-ATM they must provide support for the mobility services. Depending on architectural choices all of the ATM nodes, or just some of them, must be "mobility aware". There are two possibilities for the provision of mobility support in ATM nodes. In the first one, the mobility support functions are directly embedded within

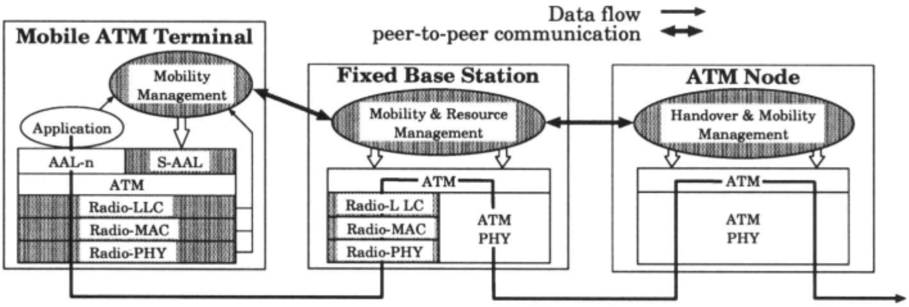


Figure 13.2 Protocol stack of mobile terminals, base stations and ATM nodes in W-ATM networks with fixed base stations

the node. In the second one, the mobility support functions are located in a *mobility server*, i.e., a special purpose entity that can interact with the control and management plane of the ATM node.

The network layout illustrated in Fig. 13.1 can indeed be appropriate for any integrated network comprising mobile terminals and does not give any insight on what distinguishes W-ATM from other integrated networks, like for instance the UMTS [6, 7]. The key difference between different proposals for integrated networks lies in the*protocol architecture* of the network. In W-ATM the information flows enter the ATM layer at the user premises (the User to Network Interface or UNI, W-UNI if over a radio link) and exit from it at the UNI at the destination, so that from the user point of view the network is completely homogeneous and there is no difference between mobile and fixed users. In other architectures, the B-ISDN connection is terminated somewhere within the network (typically at the base station), so that the network is not completely homogeneous and differences exists between fixed and mobile users.

The protocol stack of W-ATM entities is illustrated in Figs. 13.2 and 13.3. The difference between the two figures lies in the fact that Fig. 13.2 assumes a fixed base station, while Fig. 13.3 assumes a mobile base station. The base stations may also have switching capabilities, but, since the switching function is embedded within the ATM layer, the impact on the protocol stack is null. In both figures there are shaded entities or layers: these are the areas where research is still needed for the provision of W-ATM or where standards are not stable. Notice that in Fig. 13.3 the shading of the radio protocol stack between the BS and the MT is different from the one of the stack between the BS and the MT, this is to indicate that the radio interface has different characteristics and needs. For instance the mobile base station can be on a low orbit satellite: the access point to the fixed, terrestrial network, will be over a radio channel and change over time. However, the mobility pattern of the satellite is deterministic,

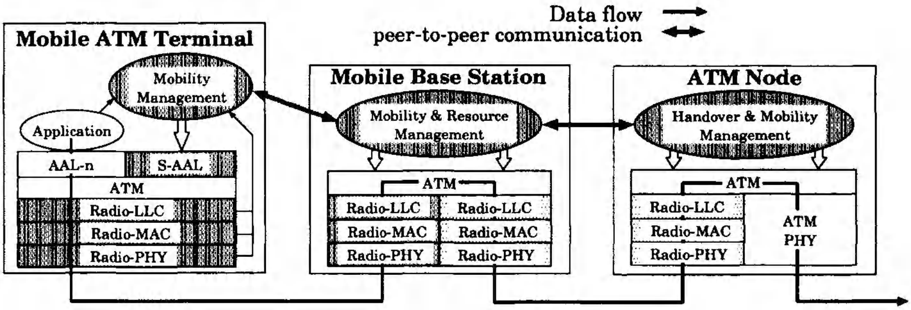


Figure 13.3 Protocol stack of mobile terminals, base stations and ATM nodes in W-ATM networks with mobile base stations

so that the procedures for changing the access points are completely different from those needed for the handover of MTs from one BS to another.

In both figures the thick solid line starting from the MT application represents the logical path followed by the information flow between the application and the remote host to which the MT is connected. It must be pointed out that all the components of the network, the MTs, the BSs and the ATM nodes[1] must have *mobility aware* entities that operate directly at the ATM level and that are devoted to the management of handover procedures when the terminal moves from one base station to another. The peer-to-peer communication between these entities in W-ATM contains control information *relevant to the ATM level* and the protocol entity at the ATM level must have primitives capable of making a suitable use of this control information. There are two possibilities for the implementation of mobility management procedures. The first one makes use of standard ATM signaling, as proposed by the ATM Forum [8]. The second one uses dedicated signaling like, for example, in-band signaling implemented through dedicated Resource Management cells, as proposed in [9]. The first solution requires the modification of the Signaling ATM Adaptation Layer S-AAL, together with the modification of ATM signaling standards like Q.2931 recommendations of the ITU-T or the UNI/NNI specification of the ATM Forum, the other solution require less modifications to the standards, but may result in a less efficient exploitation of network resources if it is not properly implemented and integrated within the network.

Fig. 13.4 reports a possible architecture of a non W-ATM integrated network. Although this is clearly not the only possibility, it still helps in pointing out some of the key features that distinguish W-ATM from other proposals.

[1] In general terms it is not necessary that all of the nodes of the fixed part of the ATM network are *mobility aware*, but it is possible that only some of them are capable of handling mobility related signaling and protocols, the other nodes are just by-passed by this information.

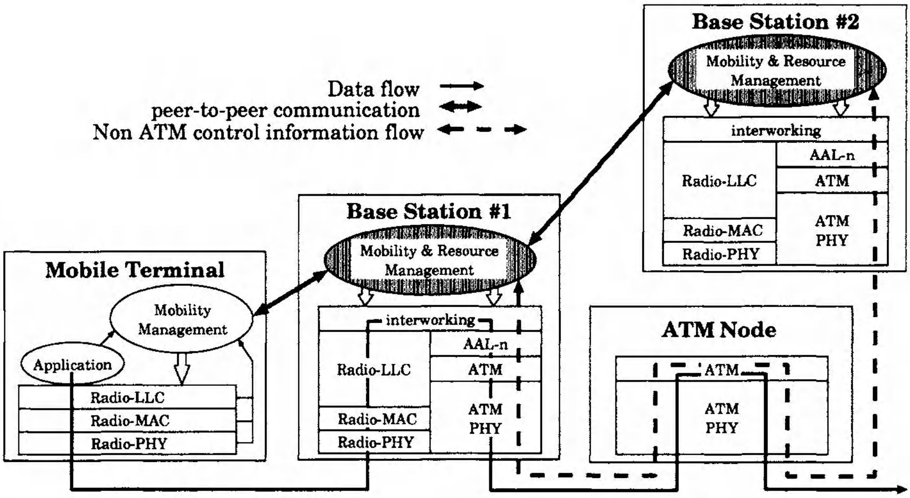


Figure 13.4 Protocol stack of mobile terminals, base stations and ATM nodes in a non W-ATM network

From the point of view of the end user the key difference is that the ATM connection is terminated at the base station and the user protocol stack is not the standard B-ISDN protocol stack, but a different one. Although this may look like a minor difference, it has quite a big impact on the user terminal. For instance, a portable computer must have the possibility of connecting both to the wired and to the wireless part of the network, and a unified protocol architecture would be a great advantage.

From the network point of view the differences are even more important. First of all the base stations must have an interworking unit linking ATM to the wireless part of the network. In addition, the ATM network need not be mobility aware (this can be an advantage, since it does not require modifications to the ATM standards). Mobility issues are completely managed by base stations or dedicated servers *above* the ATM layer. The mobility control information (represented by the thick dashed line in the figure) is thus not relevant to the ATM layer and flows through a "back-door" channel between base stations (the back-door channel can, for instance, be a semi-permanent VC). Although this may seem a simplification of the mobility management procedures, it introduces an additional problem: as roaming through the wireless network the MT needs to change its access point to the fixed network so that the connection must be re-routed. Since there are no suitable primitives for performing this task at the ATM level, it follows that for each handover the whole end-to-end connection must be torn down and built up again, leading to unacceptable delays and signaling overheads. This problem is exasperated when BSs connected to different ATM nodes are involved. Alternatively the ATM network must be involved in the handover procedure so that its architecture must be enhanced and mobility becomes embedded in the ATM network.

3 THE RADIO ACCESS

With reference to the protocol architecture of a W-ATM network sketched in the previous Section, the radio access identifies the part of the network that lies *below* the ATM layer. This Section addresses problems related to the radio channel between the MT and the BS. If the BS is mobile itself, then a radio access exists also between the BS and the ATM node; however the characteristics of this radio access are completely different from those of the channel between the MT and the BS, and are presently receiving little attention from the technical and scientific community. Problems related to the radio channel between a BS and an ATM node will not be considered further in this Section, even if Satellite ATM Networks surely offer very interesting research areas and potential commercial utilization.

3.1 MODULATION AND CODING

The concept of ATM, together with the “core and edge” architecture, is based on highly reliable transmission means, like optical fibers. Due to shadowing and multipath fading a radio channel for a Mobile Terminal is not reliable. Hence it is, in principle, incompatible with ATM. Modulation and coding techniques have the task of bringing down the cell loss rate and the cell delay variation on the radio channel, if not to values comparable with standard ATM links, at least to levels compatible with minimum ATM requirements. Admissible cell loss rates in ATM links are set below 10^{-9}, a value difficult to reach on a mobile radio channel. The protocol architecture depicted in Figs. 13.2 and 13.3, however, assumes that below the ATM layer on a radio link, a complete OSI Level 2 protocol is inserted, allowing also for the operations of automatic retransmission request (ARQ) protocols. Since the transmission delay between the mobile terminal and the base station is fairly small, ARQ protocols can be used while still ensuring low transfer delays. The target cell loss rate can be met even if the cell loss rate on the radio link is roughly $10^{-3} - -10^{-5}$, values that can be reached without too much effort with modulation and coding techniques.

Traditional techniques for narrowband, TDMA radio access for mobile networks (see for instance [10], for a thorough coverage of the argument) make use of long interleaving, codes with memory, like convolutional codes or trellis coded modulations, and frequency or spatial diversity. These techniques, however, can not be extended directly to W-ATM without some drawbacks. Interleavers introduce additional delays and extend over several ATM cells, whose transmission is thus no longer independent of one another. The same considerations apply to codes with long memory. The use of diversity may also become a problem as the transmission speed grows.

Also the use of spread spectrum techniques, like CDMA, does not scale very well with transmission speed. A spreading factor of 64 or 128, corresponding to the use of codes (or chip sequences) with length of 64 or 128 bits respectively, is acceptable on signals whose bandwidth is a few hundreds kHz. Probably it is not easy to apply the same spreading factor to signals whose bandwidth is a few tens of MHz.

An interesting proposal, that opens new fields for research, has been presented in [11], where a system that allows the channel subdivision with a technique called "*Capture-Division*" is analyzed both theoretically and via simulation, showing that it can offer advantages with respect to traditional TDMA/CDMA schemes. The authors start from the observation that the near-far effect in a picocellular environment can be exploited to dynamically connect the MT to the best possible BS instead of trying to compensate for it. This access scheme shows the best performance where the attenuation is very high, such as in picocellular environments where very high frequencies are used. It must be coupled with a scheme that allows macrodiversity to be used at least at the radio level, since the access point to the network changes continuously in time due both to the terminal movements and the attenuation fluctuation, and it can not be imagined that handovers occur with such high rates.

3.2 RADIO HANDOVER PROCEDURES

The provision of high data rates over a radio interface implies the use of high frequency carriers, in order to have enough bandwidth to accommodate services. Depending on the different standards and proposals, carriers can be accommodated in frequency bands from 5 to 60 GHz [12, 13], while other proposals foresee the use of microwave or laser techniques. In all cases the signal attenuation in air is so high that the coverage of radio cells can not be more than a few tens of meters in radius, with the implicit assumption that the frequency of radio handover events is potentially very high, even for slowly moving terminals. In addition it is possible that a stationary terminal must undergo a handover just because or random field variations. If microwave or lasers are used, moreover, the two antennas must be in line of sight, so that, if the connection must be guaranteed over time, the same area must be covered by different antennas, and handover procedures are necessary all the times an object shadows the BS from the MT view.

For this reason the radio handover must be very fast and the use of macrodiversity, i.e., the possibility for the MT to be connected to two or more BS at the same time, has a great appeal. Macrodiversity at the radio level, especially in indoor environments, where the multipath delays are not very large is an available technology, whose price is the use of rake receivers (see for instance [14], Section 8-4.5). The use of rake receivers connected to multiple antennas is a fairly straightforward extension of their use for multipath decoupling, at least if the path delay through the different antennas is close enough.

As indicated in Section 2 a base station of a W-ATM network can control a number of micro- or picocells, each one connected to a port of the BS. The control part and the receiving/transmitting parts (the ports) of the BS can be co-located, for instance at the center of a “multicell” covered with directional antennas, of the control part can be completely separate from the receiving/transmitting parts and connected to them via cables, as for instance in GSM networks. The problems involved in handover procedures can be fairly different in the case when the handover takes place between different ports of the same BS or when different BS are involved. Besides the problem of connection re-routing, that will be addressed in Section 4.1, also the radio handover can have striking differences. In fact, a radio handover performed between micro- or picocells that are controlled by the same base station involves only two *logical* entities: the mobile terminal and the base station. The handover protocol can, in this case, be very simple, and it can be easily foreseen that macrodiversity will be used. When the mobile terminal must change the base station, on the other hand, the logical entities involved in the procedures are at least three (if the two BS are directly connected to one another). The handover protocol is obviously much more complex, and the use of macrodiversity becomes a problem unless macrodiversity is provided directly at the ATM level: which is a topic that has not been tackled yet by the research community.

The performance metrics of interest in the comparison of different radio handover schemes are 2:

- the disruption time, i.e., the time lapse during which the mobile can not communicate with either base stations (clearly this time is zero if macrodiversity is used, since with macrodiversity the absence of *any* channel would mean a broken connection);
- the stability of the handover procedure, i.e., the ability of the handover protocol to avoid un-necessary handovers when the MT lies on the border between two or more cells.

3.3 MEDIUM ACCESS CONTROL PROTOCOL

Traditional mobile networks are circuit switched, and one radio channel is dedicated to each active mobile terminal. If this solution is perfectly suitable for voice services, when considering W-ATM, the workload can be fairly different and the traditional solution fails. In presence of highly bursty traffic, like in computer communicationsf, the connection is idle most of the time, so that a dedicated channel can result in unacceptable inefficiencies.

As W-ATM is intrinsically packetized (ATM cells), the use of a random access broadcast channel seems to be a possible solution, if not the only possibility for some application scenarios like W-ATM LANs.

Research on MAC protocols for the access of wireless and cellular networks has been going on for many years. Voice compression techniques and the use of packet networks for voice services suggested the idea of transmitting only *talkspurts*, suppressing the silence periods. During silences the channel can be used for other communications. Of course most of the proposals for MAC protocols suited for wireless cellular networks are not limited to W-ATM, but have broader application to any slotted packetized wireless access network.

The proposals can be broadly grouped in centralized "polling-like" protocols, and contention resolution protocols, with many intermediate solutions that make it difficult to clearly separate the two groups. An interesting overview and comparisons between different approaches can be found in [15], even if this paper does not deal with W-ATM access schemes, but with packetized cellular networks in general. The description of two MAC protocols specifically designed for W-ATM can be found in [16, 17], while [18] reports the problems and design objectives of MAC protocols for W-ATM.

When considering the MAC protocols, the grounds of comparison between different protocols are the transmission medium exploitation and both the average and the jitter of the channel access delay. The former is particularly critical in wireless networks, because the capacity on radio channels is a really precious resource. One further issue must be considered, especially in public network: the fairness. When accessing a resource users with the same characteristics and requirements must receive the same Quality of Service, regardless of the congestion state of the network. The fairness can not be defined only for the average value as the time goes to infinity as, for instance in CSMA-CD[2] protocols, but must be granted also during short time periods.

It must be pointed out that the transmission channel for W-ATM, like the one for most cellular networks, can be split in two separate subchannels with drastically different characteristics: the "*downlink*" from the base station to the mobile terminals, and the "*uplink*" from the mobile terminals to the base station.

The downlink is essentially a point-to-multipoint broadcast channel, and there is indeed no need for a proper protocol in order to exploit it. The efficient use of this channel is in practice a *scheduling* problem, and the mobile terminals are not involved in its management.

[2] Carrier Sense Multiple Access with Collision Detection (CSMA-CD) is the MAC protocol used in Ethernet LANs. It is well known that it ensures fairness among different stations, but only in a statistical way and regarding averaged values

The uplink, on the other hand, is a multipoint-to-point channel, and an efficient protocol for the coordination of mobile terminals' transmission, as well as for the resolution of contention when they occur, is essential for the efficient exploitation of the channel.

In the former case the exploitation of the transmission resources is easy to obtain, and the main performance metrics are the channel access delay and the per-service QoS that the scheduling algorithm is able to guarantee. In the latter case, beside the QoS and access delay to the channel, the resource exploitation is also a very important metric.

Contention Resolution Protocols are a direct extension of protocols like ALOHA and CSMA-DC widely used in wired LANS. The basic idea that has lead research in this area is the exploitation of the correlation in voice or data transmission, limiting the contention phase in early stages of an activity period and, when the contention is resolved, somehow reserving slots for the subsequent transmissions.

The first ideas relative to such protocols can be found in the Reservation ALOHA protocol [19], proposed in 1973: a number of years before even the idea of ATM was born! The same concepts have been extended and refined in a protocol named Packet Reservation Multiple Access (PRMA) [20, 21, 22].

The baseline of all such algorithms is the observation that packet radio slots are generally very small, so that even a small amount of information will occupy several slots. It is assumed that the channel, beside being slotted, is also organized in*frames*, i.e., the sequence of slots has a structure that repeats over time. For instance the frame can be made of N slots, the first R being reserved for signaling and maintenance purposes, for the distribution of the clock and so on. Of the remaining $N - R$, the first K are used by the base station to send information to the mobile terminals and the last $N - R - K$ can be used by MTs. Among the first R slots one or more are dedicated to signal the slots in the frame that are currently in use and are hence not available for contention. When an MT, say MT$_j$, wants to start or resume transmission, it transmits on one of the non already occupied slots, with a simple ALOHA protocol. If a collision occurs the transmission is re-scheduled in the following frame. As soon as the first transmission is successful in slot i, the ith slot of each subsequent frame is reserved to MTj. The various proposals based on this scheme differ mainly in the frame organization and in the specific protocol that allows the reservation of slots. The more sophisticated versions of PRMA access, like C-PRMA [21], include a centralized scheduling algorithm that allows the support of different services and different QoS on the same shared transmission medium.

Centralized Protocols exploit the negligible propagation delay granted by a micro- or picocellular environment. Indeed, if the propagation delay is zero, a protocol based on polling is almost ideal, so that polling can really be a solution

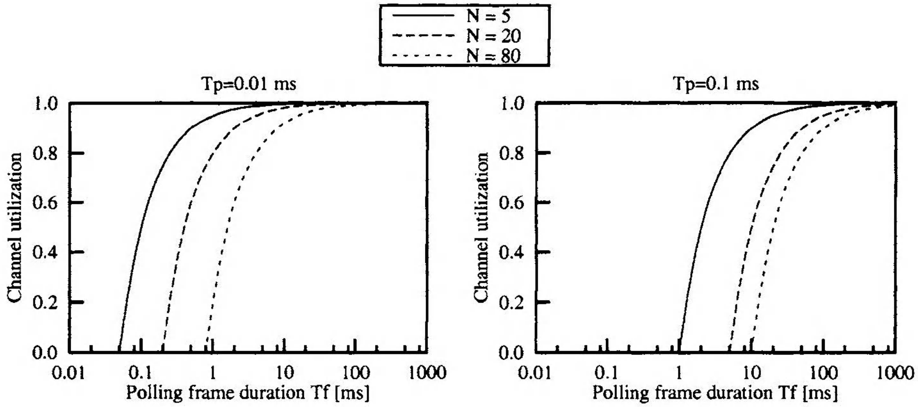


Figure 13.5 Upper bound to the channel utilization with polling MAC protocols

for voice and data transmission in picocellular environments. Examples of proposals based on polling can be found in [23, 24].

As pointed out in [25] the efficiency of any polling scheme is upper bounded by $\eta = \frac{T_f - N \cdot T_p}{T_f}$, where T_f is the duration of the polling frame, N is the number of mobile terminals to be polled, and T_p is the time "wasted" to poll a single MT, that depends on the transmission technology and conditions. Fig. 13.5 reports the plots of η as a function of T_f for different values of N and T_p ; the propagation delay, if not negligible, can be included in T_p.

From Fig. 13.5 it is quite clear that polling systems are heavily influenced by the minimum polling interval T_f required by services, as well as by the technological issues that define T_p. The number of active mobile terminals, on the other hand, seems to be a minor issue since in micro- and picocellular environments the number of MTs within the same service area will be rather small.

4 HANDOVER MANAGEMENT

The introduction of mobile terminals *within* the B-ISDN, poses a number of problems in the management of the network itself that were not considered during the standardization process.

Some of these problems are similar to those encountered in present days cellular networks and will not be considered in detail in this paper. For instance all problems related to addressing for mobile networks, location management, tariffing transparency and similar issues are not different in nature when W-ATM is considered instead of other cellular networks.

On the other hand the management of mobility is completely specific to W-ATM. Traditional cellular networks, in fact, were born with the specific aim of offering services to mobile terminals. ATM instead was born as an unifying technique for wired networks, hence, in the whole architecture of "traditional" ATM the possibility that a terminal is mobile is not even considered. For this reason new ideas on how to manage mobile terminals within an ATM network are badly needed, so that standardization bodies can have all the technical support needed to define sound and durable standards.

4.1 APPROACHES FOR NETWORK HANDOVERS

When a mobile terminal roams through the network hopping from one base station to the next, it is necessary to provide proper procedures for handling the connection re-routing through the network. This procedure can be termed *network handover* as opposed to the *radio handover* procedure briefly discussed in Section 3.2. A radio handover deals with the problem of changing the transmission channel, and is basically a procedure limited to the lower layers of the protocol stack (with reference to Fig. 13.2 we can assume that only the radio layers are involved). A network handover deals instead with the problem of modifying the connection route within the fixed part of the network in order to follow the mobile terminal.

The network handover procedure is clearly dependent from the network architecture, so that network handovers in W-ATM are different from those in non ATM wireless networks. It must be pointed out that the radio and network handover procedures are, at least in principle, independent from one another, so that the handover procedure architecture adopted at the radio level does not influence the handover procedure architecture at the network level, with the notable exception that during the execution of a network handover, a radio handover must be executed too.

A remarkable exception to the independence of radio and network handover is the use of macrodiversity at the ATM level, in this case the radio handover procedure must have macrodiversity capabilities too, otherwise it would be impossible to have multiple channels at the ATM level. This case will be discussed in some detail later on.

The different approaches proposed to handle network handovers can be broadly subdivided into four categories, which have completely different characteristics, performance, and impact on the ATM standards and definitions:

1. full-establishment
2. connection extension
3. incremental re-establishment
4. multichannel establishment.

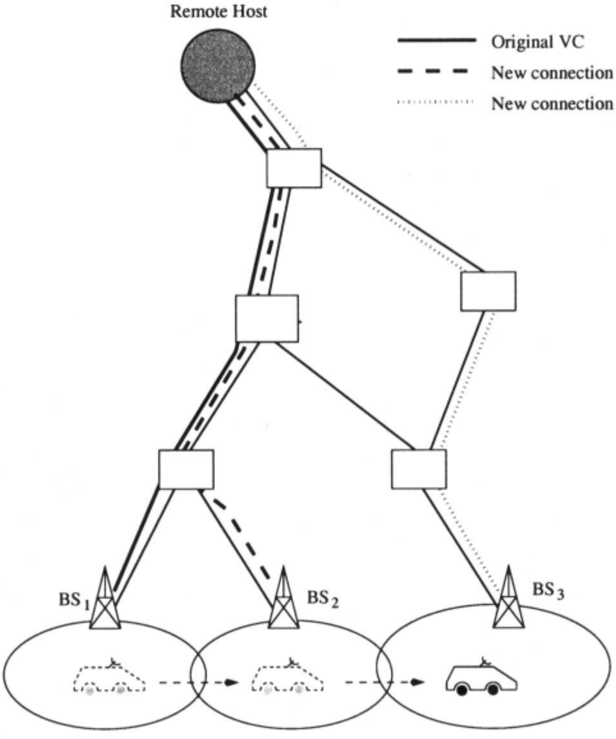


Figure 13.6 Path evolution of the VC while the MT roams through three macrocells: full-establishment case

Similar subdivisions are also found in [9, 26, 27, 28], but in [27], the last category is termed *multicast establishment*, and takes into account only the case of macrodiversity at the ATM layer, while in [26] it is not mentioned.

In the following the term macrocell will always be used referring to the area covered by a base station, with the implicit assumption that the macrocell can be divided in microcells, but handovers between microcells do not affect the ATM layer. In addition the two BS involved in the handover procedure are named *source* BS and *destination* BS, with respect to the MT movement.

The full-establishment approach requires the setup of a completely new connection between the end terminals. This is one of the earliest proposals, and it has a minor impact on the fixed network architecture. However, this procedure may not be sufficiently fast to guarantee that handovers do not cause timeouts to expire and connections to be abruptly terminated. In addition, both terminals must be involved in the path re-establishment operation. Fig. 13.6 pictorially represents the VC evolution for a mobile terminal that crosses three macrocells in a network adopting the full-establishment approach.

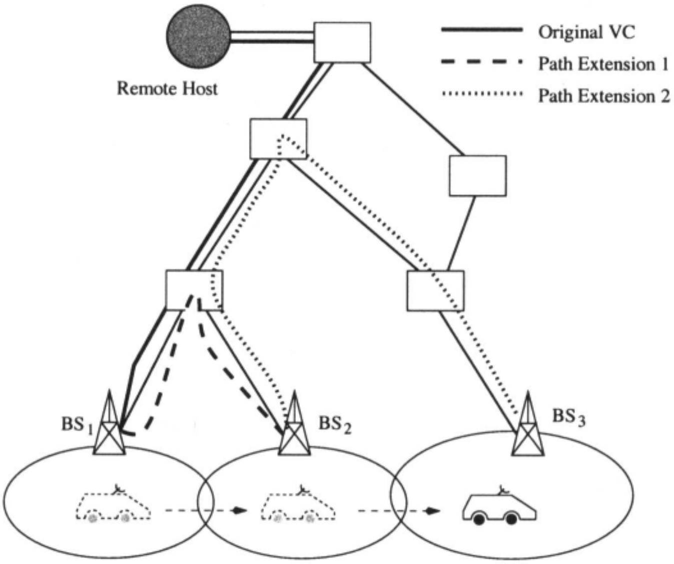


Figure 13.7 Path evolution of the VC while the MT roams through three macrocells: connection extension case

The connection extension technique extends the VC between the terminals at each handover by adding one hop that provides the connection from the source BS to the destination BS through the fixed network. As proposed in [29], this path extension can be performed by the source BS as shown in Fig. 13.7, or, as proposed in [30], it can be performed by the node where the base station is connected. The advantage of this approach is twofold: simple and reasonably fast execution, and intrinsic preservation of ATM cell sequence. As no re-routing is performed, some inefficiency may arise, specially when the mobile user circulates in a limited area, possibly returning to previously visited BSs. In this case closed loops may form in the connection path. Fig. 13.7 shows the VC modifications needed to follow the roaming terminal. It is quite evident that, although no closed loop arises in the illustrated situation, the resource waste is remarkable.

It is interesting to notice that in GSM networks the call control is maintained, throughout the connection duration, by the MSC (Mobile Switching Center) where the call has been established. In other terms the connection control is kept by the first node the MT has contacted, even if multiple handovers bring the MT very far from it and additional MSC are involved. Hence GSM handover belong to the connection extension category. The same can be said for all other first and second generation mobile networks.

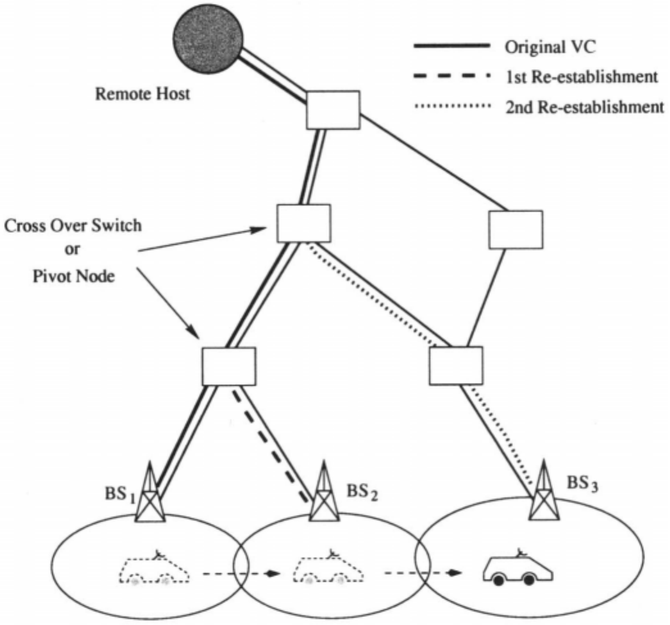


Figure 13.8 Path evolution of the VC while the MT roams through three macrocells: incremental re-establishment case

The incremental re-establishment is the handover category where most research work is performed. This technique is appealing because it requires only the establishment of a new partial path (without the involvement of the remote host) which connects to a portion of the original connection path, therefore allowing virtual circuits to be partly reused [27, 31, 32]. Note that, because of spatial locality in movement, it is very likely that the re-established path to the new location of the mobile user shares most of the VPs in the original path. As a consequence, this technique is expected to be fast, efficient and transparent, so that it can be imagined that the end user does not perceive the network handover as a service interruption. Fig. 13.8 shows the path rerouting performed while the terminal moves through the network. At each handover the optimal path is established, thus avoiding resource waste. The figure also indicates the Pivot Node (PN), i.e., the ATM switch that connects the original path to the incremental path for the handover occurring from BS$_2$ to BS$_3$. Some authors call this switch the Cross-Over Switch (COS or CX) and specific algorithms to find out the best Pivot Node among all the nodes along the connections have already been studied [33].

A two-phase handover was recently proposed in [34], that combines the advantages of both connection extension and incremental re-establishment. The rationale behind this hybrid approach is the use of a fast procedure to

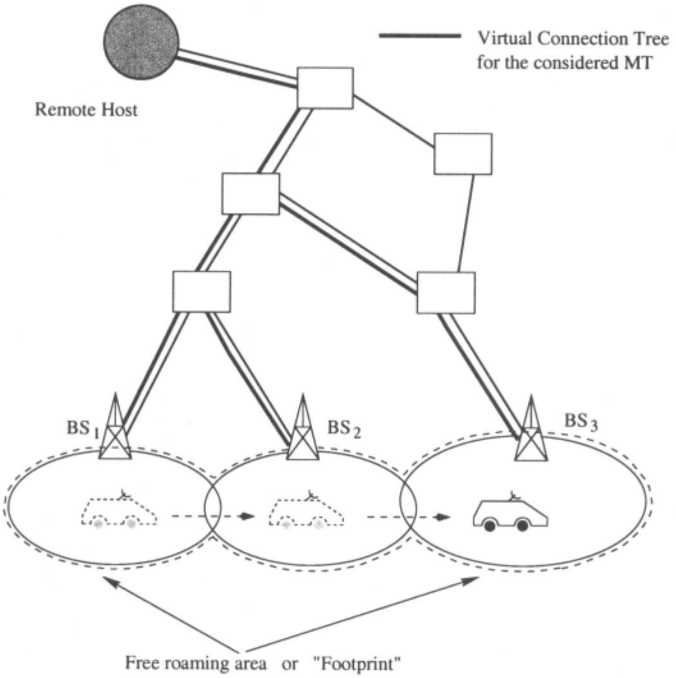

Virtual Connection Tree for the considered MT
Remote Host
BS1
BS2
BS3
Free roaming area or "Footprint"

Figure 13.9 Virtual Tree Path of the VC while the MT roams through three macrocells: multichannel establishment case

handle the connection extension during handover, followed by the optimal VC re-establishment procedure, that is activated once the MT is already connected to the destination BS. Such an approach is particularly prone to cell misordering, so that specific procedures must be devised to avoid it.

The multichannel establishment approach, finally, preallocates resources in the network portion surrounding the macrocell where the mobile user is located. When a new mobile connection is established, a set of virtual connections, called *virtual connection tree*, is created, reaching all BSs managing the macrocells towards which the MT might move in the future. Thus, the mobile user can freely roam in the area covered by the *tree* (some authors call this area the "footprint"), without invoking the network call acceptance capabilities during handover. The allocation of the *virtual connection tree* may be static [25] or dynamic [35] during the connection lifetime. This approach is fast and statistically guarantees the Quality of Service (QoS) contract in case of network handover, since the QoS negotiation is executed only once, at connection establishment, allocating resources in all the macrocells where the mobile user is expected to roam. However, this approach may not be efficient in terms of network bandwidth utilization, since it introduces the possibility of refusing a connection because of lack of resources that may never be needed, and it introduces high signaling overheads, specially in the case of dynamic tree allocation. Fig. 13.9 shows a multichannel establishment, assuming that the MT moves within three macrocells. All the paths to the macrocells where the mobile is likely to roam are open all the time, although only one is used at any given time.

One particular case of multichannel establishment is the *multicast establishment* [36, 37], that actually must exploit macrodiversity at the ATM level; such an approach, at least in WAN, public networks, is completely different from the others and deserves some specific attention. Indeed, in ATM standards, there is no means to support macrodiversity, and the duplication of cells is a potentially destructive event. Support for multicast transmission in ATM generally assumes different receivers (different VCI at the switch) for each duplicated cell. This fact however give a hint on how to solve the problem, in fact multiple VC can be opened to and from the MT, and multiple cells due to macrodiversity can be sent over different VCs. A means is needed to handle flows alignment and the redundancy reduction before the cells enter the fixed network. However *connection segmentation* concepts by means of dedicated Resource Management cells (similar to those proposed in [38] in a slightly different scenario) could solve the problem.

4.2 DIFFERENT HANDOVER TYPES

Traditional cellular networks are dedicated to a single service; besides they support a single radio network architecture. Hence they support a single handover type, that depends on the radio layer and signaling architecture. For instance GSM makes use of a single radio transceiver. During handovers, it first tears down the connection between MT and the source BS, then sets the connection between MT and the destination BS. All the signaling takes place on the new channel. This type of handover can be called a *hard, forward* handover, since the change in radio channel is abrupt (hard) and the signaling is done on the new channel with the destination BS (forward). CDMA networks handovers are clearly different. Multiple radio channels are allowed (soft handover) and the signaling is used to define how many and which channels are currently used by the MT.

ATM mobile networks should provide for many different services and support several radio architectures. This means that ATM mobile networks must be capable of handling different handover types and procedures. This fact has been recognized both by the ATM Forum and by the research community. Handover procedures classification and possible protocols can be found in [8, 41, 42].

4.3 PERFORMANCE OF NETWORK HANDOVERS

A mobile B-ISDN terminal should have the perception of being just like any other B-ISDN terminal, hence having exactly the same QoS during the whole connection duration. As a matter of fact performance issues related to the ATM mobility management are mainly concerned with handover. Whatever is the specific technique chosen for handover management, both at the radio level and at the network level, the resulting procedure must offer a seamless service to the connections of the roaming terminal. The case of unavailable resources (e.g., the transmission channel at the destination BS) is not considered. Resources availability is a matter of network planning and not a performance figure of the handover procedure.

Given this performance target, the main metrics against which the different techniques should be compared are essentially the following.

- The handover disruption time. This is essentially the time during which the communication channel is not available for transmission/reception of information. If soft handover techniques are used both at the radio and at the network level this time can be reduced to zero.
- The information loss rate. This metrics summarizes the information degradation due to different phenomena, starting from the cells loss rate during handovers, to the probability of duplicating information at the ATM level, to the delivery of information with too much delay for real-time applications. Cell misordering must in any case be avoided with proper protocols, so that a cell mis-sequence is eventually transformed into cell loss or delay jitter.
- The additional buffering required for handover management. If the handover procedure is something more sophisticated than a simple connection "*break and re-make*", then the network must provide additional buffering capabilities to store and retrieve the information that can not be transmitted during handovers. Even if soft handovers are considered it might be possible that the information must be stored in order to be able to re-align the information flows that have followed different paths.
- The procedure complexity. As always the complexity comparison is somewhat more difficult than other performance comparison, since its definition is not unique. However it is quite clear that a procedure simple to implement, that require the exchange of a small number of messages will be less expensive and more robust than a cumbersome procedure with many messages to exchange.
- The required resources. Besides the protocol complexity, another measure of the procedure efficiency is the amount of resources, e.g., signal

processors, RF transmitters/receivers, that are required for the procedure to work properly. As an example it can be easily foreseen that soft handover procedures at the radio level will be less complex and expensive using CDMA techniques that using FDMA/TDMA techniques, since in the former case only a single transmitter/receiver is required, while in the latter case, al least in principle, more than one is required.

The performance comparison of different mobility management schemes, together with the embedded handover procedures, is a topic that is just starting to be addressed and discussed by the research community. Examples of these preliminary studies can be found in [27, 32, 38, 39, 40], but much more research is needed in order to analyze and comprehend the problem.

5 CONCLUSIONS & HOT RESEARCH TOPICS

This paper has presented an introduction to the basic concepts of W-ATM: one of the most active research areas in telecommunication s point) in Japan, and has strong commitment to contribute to the networks.

The research topics in W-ATM can be broadly divided into two categories: topics related to the radio access, i.e., the transmission channel between the mobile terminals and the base stations, and topics related to network management in presence of mobiles. Of course, many research areas cover topics in both categories and, given the complexity of the scenario, solutions taking into account the overall system must be sought for.

The areas where research is more active range from the study of suitable MAC protocols for high radio resources exploitation, to the study of handover procedures directly embedded within the ATM layer, to the provision of high data rates over the radio interface.

Acknowledgments

I wish to thank the whole Telecommunication Networks Research Group of Politecnico di Torino, whose support has made this work possible. In particular Prof. Marco Ajmone Marsan for his continuous incitement and advice; Andrea Bianco, Andrea Fumagalli and Maurizio Munafò for the useful comments and discussions, and most of all Carla Fabiana Chiasserini, whose stubborn determination in pursuing research in this area has forced me to enter the field.

References

[1] *IEEE Journal on Sel. Areas in Communications* Vol. 12, No. 8, Oct. 1994.

[2] *IEEE Journal on Sel. Areas in Communications* Vol. 14, No. 4, May 1996.

[3] *IEEE Personal Communications – Special Issue on Wireless ATM*, Vol. 3, No. 4, Aug. 1996.

[4] *Mobile Networks and Applications (MONET) – Special Issue on Wireless ATM*, Vol. 1, No. 3, ACM/Baltzer, Dec. 1996.

[5] *IEEE Communication Magazine – Introduction to Mobile and Wireless ATM*, Vol. 35, No. 11, Nov. 1997.

[6] J. Rapeli, "UMTS: Targets, System Concept, and Standardization in a Global Framework", *IEEE Personal Comm.*, pp. 20–28, Feb. 1995.

[7] Buitenwerf, G. Colombo, H. Mitts, P. Wright, "UMTS: Fixed Network Issues and Design Options," *IEEE Personal Comm.,* pp. 30-37, Feb. 1995.

[8] R. R. Bhat (Editor), ATM Forum BTD-WATM-01.08, Wireless ATM Baseline Text.

[9] M. Ajmone Marsan, C. F. Chiasserini, A. Fumagalli, R. Lo Cigno, M. Munafò, "Local and Global Handovers for Mobile Management in Wireless ATM Networks", *IEEE Personal Comm.,* Vol. 4, No. 5, pp. 16-24, Oct. 1997.

[10] S. H. Jamali, T. Le Ngoc, Coded-Modulation Techniques for Fading Channels", Kluwer Academic Publisher, Boston, MA, USA, 1995.

[11] F. Borgonovo, M. Zorzi, L. Fratta, V. Trecordi, G. Bianchi, "Capture-Division Packet Access for Wireless Personal Communications", *IEEE JSAC,* Vol. 14, No. 4, May 1996.

[12] L. Fernandes, "Developing a System Concept and Technologies for Mobile Broadband Communications", *IEEE Personal Comm.*, pp. 54–59, Feb. 1995.

[13] P. F. Driessen, L. J. Greenstain, "Modulation Techniques for High-Speed Wireless Indoor Systems Using Narrowbeam Antennas", *IEEE Trans. on Comm.,* pp. 2605-2612, Vol. 43, No. 10, Oct. 1995.

[14] R. L. Peterson, R.E. Ziemer, D.E. Borth, "Introduction to Spread Spectrum Communications", Prentice Hall, NJ, USA, 1995.

[15] C. G. Choudary, S. S. Rappaport, "Cellular Communication Schemes Using Generalized Fixed Channel Assignment and Collision Type Request Channels", *IEEE Trans. on Vehicular Technology*, Vol. 31, pp. 53–65, May 1982.

[16] N. Passas, S. Paskalis, D. Vali, L. Merakos, "Quality-of-Service Oriented Medium Access Control for Wireless ATM Networks", *IEEE Communication Magazine*, Vol. 35, No. 11, Nov. 1997.

[17] L. Dellaverson, W. Dellaverson, "Distributed Channel Access on Wireless ATM Links" *IEEE Communication Magazine,* Vol. 35, No. 11, Nov. 1997.

[18] O. Kubbar, H. T. Mouftah, "Multiple Access Control Protocols for Wireless ATM: Problems, Definition, and Design Objectives", *IEEE Communication Magazine*, Vol. 35, No. 11, Nov. 1997.

[19] W. Crowther, R. Rettberg, D. Walden, S. Ornstain, F. Heart, "A System for Broadband Communication: Reservation-ALOHA", *Proc. 6th Hawaii Int. Conf. Syst. Sci.,* pp. 596–603, Jan. 1973.

[20] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, B. Ramamurthi, "Packet Reservation Multiple Access for Local Wireless Communications", *IEEE Trans. Comm.*, Vol. 37, pp. 885-890, Aug. 1989.

[21] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci, M. Zorzi, "C-PRMA: the Centralized Packet Reservation Multiple Access for Local Wireless Communications", *Proc. IEEE GLOBCOM '94,* San Francisco, CA, U.S.A., pp. 1340–1994, Nov. 1994.

[22] P. Narasimhan, R. D. Yates, "A New Protocol for the Integration of Voice and Data over PRMA", *IEEE JSAC,* Vol. 14, No. 4, May 1996.

[23] Z. Zhang, A. S. Acampora, "Performance of a Modified Polling Strategy for Broadband Wireless LANs in a Harsh Fading Environment", *Telecommun. Syst.*, Vol. 1, pp. 279-294, Feb. 1993.

[24] A.S.Mahmoud, D. D. Falconer, S.A. Mahmoud, "A Multiple Access Scheme for Wireless Access to a Broadband ATM LAN Based on Polling and Sectored Antennas", *IEEE J*SAC, Vol. 14, No. 4, pp. 596–608, May 1996.

[25] A. S. Acampora, M. Naghshineh, "An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks", *IEEE JSAC*, Vol. 12, No. 8, pp. 1365-1375, Oct. 1994.

[26] B. Rajagopalan, "Mobility Management in Integrated Wireless-ATM Networks", *ACM/Baltzer MONET Special Issue on Wireless ATM*, Vol. 1, No. 3, pp. 273-285, Dec. 1996.

[27] C.-K. Toh, a Hybrid Handover Protocol for Local Area Wireless ATM Networks", *ACM/Baltzer MONET Special Issue on Wireless ATM*, Vol. 1, No. 3, pp. 313-334, Dec. 1996.

[28] A. Acharya, J. Li, B. Rajagopalan, D. Raychaudhuri, "Mobility Management in Wireless ATM Networks", *IEEE Communication Magazine,* Vol. 35, No. 11, Nov. 1997.

[29] M.J. Karol, K.Y. Eng, M. Veeraghavan, E. Ayanoglu, "BAHAMA: A Broadband Ad-Hoc Wireless ATM Local-Area Network", *ACM/Baltzer Wireless Networks Journal,* Vol. 1, Issue 2, pp. 161-174, 1995.

[30] T. La Porta, "Distributed Processing for Mobility and Service Management in Mobile ATM Networks", *Wireless ATM Networking Workshop*, New York, Jun. 1996.

[31] R. Yuan, S. K. Biswas, L. J. French, J. Li, D. Raychaudhuri, "A Signaling and Control Architecture for Mobility Support in Wireless ATM Net-

works", *ACM/Baltzer MONET – Special Issue on Wireless ATM*, Vol. 1, No. 3, pp. 287-298, Dec. 1996.

[32] M. Ajmone Marsan, C. F. Chiasserini, A. Fumagalli, R. Lo Cigno, M. Munafò, 'Buffer Requirements for Loss-Free Handovers in Wireless ATM Networks" IEEE ATM'97 Workshop, Lisboa, PL, May 1997.

[33] C.-K. Toh, "Crossover Switch Discovery for Wireless ATM LANs", *Journal on Mobile Networks and Applications*, No. 1, pp. 141-165, 1996.

[34] M. Veeraraghavan, M. Karol, K. Y. Eng, "A combined Handoff Scheme for Mobile ATM Networks", ATM Forum/WATM - ATM_Forum/96-1700, Vancouver, Canada, Dec. 1996.

[35] O. Yu, and V. Leung, "Extending B-ISDN to Support User Terminal Mobility over an ATM-Based Personal Communications Network", *Proc. GLOBCOM'95,pp.* 2289-2293, 1995.

[36] R. Earnshaw, "Footprints for Mobile Communications", *Proc. of the 8th IEEE U.K. Tele-Traffic Symposium* Apr. 1991.

[37] R. Ghai, S. Singh, "A Protocol for Seamless Communication in a Picocellular Network", *Proc. of Supercomm/ICC'94,* May 1994.

[38] H. Mitts, H. Hansen, J. Immonen, S. Veikkolainen, "Lossless Handover for Wireless ATM," *ACM/Baltzer MONET Special Issue on Wireless ATM,* Vol. 1, No. 3, pp. 299–312, Dec. 1996.

[39] M. Ajmone Marsan, C.F. Chiasserini, A. Fumagalli, R. Lo Cigno, M. Munafò, "Local and Global Handovers Based on In-Band Signaling in Wireless ATM Networks", *ACM Mobile Computing and Communications Review*, Vol. 2, No. 3, Jul. 1998.

[40] M. Cheng, S. Rajagopalan, L. Fung Chang, G. P. Pollini, M. Barton, "PCS Mobility Support over Fixed ATM Networks", *IEEE Communication Magazine*, Vol. 35, No. 11, Nov. 1997.

[41] C. F. Chiasserini, R. Lo Cigno, E. Scarrone, "Handovers in Wireless ATM: An In-Band Signaling Solution", *Proc. of IEEE ICUPC'98,* Florence, Italy, Oct. 1998.

[42] M. Ajmone Marsan, C. F. Chiasserini, P. Di Viesti, A. Fumagalli, R. Lo Cigno, E.Scarrone, "In-Band Signaling for Handover and Mobility Management in Wireless ATM Networks", *Technical Report* DTD 98.0446, CSELT, Jun. 1998.

Chapter 14

SATELLITE ATM NETWORKS

Zhili Sun
Centre for Communication Systems Research,
University of Surrey, Guildford, Surrey GU2 5XH, UK,
E-mail: Z.Sun@ee.surrey.ac.uk, Tel: (+44) (0)1483 87 9493, Fax: (+44) (0)1483 87 9504

Abstract: This paper is to provide an introduction to satellite ATM networks. It presents an overview of the important issues and the recent development of satellite systems for ATM networks and broadband communications. Particularly, it discusses the architecture and performance of broadband network interconnection and terminal access using ATM over satellite. It covers a range of topics including: the major issues on the role of satellites in ATM networks, satellite ATM system structure and architecture, management and control over satellite, performance aspects of ATM over satellites, satellite bandwidth resource management, future satellite systems and convergence of ATM and Internet.

Keywords: Satellite, ATM, B-ISDN, Network, Protocol, Internet.

1. INTRODUCTION

The space era started in 1957 with the launching of the first artificial satellite followed by various experimental satellites. In 1965 the first commercial geostationary orbit satellite INTELSAT I (or Early Bird) inaugurated the long series of INTELSAT satellite services; in the same year, the first Soviet communication satellite of the MOLNYA services was launched. Since then, satellites have played more and more important role in the world communications infrastructure.

In the recent years, significant progress has been made in the research into broadband communications based on ATM and fibre optic cable. It generates an increasing demand for cost-effective interconnection of private and public broadband islands including ATM LANs, DQDB MANs and experimental ATM networks and testbeds and also for cost-effective access to these broadband islands [1] [5]. However there is a shortage of broadband terrestrial connections in wide areas, particularly in more remote or rural areas where terrestrial lines are expensive to install and operate. Therefore, significant research and development have been carried out on satellite systems to complement terrestrial networks by extending the broadband networks with its flexibility and immediate wide coverage.

This paper aims to provide an introduction on how satellite ATM systems provide interconnection and also access to geographically dispersed broadband islands, and how this could further stimulate the introduction of broadband applications and services across Europe in a wide area and large scale.

Due to the global coverage and broadcasting nature of satellite systems, satellite can also be best used for broadband mobile and broadcasting services. The major technology challenge is how to design a mobile terminal for broadband services that it has to be small and capable of high speed transmission.

The satellite ATM system should provide direct compatibility with the future ATM based B-ISDN. It is widely recognised that development of B-ISDN based ATM will not be a revolution but an evolution. This also requires that satellite ATM system has to be designed to be able to interconnect the ATM networks as well as existing networks such as the LANs and MANs.

By interconnecting these broadband islands and providing terminal access to broadband networks, the initial ATM based B-ISDN can be introduced, thus getting the B-ISDN started. In this way, the satellite ATM system can support data, voice, video and multimedia applications. Some experiments have been done to demonstrate such broadband services and application over satellite ATM system. In the light of the experiments, relevant issues and the impact of ATM over satellite on the applications and the protocols can be discussed.

2. SATELLITE SERVICES AND THE ROLE IN THE B-ISDN ENVIRONMENT

The principal advantages of satellite systems are their wide coverage and broadcasting capabilities. European satellites can provide broadband connections anywhere in Europe and some peripheral countries. The cost and complexity are independent of distance. They enable the broadband capabilities to be extended from the beginning to rural and remote areas. Satellite links are quick and easy to install irrespective of geographical constraints. It makes long distance connections more cost-effective within the coverage areas, particularly for point-to-multipoint and broadcasting services. Satellites can also be complementary to the terrestrial networks and suitable for providing interconnection of networks and mobile services.

From the radio communications point of view, there are three main classes of satellite services: fixed satellite services, broadcast satellite services, and mobile satellite services [14]:

- Fixed satellite services (FSS) concern all radio communication services between earth stations at given positions. The given position can be a specified fixed point or any fixed point within specified areas. These services provide transmission nationally or internationally on the basis of a network topology which can be transit, distribution or contribution type. They include, video, TV, sound and data type, primarily on a point-to-point basis (transit mode).
- Broadcast satellite services (BSS) gather video, TV, sound, data, and other types of transmissions intended to broadcast for direct reception by the general public. A common specification for FSS and BSS would be beneficial to service integration, sharing and flexibility. Broadcast involves one feeder uplink and a broadcast down link to home.
- Mobile satellite services (MSS) include all radio communications between a mobile earth station and the space station, or between mobile stations by the intermediate of one or more space stations. The class of transportable services seems fall partly between MSS and FSS with examples of each being currently used.

Table 14.1. Frequency allocations.

	Typical frequency bands for up/down link	Usual terminology
FSS	6/4 GHz 8/7 GHz 14/12 GHz 30/20 GHz	C band X band Ku band Ka band
MSS	1.6/1.5 GHz 30/20 GHz	L band Ka band
BSS	12 GHz	Ku band

From the networks point of view, the satellite system can be applied in two modes: user access and network transit. In broadband systems terminology, the following is applicable:

- In the user access mode, the satellite system is allocated at the border of the B-ISDN. The satellite network provides access links to a large number of users and the earth station provides a concentration point for multiplexing and de-multiplexing functions. The interfaces to the satellite system in this mode are of the User Network Interface (UNI) type on one side and the Network Node Interface (NNI) type on the other side.
- In the network transit mode, the satellite systems provide high bit rate links to interconnect the B-ISDN network nodes or network islands. The interfaces on both sides are NNI type [15] [17].

3. B-ISDN SERVICES AND QUALITY OF SERVICES

All satellite services can be extended to the B-ISDN environment for the future broadband communications to support B-ISDN services. Two main categories for the B-ISDN services has been specified from the point of view of the network: interactive services and distribution services [16].

The interactive services are subdivided into three classes of services:

- Conversational services: some typical examples are video telephony, video-conferences, video/audio information transmission, high speed digital information, file and document transfer;
- Message services: some typical examples are video mail and document mail; and
- Retrieval services: some typical examples are video, high resolution image, document and data.

The distribution services are subdivided into two classes:

- The class without user individual representation control (such as TV, document, video and audio distribution); and
- The class with user individual representation control (such as full channel broadcast videos).

In the ITU-T recommendation, a guideline has been provided for classification of specific standardised services to be supported by the B-ISDN. Some studies of the characteristics of these services have also been carried out in the areas of traffic engineering.

The traffic bit rates generated by the current services are in the range of 64 Kbit/s to 2 Mbit/s (such as telephony, data retrieval, video telephone and video conference). In the future, some services, such as high quality video telephony, high quality video conference and high speed data retrieval, the bit rate can be up to the range of 2 - 100 Mbit/s, and HDTV may generate a traffic with a bit rate of 140 Mbit/s.

Source coding algorithms can be used to compress the traffic. The traffic may be multiplexed/demultiplexed when passing through the networks. These may reduce the amount of traffic getting into the networks and change the characteristics of the traffic.

The ATM networks are able to handle both constant bit rate (CBR) services and variable bit rate (VBR) services. Different services with different coding and decoding techniques may generate a wide range of ATM traffic with different characteristics. Some of the services produce CBR traffic, some VBR traffic and some "burst" traffic. The traffic may be auto-correlated and correlated with each other.

Traffic sources can be described using traffic parameters such as peak cell rate (PCR), sustainable cell rate (SCR), maximum burst size (MBS) and minimum cell rate (MCR). The quality of service (QoS) is specified using parameters such as maximum cell transfer delay (maxCTD), cell delay variation (CDV) and Cell Loss Ratio (CLR) [8].

4. CHARACTERISTICS AND DESIGN OF SATELLITE ATM SYSTEMS

4.1 SATELLITE AVAILABILITY AND CONSTRAINS

The satellite ATM networks have been fundamentally different from terrestrial networks in terms of delay, error and bandwidth [6]. Satellite communication bandwidth being a limited resource will continue to be a precious asset. Achieving availability rates of 99.95% at very low bit error rate (BER) is costly. Lowering required availability rates by even 0.05% dramatically lowers satellite link costs. An optimum availability level must be a compromise between cost and performance.

There are constrains in general in choosing the satellite link parameters due to regulations, operational constrain and propagation conditions. The regulations are administered by ITU-R, ITU-T and ITU-D. They define space radio-communication services in terms of transmission and/or reception of radio waves for specific telecommunication applications. The concept of a radio communication service is applied to the allocation of frequency bands and analysis of conditions for sharing a given band among compatible services. A co-ordination procedure has been constituted between earth and terrestrial stations. The operational constrains related to realisation of a C/N_0 ratio, provision of an adequate satellite antenna beam for coverage of service area with a specified value of satellite antenna gain, level of interference between satellite systems, orbital separation between satellite operating in identical frequency bands and minimum of total cost.

4.2 QUALITY OF SERVICES (QOS)

The satellite long propagation delay can have a big impact on applications and services. For example, voice and video applications are more sensitive to the long delay than data applications. Delay variations can significantly degrade the QoS. The delay also affects throughput of the connections based on different protocols such as connection oriented and connectionless protocols. The connection oriented protocols requiring acknowledgements of packet arrival may need to increase the time-out parameter or window size to accommodate the long propagation delay (see [3] for TCP extensions). Hence adjustment of existing protocols or development of new ones are required to support the B-ISDN applications efficiently.

4.3 SATELLITE OBITS

Geostationary Orbit (GEO) satellites and satellite-based access scenarios have been used up to now in the existing operation satellites. GEO satellites have coverage areas spanning thousands of miles thus eliminating the need for call hand-off and minimising (or eliminating) the need for antenna tracking. However, in the new generation satellite constellations, the scenario involving Low Earth Orbit (LEO) or Medium Earth Orbit (MEO) satellites will have to address these issues in addition to the issues pertinent to the GEO case. Investigation of point-to-point links via GEO satellites for interconnection of broadband islands is an appropriate starting point. There is also near-term market need for this class of satellite networks.

5. GEO SATELLITE ATM SYSTEM ARCHITECTURE

5.1 THE GROUND SEGMENT

To make use of the existing satellite systems, development has been mainly on the ground segment [21] [22]. A modular approach was used in the design, where each module had buffer(s) for packet/cell conversion and/or traffic multiplexing. The buffers are also used for absorbing high speed bursty traffic.

Therefore, the satellite ATM system can be designed to be capable of interconnect different networks with the capacities in the range of 10 to 150 Mbit/s (10 Mbit/s for Ethernet, 34 Mbit/s for DQDB, 100 Mbit/s for FDDI and 150 Mbit/s for ATM networks). Figure 14.1 illustrates the model of the ground equipment. A brief description of these modules is also given in the following.

The ATM-LT provides an interface with a speed of 155 Mbit/s between the ATM network and the ground-station ATM equipment. It is also the termination point of the ATM network and passed the ATM cells to the ATM-AM module.

The Ethernet LAN Adaptation Module (E-LAM) provides an interface to the Ethernet local area network.

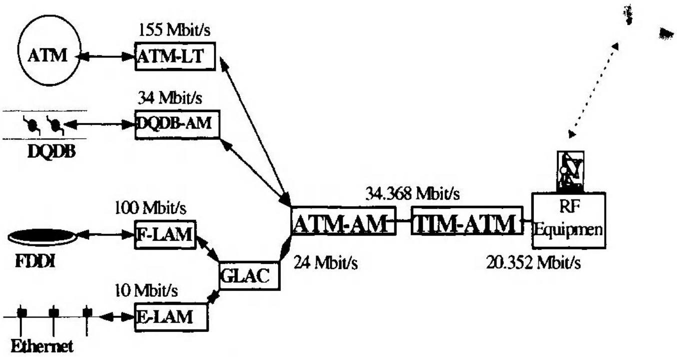


Figure 14.1. Ground Segment Model.

The FDDI LAN Adaptation Module (F-LAM) provides an interface to the FDDI network.

The Generic LAN ATM Converter (GLAC) module converts the FDDI and Ethernet packets into ATM cells, then passes the cells to the ATM Adaptation Module (ATM-AM).

The DQDB Adaptation Module (DQDB-AM) provides an interface to the DQDB network with a small buffer. It converts DQDB packets into ATM cells and then passes to the ATM-AM.

The ATM-AM is an ATM Adapter. It multiplexes the ATM cell streams from the two ports into one ATM cell stream. This module passes the cells to the Terrestrial Interface Module for ATM (TIM-ATM) and provides an interface between the terrestrial network and the satellite ground-station.

The TIM-ATM had two buffers with a "ping-pong" configuration. Each buffer can store up to 960 cells. The cells are transmitted from one buffer while the ATM-AM is feeding the cells into the other buffer. Transmissions of the buffers are switched every 20 ms.

5.2 THE SPACE SEGMENT

In the demonstrator system [21] [22], the EUTELSAT II satellite was used at a BER of 10^{-10} (99% of time in good weather conditions) with a 36 MHz (25 Mbit/s) bandwidth. The satellite propagation delay is a function of satellite orbit and earth station location, was about 250 ms.

The satellite has a link capacity of approximately 25 Mbit/s per transponder at present and will perhaps never be able to match the speed of optical fibre terrestrial networks. The satellite link capacity has to be shared by a number of earth stations when multiple broadband islands are interconnected. It is important for satellite to provide the required Quality of Service (QoS) with efficient utilisation of the satellite resources.

Compared to the propagation delay, the delay within the ground segment was insignificant. Buffering in the ground segment modules could cause variation of delay which was affected by the traffic load on the buffer. Most of the variation was caused in the TIM-ATM buffer. It could cause an estimated average delay of 10 ms and worst case delay of 20 ms. Cell loss occurred when buffer overflow. The effects of delay, delay variation and cell loss in the system could be controlled to the minimum by controlling the number of applications, the amount of traffic load and allocating adequate bandwidth for each application.

5.2.1 The TDMA as the multiple access control (MAC) scheme

There are three multiple access schemes: Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), and Code Division Multiple Access (CDMA). To interconnect broadband islands and support broadband services requires the satellite system and the multiple access scheme to be highly efficient, capable of supporting high speed, point to point and point to multipoint connections. TDMA was the most suitable solution for a small number of terminal with relatively high bit rates, hence it was chosen for the satellite ATM system.

In TDMA system, stations transmit traffic bursts that are synchronised so that they occupy non-overlapping time slots. These time slots are organised within periodic frames. All stations receive the down-link bursts, and a particular station can extract its traffic from these. The general TDMA format is shown in Figure 14.2.

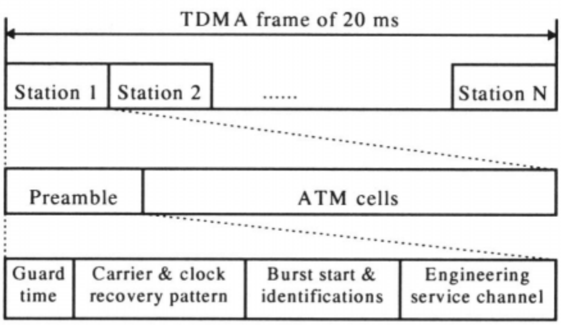

TDMA frame of 20 ms
Station 1
Station 2
......
Station N
Preamble
ATM cells
Guard time
Carrier & clock recovery pattern
Burst start & identifications
Engineering service channel

Figure 14.3 TDMA frame format (earth station to satellite).

The TDMA frame had a length of 20 ms which was shared by the earth stations. Each earth station was limited to the time slots corresponding to the allocated transmission capacity up to maximum 960 cells (equivalent to 20.352 Mbit/s).

5.2.2 Satellite link error control mechanisms

The commonly used mechanisms in addition to the re-transmission mechanism for error control used in satellite communications are Forward Error Control (FEC) and interleaving to provide high quality for ATM traffic over satellite. COMSAT has built these error control mechanisms into its satellite ATM interface equipment named ATM link Accelerator (ALA) and ATM link Enhancement (ALE) [7]. In the ALA and ALE the adaptive Reed-Solomon codings and specialised cell-based interleaving algorithms are used for error control. These generate 0-8% overhead depending on the dynamically measured satellite link quality. The satellite could maintain BER below 1×10^{-8} in clear sky operation 96% of the time. The interleaving mechanism reduced the burst error effect of the satellite links.

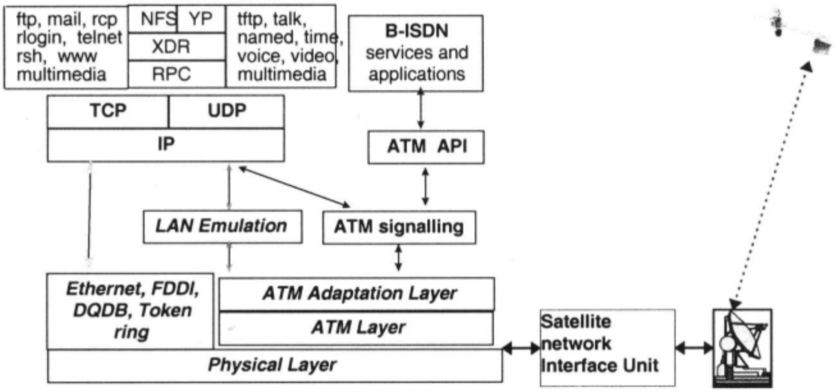

ftp, mail, rcp rlogin, telnet rsh, www multimedia
NFS
YP
XDR
RPC
tftp, talk, named, time, voice, video, multimedia
B-ISDN services and applications
TCP
UDP
IP
ATM API
LAN Emulation
ATM signalling
Ethernet, FDDI, DQDB, Token ring
ATM Adaptation Layer
ATM Layer
Physical Layer
Satellite network Interface Unit

Figure 14.4. Architecture of existing and ATM networks over the satellite system.

5.3 PROTOCOL STACK AND ARCHITECTURE

Currently most of the applications and services are based on the existing network protocols such as TCP/IP and UDP/IP. It is expected that in the future B-ISDN services will directly use the ATM Application Programming Interface (API) which has the advantage that the application can specify the required bandwidth and quality of services.

Figure 14.4 illustrates the relationship between the existing network architecture and the ATM network architecture used in the satellite ATM system. It shows how the services and applications of the existing network architecture can be transmitted transparently by ATM over satellite. There was a restriction in the current implementation of the demonstrator that it allowed only homogeneous network interconnections such as Ethernet to Ethernet, DQDB to DQDB, FDDI to FDDI and ATM to ATM connections. But it would be possible to have gateway function in the ground segment to interconnect heterogeneous networks.

5.4 PERFORMANCE ON THE SERVICES AND APPLICATIONS

The ATM performance parameters are related to the link bit error rate and are also dependent on the bit error distribution. In the case of random distribution of errors as in optical fibre links, the ATM header error correction (HEC) mechanism which is capable of correcting single-bit errors corrects most errors encountered. However in satellite links, the link coding mechanism used produces burst of errors. More than one error in the header can not be corrected by the ATM HEC.

The CLR is proportional to the BER and is higher than for links with random error distribution. The link coding is however necessary to reduce the error rate of payload data. To avoid the effect of burst error in the ATM QoS, improved coding techniques (such as Reed-Solomon code and interleaving) could be used to spread bit errors over the ATM cell headers.

Although satellites can not compete with optical fibre in total bandwidth available to applications, they still provide enough bandwidth for quite a few numbers of applications. Most of the protocols designed for data communication re-transmit the error or lost packets. Long delays made these protocols very inefficient. Therefore, the long delay due to the nature of satellite link had a significant impact on different aspects of the applications that included:

Throughput: Applications using connection oriented transport level protocols (such as the TCP/IP) needed to wait for the acknowledgements of packet arrival to support the flow control mechanism and to provide a reliable transport layer service. If a packet was lost, the protocol would re-transmit the packet. The throughput was restricted by the waiting for acknowledgement. The window size of the protocol can be used to adjust the amount of data to be sent before waiting for the acknowledgement. If connectionless protocols such as the UDP/IP were used, there was no re-transmission and no guarantee that the packet will arrive at its destination. Throughput can be estimated as:

$$\text{Throughput} = \text{WindowSize} / \text{RTT}$$

where WindowSize is the maximum number of data to be transferred before getting acknowledgements, and the RTT is Round trip time.

Request and response services: the long delay affects the throughput of the request and response type services (for example, the interactive service of login to a remote system). Users experience slow response time and slow information retrieval.

Video and voice services: real time services are more sensitive to the delay and waiting time for acknowledgement. As long as the delay variation is restricted to a very small value or the signal timing can be recovered at the destination, the satellite can still provide the connection at high quality. The extra delay for data waiting for a time slot in the TDMA frame can be up to a TDMA frame time (20 ms in the demonstration).

Text or data services: These are not sensitive to the delay and often require a reliable transport level protocol. The throughput is restricted and parameters of the protocols need to be adjusted or new protocol designed to suit the feature of satellite communications.

Buffer Requirement: Since the satellite ATM equipment interfaces to the high speed networks, it was important to allow these networks to transmit burst traffic at a high speed, to take the advantages of ATM technology. If the transmission link capacity was higher than the satellite link, buffers are required to absorb the traffic. Larger buffers resulted in increased delays. Traffic management is important to allow the satellite system to support high speed networks and limiting the probability of buffer overflow and extra buffering delay. The buffer requirement should take into account the maximum packet size and the differences between the network speed and the satellite link speed.

6. RESOURCE MANAGEMENT AND TRAFFIC CONTROL FOR THE SATELLITE ATM SYSTEM

6.1 RESOURCE MANAGEMENT

There are three levels of Resource Management (RM) mechanisms in the satellite system. The first level is controlled by the Network Control Centre (NCC) to allocate the bandwidth capacity to each earth station. The allocation is in the form of Burst Time Plans (BTPs). Within each BTP, burst times are specified for the earth station that limit the number of cells in bursts the earth stations can transmit. In the CATALYST demonstrator, the limit is that each BTP is less than or equal to 960 ATM cell and the sum of the total burst times is less than or equal to 1104 cells.

The second level is the management of the virtual paths (VPs) within each BTP. The bandwidth capacity which can be allocated to the VP is restricted by the BTP. The third level is the management of the virtual channels (VCs). It is subject to the available bandwidth resource of the VP. Figure 14.5 illustrates the resource management mechanisms of the bandwidth capacity. The allocation of the satellite bandwidth is done when the connections are established. Dynamic changing, allocation, sharing, or re-negotiation of the bandwidth during the connection is for further study.

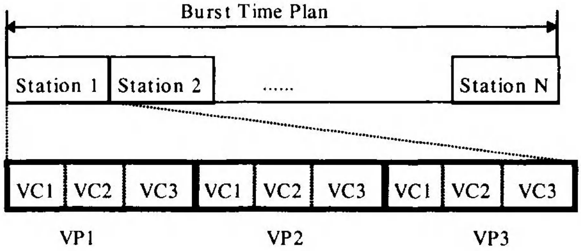


Figure 14.5. Satellite resource management.

To effectively implement resource management, the allocation of the satellite link bandwidth can be mapped into the virtual path (VP) architecture in the ATM networks and the each connection mapped into the virtual connection (VC) architecture. The BTP can be a continuous burst or a combination of a number of sub-burst times from the TDMA frame.

The burst time plan, data arrival rate and buffer size of the ground station had an important impact on the system performance. To avoid buffer overflow the system needs to control the traffic arrival rate, burst size, or allocation of the burst time plan. The maximum traffic rate allowed, to prevent the buffer overflow, is a function of the bursts time plan and burst size for a given buffer size, and the cell loss ratio is a function of traffic arrival rate and allocated burst time plan for a given buffer size.

6.2 TRAFFIC AND CONGESTION CONTROL

The demonstration system can efficiently cope with traffic flowing from the network with bit rates up to 20.352 Mbit/s (excluding the overhead of the ATM cells) and even higher bit rates in a short burst if traffic control mechanisms are used. The demonstrator did not use any traffic control functions apart from resource management, which can be used to allocate network resources to separate traffic according to service characteristics. Thus this section will describe methods by which the system performance can be improved

6.3 CONNECTION ADMISSION CONTROL (CAC)

The CAC is defined as the set of actions taken by the network at the call set up phase in order to establish if sufficient resources are available to establish the call through the whole network at its required quality of service (QoS) and maintaining the agreed QoS of existing calls. This applies as well to re-negotiation of connection parameters within a given call. In a B-ISDN environment, a call can require more than one connection for multimedia or multiparty services such as video-telephony or video-conference.

A connection may be required by an on-demand service, or by permanent or reserved services. The information about the traffic descriptor and QoS is required by the CAC mechanism to determine whether the connection can be accepted or not. The CAC in the satellite has to be the integrated part of the whole network CAC mechanisms.

6.4 USAGE PARAMETER CONTROL (UPC) AND NETWORK PARAMETER CONTROL (NPC)

UPC and NPC monitor and control traffic to protect the network (particularly the satellite link) and enforce the negotiated traffic contract during the call. The peak cell rate has to be controlled for all types of connections. Other traffic parameters may be subject to control such as average cell rate, burstiness and peak duration.

At cell level, cells are allowed to pass through the connection if they comply with the negotiated traffic contract. If violations are detected actions such as cell tagging or discarding is taken. At connection level, violations may lead to the connection being released.

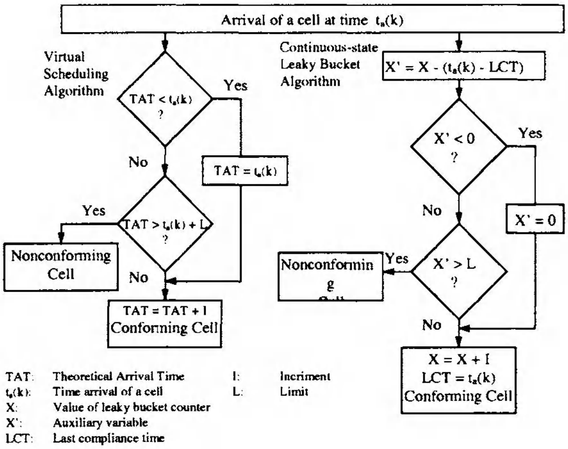

Arrival of a cell at time $t_a(k)$
Virtual Scheduling Algorithm
TAT < $t_a(k)$?
Yes
No
TAT = $t_a(k)$
Yes
TAT > $t_a(k)$ + L ?
Nonconforming Cell
No
TAT = TAT + I
Conforming Cell
Continuous-state Leaky Bucket Algorithm
X' = X - ($t_a(k)$ - LCT)
X' < 0 ?
Yes
No
X' = 0
Nonconformin g
Yes
X' > L ?
No
X = X + I
LCT = $t_a(k)$
Conforming Cell
TAT: Theoretical Arrival Time
$t_a(k)$: Time arrival of a cell
X: Value of leaky bucket counter
X': Auxiliary variable
LCT: Last compliance time
I: Incriment
L: Limit

Figure 14.6. Generic Cell Rate Algorithm (GCRA).

Figure 14.5 illustrates the Generic Cell Rate Algorithm (GCRA) is recommended as UPC/NPC in [8] [19]. The non-confirming cells violate the contract to be discarded or tagged for discarding when network becomes congested.

Apart from UPC/NPC tagging users may also generate different priority traffic flows by using the cell loss priority bit. This is called Priority Control (PC). Thus the traffic with low priority of a user may not be distinguished by a tagged cell, since both use the same CLP bit in the ATM header. Traffic shaping can also be implemented in the satellite equipment to achieve a desired modification of the traffic characteristics. For example, it can be used to reduce peak cell rate, limit burst length and reduce delay variation by suitably spacing cells in time.

6.5 REACTIVE CONGESTION CONTROL

Although preventive control tries to prevent congestion before it actually occurs the satellite system may experience congestion due to the earth station multiplexing buffer or switch output buffer overflow. In this case, where the network relies only on the UPC and no feedback information is exchanged between the network and the source, no action can be taken once congestion has occurred. Congestion is defined as the state where the network is not able to meet the negotiated QoS objectives for the connections already established. Congestion Control (CC) is the set of actions taken by the network to minimise the intensity, spread and duration of congestion.

Many applications, mainly handling data transfer, have the ability to reduce their sending rate if the network requires them to do so. Likewise, they may wish to increase their sending rate if there is extra bandwidth available within the network. These kinds of applications are supported by the ABR service class [19]. The bandwidth allocated for such applications is dependent on the congestion state of the network. Rate-based control was recommended for ABR services, where information about the state of the network is conveyed to the source through special control cells called Resource Management (RM) cells [8]. Rate information can be conveyed back to the source in two forms:

Binary Congestion Notification (BCN) using a single bit for marking the congested and not congested states. BCN is particularly attractive for satellites due to their broadcast capability. Explicit Rate (ER) indication, with which the network notifies the source of the exact bandwidth share it should be used to avoid congestion. The earth stations can determine congestion either by measuring the traffic arrival rate or by monitoring the buffer status.

7. FUTURE SATELLITE SYSTEMS

Until the launching of the first regenerative INTELSAT satellite in January 1991 and the ACTS satellite in September of 1993, all the satellites are transparent satellites. Though the regenerative, multibeam and on-board ATM switch satellites have potential advantages, they increased the complexity on reliability, the effect on flexibility of use, the ability to cope with unexpected changes in traffic demand (both volume and nature) and new operation procedure. So far, the ATM experiments and demonstrators have been based on the transparent and regenerative satellites, hence the research and development have been mainly on the ground segment. The on-board satellite ATM switches will be the new development of future satellite systems together with multibeam and LEO/MEO constellation.

7.1 THE ATM ON-BOARD SWITCH SATELLITE

There are potential advantages in performance and flexibility for the support of services by placing the processing and switching functions on board of the satellites, with respect to the use of a satellite with a transparent payload and routing functions. It is particularly important for satellite constellations with spot beam coverage and/or inter-satellite communications.

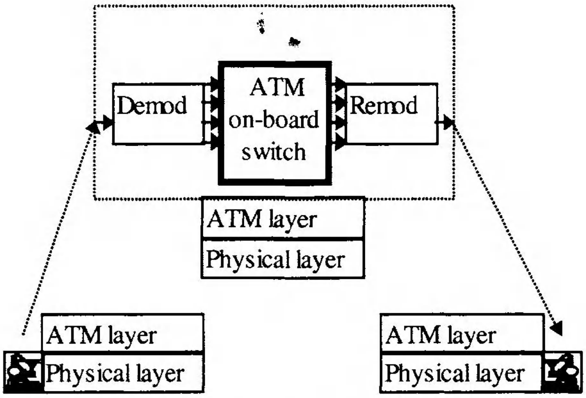


Figure 14.7. Model of the user plan for ATM on board switch satellite

In the case of ATM on-board switch satellite, the satellite acts as a switching point within the network (as illustrated by Figure 14.7) and is interconnected with more than two terrestrial network end-points. The on-board switch routes ATM cells according to the VPI/VCI of the header and the routing table when connections are set up. It also supports the signalling protocols used for UNI as access links and for NNI as transit links.

7.2 MULTIBEAM SATELLITE

A multibeam satellite features several antenna beams which provide coverage of different services zone as illustrated by Figure 14.8. As received on board satellite, the signals appear at the output of one or more receiving antennas. The signals at the repeater outputs must be fed to various transmitting antennas.

The spot beam satellites provide advantages to the earth station segment by its improving the figure of merit G/T on the satellite.

It is also possible to reuse the same frequency band several times in different spot beams to increase the total capacity of the network without increasing the allocated bandwidth. But there is interference between the beams.

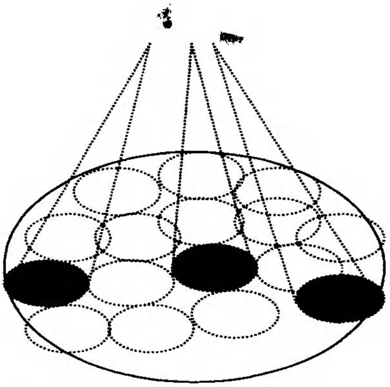


Figure 14.8. Multibeam satellite.

One of the current techniques for interconnections between coverage areas is on-board switching - satellite switched TDMA (SS/TDMA). It is also possible to have ATM switch on-board multibeam satellite.

7.3 LEO/MEO SATELLITE CONSTELLATIONS

One of the major disadvantages of GEO satellites is caused by the distance between the satellites and the earth stations. They have traditionally mainly been used to offer fixed telecommunication and broadcast services. In the recent years, satellite constellations of Low/Medium-Earth-Orbit (LEO/MEO) for global communication have been developed and will be in operations in year 2000s. The distance is greatly reduced. A typical MEO satellite constellation such as ICO has 10 satellites plus 2 spares, and LEO such as SKYBRIDGE 64 satellites plus spares.

Comparing to GEO network, LEO/MEO network is much complicated, but provide a lower end-to-end delay, less free space loss and higher overall capacity. However due to the relatively fast movement of the satellites in LEO/MEO orbit relative to the user, satellite handover is an important issue.

Constellations of LEO/MEO satellites can also be an efficient solution to offer highly interactive services with a very short round-trip propagation time over the space segment (typically 20/100 ms for LEO/MEO as compared to 500 ms for geostationary systems). The systems can offer similar performances to terrestrial networks, thus allowing the use of common communication protocols and applications and standards. Protocols such as TCP/IP are latency sensitive, which significantly reduces the throughput on geostationary systems.

7.4 USE OF HIGHER FREQUENCY SPECTRUM

Satellite constellations can use the Ku band (11/14 GHz) for connections between user terminals and gateways. High speed transit links between gateways will be established using either the Ku or the Ka band (20/30 GHz).

According to ITU radio regulation, geostationary networks have to be protected from any harmful interference from non geostationary systems. This protection is achieved through angular separation using a predetermined hand-over procedure based on the fact that the positions of geostationary and constellation satellites are permanently known and predictable. When the angle between a gateway, the LEO/MEO satellite in use by the gateway and the geostationary satellite is smaller than 10, the LEO/MEO transmissions are stopped and handed over to an other LEO/MEO satellite which is not in similar interference conditions.

The constellations are to provide a cost-effective solution to offer a global access to broadband services. The architectures are capable of supporting a large variety of services, reducing costs and technical risks related to the implementation of the system, ensuring a seamless compatibility and complementary with terrestrial networks, providing flexibility to accommodate service evolution with time as well as differences in service requirements across regions, and optimising the use of the frequency spectrum.

7.5 SUPPORTING INTERNET OVER SATELLITES

In the last few years, the Internet has had a very rapid expansion worldwide. A large number of new multimedia applications, such as real time video and audio communications and distributions, have been developed based on the Internet. These applications have variable requirements in terms of data rate and sensitivity to delay, several of them requiring a large bandwidth to function satisfactorily. The connection links for Internet network are being upgraded rapidly to support high data rates.

It becomes important for the satellites to support Internet protocols and services. A satellite architecture (including the existing GEO satellites and the MEO/LEO satellite constellations) which can support Internet Protocols will provide a good alternative to take advantage of the intrinsic capability of satellites for Internet connections and Internet access covering very wide areas and large population.

Future satellite network architectures will have to cope with a tremendous increase in connected end-systems, and with a large diversity in service types and quality-of-service.

While originally the Internet Protocols were conceived for the transfer of data, with the emergence of multimedia application, there is a sharp increase in the use of IP-based applications which present real-time (or near real-time) characteristics. Satellite based systems have significant impact on the use of these IP protocols to support multimedia applications such as streaming audio, streaming video, and audio-video conferencing. The new Internet Protocol, IPv6, has the potential to the support these applications. It is important to understand various classes of service provided by the new protocols, and specific issues including Quality of Service guarantees, scaleable routing, mobility and addressing.

The future satellite systems can allow services such as high-speed network access and interconnection take place anywhere in the world. They can support broadband asymmetrical connections from terminals to the fixed network, for example, up to 60 Mbit/s from the network to the users with increments of 16 Kbit/s and up to 2 Mbit/s in the return link. This can be optimised for Internet communications which are characterised by random bursts of asymmetrical data transmission. In addition, the small size of the increments will provide the user with bandwidth on demand. The highly interactive applications and services include: high-speed access to Internet, on-line services, telecommuting, electronic mail, file transfers, video conference, telemedecine, video on demand and electronic games.

8. SUMMARY

This paper presented an introduction to “satellite ATM networks” covering a range of topics. It is based on mainly on the results of European projects focusing on the GEO satellite systems and also taking into account the future LEO/MEO satellite constellations and new applications.

Many experiments have been carried out using GEO satellite systems to demonstrate of how satellite ATM systems can be developed based on the existing satellites to support data, voice, video and multimedia communications. They have some limitations, but still provide useful results about the behaviour of the applications over satellite connections. They also provide good experience and some important reference values for future development of satellite systems for broadband communications. The limitations are mainly due to the characteristics and nature of the GEO satellites.

Recently, the requirement to support integrated services and smaller user terminals with mobility will result in ATM switches being deployed on-board satellite. Furthermore there is trend towards lower orbits such as MEO/LEO constellations to achieve lower delays and lower power requirements for mobile terminals. Thus the changing has started from where satellites are used to interconnect a small number of earth stations to where satellites will be used as access to the B-ISDN by a large number of small, portable and/or mobile terminals.

Therefore, satellites will in the near future to provide a practical and economical alternative for interconnections of the broadband networks and for remote user access to broadband services. The will complement the terrestrial networks and provide mobile and broadband services as an integrated part of the broadband communication infrastructure.

New advanced satellite systems, particularly the new LEO/MEO constellation systems for their international coverage and applications, will bring high-speed multimedia services to business and residential users at significantly lower costs than existing systems. These systems will offer voice, data, video, imaging, video-teleconferencing, interactive video, TV broadcast, multimedia, global Internet, messaging, and trunking services. As more and more new commercial actives, applications and services are developed on the Internet, we will a significant development and research in Internet over satellite or IP over satellite.

References

[1] Cuthbert, L.G. and J.C. Sapanel, "ATM: The Broadband Communication Solution", Institution of Electrical Engineers, 1993.

[2] Evans, E.G. and R. Tafazolli., "Future multimedia communications via satellite", 2nd Ka band utilisation conference, Florence-Italy, 24-26 September, 1996.

[3] Jacobsen, V., R. Braden and D. Borman, "TCP Extensions for High Performance", RFC1323, 1992.

[4] Louvet, B. and S. Chellingsworth, "Satellite integration into broadband networks", Electrical Communications, 3rd Quarter, 1994.

[5] Luckenbach, T.., R. Ruppelt., and F. Schulz, "Performance Experiments within Local ATM Networks", Twelfth Annual Conference on European Fibre Optic Communication and Networks, Heidelberg-Germany, June 1994.

[6] Maral, G. and M. Bousquet, "Satellite Communications Systems - System Techniques and Technology", 2nd ed., John Wiley, 1993.

[7] Miller, S P and D. M. Chitre, "COMSAT's ATM Satellite Services", IEE Colloquium on "ATM over satellite" organised by Professional Group E9 (Satellite systems and applications), LONDON, 27 November 1996.

[8] ATM Forum, "Traffic Management Specification, version 4.0", Document Number: af-tm-0056.00, April 1996.

[9] ATM Forum, "Work Items for Wireless ATM Access over Geosynchronous Satellite Links", Document Number: ATM_Forum/96-1109, 1996.

[10] ATM Forum, "Satellite ATM Utilisation", Document Number: ATM_Forum/96-1109, 1996.

[11] ATM Forum, “Satellite Access Service Descriptions”, Document Number: ATM_Forum/96-1452, 1996.

[12] ATM Forum, “Extensions to proposed charter, scope, and work plan for WATM working group”, Document Number: ATM_Forum/96-0672, 1996.

[13] ATM-Forum, “ATM User-Network Interface Specification”, Version 3.1, September 1994.

[14] CFS, “Satellite in B-ISDN: General Aspects”, RACE Common Functional Specification and Common Practice Recommendations, Issue D, D751,1993,.

[15] ITU-T, “B-ISDN General Network Aspects”, ITU-T Rec. I.311, March 1993.

[16] ITU-T, “B-ISDN Service Aspects”, ITU-T Rec. I.211, March 1993.

[17] ITU-T, “B-ISDN ATM Functional Characteristics”, ITU-T Rec. I.150, November 1995.

[18] ITU-T, “B-ISDN ATM Layer Specification”, ITU-T Rec. I.361, November 1995.

[19] ITU-T, “Traffic Control and Congestion Control in B-ISDN”, ITU-T Rec. I.371, May 1996.

[20] RACE Common Functional Specifications D751, "Satellites in the B-ISDN, General Aspects", Issue D, December 1993.

[21] Sun, Z., T. Ors and B.G. Evans, “Interconnection of Broadband Islands via Satellite - Experiments on the RACE II CATALYST Project”, Transport Protocols for High-Speed Broadband Networks Workshop at Globecom’96, London, November 1996.

[22] Sun, Z., T.Ors and B.G.Evans, “ATM-over-satellite demonstration of broadband network interconnection”, Computer Communications, Special Issue on Transport protocols for high speed broadband Networks, Volume 21 number 12 25 August 1998, page 1091-1101.

Chapter 15

PERFORMANCE MODELING AND NETWORK MANAGEMENT FOR SELF-SIMILAR TRAFFIC

Gilberto Mayor
McKinsey & Company, Inc.
Sao Paulo, Sao Paulo, Brazil 04717-004
Gilberto_Mayor@MCKINSEY.COM

John Silvester
Department of Electrical Engineering-Systems
University of Southern California
Los Angeles, California, USA
90089-2562
silveste@usc.edu

Abstract Since the discovery of the self-similar nature of network traffic, researchers were able to propose new traffic models [Mayor96d, Norros94] that are better able to mimic the long-range dependence phenomenon exhibited by real network traffic. Nevertheless, since most of the existing queueing theory is based on the assumption of Markovian models, there are few analytical results dealing with an ATM queueing system driven by a self-similar process [Addie95b, Duffield95, Likhanov95, Mayor96d, Parulekar96, Ryu96a]. In this work, we give an overview of traffic models and analytical tools capable of computing tail probabilities of an ATM queueing system driven by a self-similar process. We also explain the meaning of long-range dependence and its impact on network performance and network management protocols, by revisiting Mandelbrot's work[Mandelbrot69]. We propose a traffic characterization based on a fractional Brownian motion envelope process. By using this characterization, we show a framework derived in [Mayor96d] capable of computing bandwidth and buffer requirements in ATM networks driven by aggregate, heterogeneous, self-similar processes.

Keywords: Self-similar, ATM envelope process and fractional brownian motion

1. INTRODUCTION

The discovery of the self-similar nature of network traffic [Leland94] greatly improved our understanding of network performance by i) describing real network traffic's behavior throughout different time-scales, ii) showing the importance of choosing the right time-scale when designing traffic models, which ultimately lead to the development of more accurate network queueing systems. First, we give a brief overview of self-similar processes' main properties. We revisit Mandlbrot's work [Mandelbrot69] in order to explain the impact of long-range dependence on traffic behavior on queueing performance. We also give an overview of traffic models and analytical tools [Huebner, Mayor96d, Narayan, Norros94] dealing with queueing systems driven by self-similar processes. We propose a traffic characterization based on a fractional Brownian motion envelope process. By using this characterization, we show a framework derived in [Mayor96d] capable of computing bandwidth and buffer requirements in ATM networks driven by aggregate, heterogeneous, self-similar processes.

In section 2, we give an overview of self-similar processes and heavy-tailed On-Off sources. In section 3, we quantify the impact of long-range dependence (LRD) on queueing performance. In section 4, we use a fractional Brownian motion envelope process to compute tail probabilities. In section 5, we quantify the statistical multiplexing gain of a queueing system driven by LRD sources. In section 6, we discuss the impact of LRD on ATM flow control and congestion detection protocols.

2. SELF-SIMILAR PROCESSES: BASIC DEFINITIONS

A self-similar process is invariant in distribution under scaling of time [Samorodnitsky94]. Intuitively, if we look at several pictures of a self-similar process at different time-scales they will all look *similar*. The real valued-process $X(t), t \epsilon T$ is *self-similar* with Hurst parameter $H > 0$ if for all $a > 0$, $X(at) \doteq a^H X(t)$. This definition says that for any sequence of time points t_1, ..., t_k and any positive constants a^H, $a^H (X(ct_1), X(ct_2), ..., X(ct_k))$ has the same distribution as $((X(t_1), X(t_2))..., X(t_k))$. Therefore, typical sample paths of a self-similar process look qualitatively the same (similar), irrespectively to the distance from which we look at them[Beran94].

Following [Leland94], we also define a second-order self-similar process. Let $X = (X_t : t = 0, 1, 2, ...)$ be a covariance stationary stochastic process with mean μ, finite variance σ^2, and autocorrelation function $r(k), k \geq 0$. We assume that X has the autocorrelation function

$$r(k) \approx k^{-\beta}, k \to \infty \tag{15.1}$$

where $0 < \beta < 1$ is given by $H = 1 - \beta/2$. Let $X_k^{(m)}$ be the new Covariance stationary process with autocorrelation function $r^{(m)}$ obtained by averaging the original process X over non-overlapping blocks of size m.

$$X_k{}^{(m)} = 1/m(X_{km-m+1} + ... + X_{km}), k \geq 1$$

X is called *second-order self-similar* with self-similarity parameter $H = 1 - \beta/2$ if for all m = 1,2,..., these properties apply:

1. $var(X^{(m)}) = \sigma^2 m^{-\beta}$
2. $r^{(m)}(k) = r(k), k \geq 0.$

X is called *asymptotically second-order self-similar* with self-similarity parameter $H = 1 - \beta/2$ if for all k large enough

1. $\lim_{m \to \infty} var(X^{(m)}) = \sigma^2 m^{-\beta}$
2. $r^{(m)}(k) = r(k), k \geq 0$

The first important characteristic manifested by the LAN traces, and identified by Bellcore researchers, is called **Long-Range Dependence** (LRD)[Beran94]. Mathematically, LRD implies that the autocorrelation function of the process decays hyperbolically fast, *i.e.*, the same behavior predicted by equation (15.1). In this case, $1/2 < H < 1$ implying a non-summable autocorrelation function, *i.e.* $\sum_k r(k) = \infty$. In the frequency domain, LRD implies that the spectral density obeys a power-law behavior near the origin. On the other hand, traditional Markov models exhibit **Short-Range Dependence** (SRD), i.e., the autocorrelation function decays exponentially fast

$$r(k) \approx a^{|k|}, k \to \infty$$

In this case, $0 < H < 1/2$, implying a summable autocorrelation function, *i.e.*, $\sum_k r(k) < \infty$. For $H = 1/2$, we have the case of uncorrelated arrivals. Since traditional Markovian models are SRD processes, they usually underestimate the dependence among packet arrivals over long periods of time. Even though it is hard to show that network traffic is a self-similar process, it is relatively simple to show that several types of network traffic exit LRD over the time scales of interest. We need only to analyze its autocorrelation structure in order to verify if it behaves like an LRD process. Moreover, recent studies showed that LRD might have a pervasive impact on queueing performance. Therefore, it is our view that a self-similar process is indeed a very accurate model for network traffic, since it can mimic the long-range dependent behavior exhibited by the real traffic.

The second phenomenon exhibited by a self-similar process is called the **Slowly Decaying Variance.** In this case, the variance of the sample mean decays more slowly than the reciprocal of the sample size:

$$var(X^{(m)}) \approx a_1 m^{-\beta}$$

as $m \to \infty$, $H = 1 - \beta/2$, with a_1 being a positive constant. This result also differs from traditional Markovian models where the variance of the sample mean is given by

$$var(X^{(m)}) \approx a_1 m^{-1}$$

This mathematical result matches our knowledge that network traffic usually has a very large variability. In fact, Mandelbrot [Mandelbrot69] proposed the *Infinite Variance Hypothesis* (IFV) in order to account for the erratic variability of the sample variances without giving up stationarity. Intuitively, instead of assuming that network traffic is not stationary, the self-similar hypothesis allows the assumption that it has *infinite* variance.

2.1 HEAVY-TAILED ON-OFF SOURCES

Although, there is still no clear explanation for the self-similar nature of network traffic, Bellcore researchers claim that it derives from the aggregation of heavy-tailed (HT) On-Off sources. Based on an early theorem derived by Mandelbrot and on empirical results, they claim that individual sources can be modeled by Heavy-Tailed (HT) On-Off sources so that the aggregate traffic converges to a self-similar process. We give an overview of HT On-Off sources here.

Willinger *et al.* [Willinger95] investigated Bellcore's LAN traces. These trace were shown to be self-similar [Leland94] and are publicly available at ftp.bellcore.com. Whenever we refer to the LAN traces throughout this work, we are addressing these specific traces. They concentrated on the traffic generated by individual source-destination pairs instead of looking at the aggregate traffic. They concluded that individual sources can be seen as HT On-Off sources. In this case, the sojourn time spent in each state, defined by U, is not exponentially distributed, but rather has hyperbolic tail distribution satisfying

$$P(U > u) \sim cu^{-\alpha}. \tag{15.2}$$

as $u \to \infty$, for $1 < \alpha < 2$ where c is a positive constant. For example, U can have the Pareto distribution [Jain9lb]

$$F_U(u) = 1 - u^{-\alpha}.$$

Previously, Mandelbrot have shown that a sum of heavy-tailed renewal reward processes can converge to a self-similar process. Taqqu extended Mandelbrot's work and established several theorems regarding the limit sum of renewal-reward processes with infinite variance [Taqqu86]. More recently Willinger and Taqqu, revisited this previous work in order to show that a sum of HT On-Off sources can converge to a fractional Brownian motion (fBm) process [Willinger95]. We can summarize their claim by saying that:

1. Individual traffic sources can be modeled as heavy-tailed On-Off sources.

2. The aggregate traffic resulting from the superposition of those HT sources converges to an fBm process.

Therefore, we conclude by claiming that an fBm process is a natural candidate for modeling network traffic since i) it can accurately replicate the long-range dependent behavior of real network traffic and ii) it is parsimonious, *i.e.,* it only requires three parameters to fully define the model.

3. UNDERSTANDING THE IMPACT OF LRD ON QUEUEING PERFORMANCE

The fBm process was introduced by Mandelbrot in [Mandelbrot68]. It is extensively used in both simulation and analytical studies of ATM queueing systems driven by self-similar traffic. There are several algorithms for generating an fBm synthetic trace [Chi73, Hosking84, McLeod78, Mandelbrot71]. More recently, new methods have been developed. For example, Huang [Huang95a, Huang95b] proposed a simulation method based on importance sampling, Pruthi [Pruthi95] used nonlinear chaotic maps and Lau *et al.* [Lau95] used a random midpoint displacement algorithm.

The ordinary Brownian motion, *B(t)*, describes the movement of a particle in a liquid subjected to collisions and other forces. It is a real random function with independent Gaussian increments such that

$$E[B(t+s)-B(t)]=0$$

$$Var[B(t+s)-B(t)]=\sigma^2|s|.$$

Mandelbrot [Mandelbrot68] defines the fBm process as being the moving average of *dB(t)* in which past increments of *B(t)* are weighted by the kernel $(t-s)^{h-1/2}$. Let H be such that $0 < H < 1$. The fBm is defined as the Weyl's fractional integral of B(t)

$$B_H(t)=\frac{1}{\Gamma(H+1/2)}\int_{-\infty}^{0}((t-s)^{H-1/2}-(-s)^{H-1/2})dB(s)+ \int_0^t (t-s)^{H-1/2}dB(s)$$

This equation leads to the ordinary Brownian motion if $H = 1/2$. Its self-similar property is based on the fact that $B_H(\rho s)$ is identical in distribution to $\rho^H * B_H(s)$. The increments of the fBm, Y_j, form a stationary sequence called fractional Brownian noise (fBn).

$$Y_j=B_H(j+1)-B_H(j), j=\cdots,-1,0,1,\cdots$$

By using large deviation theory, we revisit Mandelbrot's work in order explain the behavior of an ATM queueing system driven by LRD traffic [Mandelbrot69]. Mandelbrot studied long-range dependence in Economic time series. He explains this phenomenon as a tendency for large values to be followed by large values, in such a way that those time series seem to go through a succession of *cycles* whose wavelength is of the order of the magnitude of the total sample size. It implies that i) traffic sources exhibiting LRD can sustain *high* transmission rates for very long intervals (strong low frequency component) leading to unexpected cell losses and ii) it is not possible to define a maximum burst size for those sources leading to the buffer inefficacy phenomenon.

3.1 QUANTIFYING LRD

An ATM node can be modeled as a single-server queueing system, with deterministic service rate given by *c*. The arrival traffic is defined by the process $A_H(t)$ with mean $\bar{a}$ and variance σ^2. We can quantify the LRD phenomenon by investigating how long the source is likely to transmit at *high* rates, *i.e.*, at rates substantially higher than its average arrival rate. In [Norros94], Norros introduced a new model for fBm arrival processes. We use this model to quantify the impact of LRD on queueing performance. Following his work, we assume that the arrival process $A_H(t)$ is a fBm process given by

$$A_H(t) = \bar{a}t + \sqrt{\bar{a}v}Z(t) \tag{15.3}$$

where $\bar{a} > 0$ is the mean input rate, $v > 0$ is the coefficient of variation, $H \in \lfloor\frac{1}{2}, 1)$ is the self-similar (Hurst) parameter and *Z(t)* is a normalized fBm. We investigate the probability that the *instantaneous* average arrival rate, defined as $\frac{A_H(t)}{t}$, exceeds k times its mean rate at time t:

$$P(\frac{A_H(t)}{t} > k\bar{a}) = P(\bar{a}t + \sqrt{\bar{a}v}Z(t) > k\bar{a}t) = P(Z(t) > \frac{t(k\bar{a} - \bar{a})}{\sqrt{\bar{a}v}}).$$

By the self-similarity property $Z(t) = t^H Z(1)$, we have

$$P(Z(1) > \frac{t\bar{a}(k-1)}{\sqrt{\bar{a}v}t^H}) = \overline{\Phi}(\frac{\bar{a}(k-1)t^{1-H}}{\sqrt{\bar{a}v}})$$

where $\overline{\Phi}(y) = P(Z(1) > y)$ is the residual distribution function of the standard Gaussian distribution. In fact, using the approximation given by the Weibull distribution [Norros94]

$$\overline{\Phi}(y) \approx (2\Pi)^{-1/2}(1+y)^{-1}exp(-y^2/2) \approx exp(-y^2/2) \tag{15.4}$$

we obtain

$$P(\frac{A_H(t)}{t} > k\bar{a}) \approx exp(-\frac{\bar{a}^2(k-1)^2t^{2-2H}}{2\bar{a}v}) \tag{15.5}$$

Equation (15.5) shows that the probability that the instantaneous average arrival rate of fbm exceeds its mean rate, decays exponentially fast with t when $H = 1/2$. For a LRD process, *e.g.*, if $H = 0.9$, this probability can decrease very slowly with t We compute the average arrival rate $\bar{a}$ and variance σ^2 parameters of Bellcore's LAN trace (pAug.TL) We substitute them in equation (15.5); Figure 15.1 shows the result.

The upper and lower dashed curves correspond to the probability that the Brownian motion's instantaneous average rate is $2.0\bar{a}$ and $3.0\bar{a}$ at time t, respectively. The upper and lower solid curves correspond to the probability that the average arrival rate is $2.0\bar{a}$ and $3.0\bar{a}$ at time t for an fBm with $H = 0.9$, respectively. We conclude that because a LRD source can transmit at high rates for very long periods of time, it might not be possible to avoid cell losses by just allocating a large buffer. In other words, a queueing system driven by an LRD source suffers from the buffer inefficacy phenomenon.

Figure 15.1 $P(\frac{A_H(t)}{t} > k\bar{a})$ for H=0.5 (dashed curve) and H=0.9 (solid curve).

3.2 DEFINING UTILIZATION

By the Strong Law of Large Numbers we know that $\frac{A_H(t)}{t}$ [1] converges to its mean $\bar{a}$ when $t \to \infty$. For LRD processes, this convergence can be *very slow* [Garret94]. Therefore, we show the rate of convergence of both a Poisson and an fBm process. Let $A(t)$ denote the cumulative number of cell arrivals at time t for a given arrival process with average arrival rate $\bar{a}$ cells per unit of time. Figure 15.2 shows the rate of convergence for an ordinary Poisson process's sample path. The three dotted curves correspond to three non-overlapping sample-paths of the normalized average rate, *i.e.*, $\frac{A(\tau)}{\bar{a}\tau}$, for this Poisson process. We also define the worst-case sample path within a trace, *i.e.*, the optimal envelope process, given by $Y(\tau) = max_{\tau>0}(A(t+\tau) - A(t))$. Intuitively, $Y(\tau)$ defines the *maximum* number of cell arrivals within an interval of size T. The solid curve corresponds to $\frac{Y(\tau)}{\bar{a}\tau}$, for a 1,000,000 points sample. We can see that even the worst-case sample path converges to the average arrival rate relatively fast, *i.e.*, within a *short* period of time.

Figure 15.3 shows the rate of convergence for the Bellcore LAN trace. Contrary to the case of Poisson arrivals, the worst-case sample path converges very slowly to its mean arrival rate. This phenomenon limits the maximum possible link utilization, since the average arrival rate is significantly higher than $\bar{a}$ for very long periods of time. In fact, real network traffic is not stationary, therefore it is not adequate to define a long-term utilization. Therefore, by using a self-similar model, we

[1] We assume that $A_H(t)$ is a covariance stationary process

Figure 15.2 Sample paths of the normalized average rate for a Poisson process.

Figure 15.3 Sample paths of the normalized instantaneous average rate for the LAN traffic.

Figure 15.4 Instantaneous utilization measured over 10,000 time-slots.

attempt to account for the large variability of the traffic without giving up stationarity [Mandelbrot69]. In this case, even though the long-term link utilization is low, the instantaneous utilization can be relatively high for very long periods of time. For example, assume that the link capacity c is given by $2\bar{a}$. We computed the instantaneous utilization, defined as $\frac{A(\tau)}{\tau c}$, for the LAN traffic over non-overlapping, consecutive, periods $\tau = 10,000$ time-slots. Figure 15.4 shows that in some intervals the link utilization achieves almost 80% even though the long-term utilization is only 50%. It shows that a LRD process can sustain utilizations as high as 80% for a long period of time. On the other hand, in a traditional queueing system driven by a SRD process, the utilization achieves high peak values only during small time intervals, *i.e.*, the input traffic is not able to sustain a high utilization rate for a very long period of time.

3.3 COMPUTING THE MAXIMUM BUSY PERIOD

A direct consequence of LRD is the presence of very long busy periods, possibly causing massive cell losses. In fact, we showed [Mayor96b] that in an ATM queueing system with LRD traffic, at low utilization, the cell losses are concentrated at the tail of the busy period. Moreover, the busy period is an upper bound for the maximum delay that a cell can occur in an ATM queueing system. Therefore, we compute a probabilistic bound for the maximum busy period of an ATM queueing system driven by an fBm process. We compare it to the busy period of a system with Brownian motion arrivals.

By using large deviation theory, we extend Chang's work [Chang94] in order to compute a probabilistic bound $\hat{d}$ for the busy period of a stochastic queueing system

$$\hat{d} \overset{\text{def}}{=} inf\{t \geq 1 : P(A_H(t) > ct) < \epsilon\}$$

where $\epsilon << 1$. Therefore, *the busy period will not exceed* $\hat{d}$ *with probability* $(1 - \epsilon)$. By following the same approach as in the previous section, we can write

$$P(A_H(t) > ct) = \overline{\Phi}(\frac{(c - \bar{a})}{\sqrt{\bar{a}v}t^{H-1}}).$$

Figure 15.5 The busy period's bound when H=0.50 (dotted curve) and H=0.90 (solid curve).

Therefore,

$$\hat{d} = inf\{t > 0 : \overline{\Phi}(\frac{(c - \bar{a})}{\sqrt{\bar{a}v}t^{H-1}}) \leq \epsilon\}$$

where $\epsilon << 1$. Using the approximation given by equation (15.7), we can write

$$\hat{d} \approx (\frac{\sqrt{(-2\log\epsilon)}\sqrt{\bar{a}v}}{(c - \bar{a})})^{\frac{1}{1-H}} = (\frac{k\sigma}{(c - \bar{a})})^{\frac{1}{1-H}} = B^{\frac{1}{1-H}} \qquad (15.6)$$

where B is given by $(\frac{\sqrt{(-2\log\epsilon)}\sqrt{\bar{a}v}}{(c-\bar{a})})$. For the case of LAN traffic, Bellcore researchers observed H to be as large as 0.9. Therefore, the dependence on H exhibited by equation (15.6) shows that the busy period of the LRD system can be several orders of magnitude larger than the case of Brownian motion arrivals. For example, for $H = 1/2$ and $H = 0.90$, $\hat{d}_H$ is given by B^2 and B^{10} respectively.

3.3.1 Example . We substitute the parameters for the LAN traffic in equation (15.6) and compare it to the case given by $H = 1/2$. Figure 15.5 shows the results. The dotted curve corresponds to the case of a Brownian motion process, *i.e.*, H=1/2. In this case, the maximum busy period is relatively *small* even if the link capacity is close to the average arrival rate. The solid curve shows the busy period bound when H=0.90. In this case, since the process exhibits LRD, the maximum busy period can be extremely large (> 100,000 time-slots) if the link rate is close to the mean arrival rate.

We conclude that ATM links can be either *busy* or *idle* for very long periods. In this case, it is necessary to allocate bandwidth dynamically in order to maximize link utilization and avoid congestion. A possible solution for this problem of *non-homogeneous link utilization,* is to dynamically change the bandwidth allocated for a given Virtual Path (VP) based on its current utilization level as suggested in [Lin96].

4. A FRACTIONAL BROWNIAN MOTION ENVELOPE PROCESS

In this section, we introduce a traffic model based on an fBm probabilistic envelope process [MayorQGd]. We show that it closely matches the behavior of real network traffic. We believe that this characterization can be widely used to model several types of input traffic, including LRD and SRD sources. Moreover, because of its simplicity, it leads to an elegant framework capable of computing tail probabilities of ATM queueing systems. Furthermore, we show that this model can be used to predict the behavior of a queueing system driven by LRD traffic accurately, with minimal computational complexity.

It is well known that for a Brownian motion (Bm) process $A(t)$ with mean $\bar{a}$ and variance σ^2, the envelope process $\hat{A}(t)$ can be defined by

$$\hat{A}(t) \overset{\text{def}}{=} \bar{a}t + k\sqrt{\sigma^2 t} = \bar{a}t + k\sigma t^{\frac{1}{2}}$$

The parameter k determines the probability that *A(t)* will exceed $\hat{A}(t)$ at time *t.* Since A(t) is a Brownian motion process we can write

$$P(\frac{A(t) - \bar{a}t}{\sigma t^H} > k) = \overline{\Phi}(k)$$

where $\overline{\Phi}(y)$ is the residual distribution function of the standard Gaussian distribution. Using the approximation

$$\overline{\Phi}(y) \approx (2\Pi)^{-1/2}(1+y)^{-1}exp(-y^2/2) \approx exp(-y^2/2) \qquad (15.7)$$

we find k such that

$$\overline{\Phi}(k) \leq \epsilon$$

Therefore, k is given by

$$k = \sqrt{-2\log\epsilon}$$

We claim that $P(A(t) > \hat{A}(t)) \approx \epsilon$, where $k = \sqrt{-2\log\epsilon}$. This approach can be extended to deal with LRD traffic. Let $A_H(t)$ be a fBm process with mean $\bar{a}$. Hurst's law states that the variance of the increment of this process is given by $Var[A_H(t+s) - A_H(t)] = \sigma^2 s^{2H}$ where $H \in [\frac{1}{2}, 1)$ is the Hurst parameter. Therefore, we can also define a fBm envelope process by

$$\hat{A}_H(t) \overset{\text{def}}{=} \bar{a}t + k\sqrt{\sigma^2 t^{2H}} = \bar{a}t + k\sigma t^H \qquad (15.8)$$

The Bm envelope process is just the special case of $H = 1/2$. Similarly, k determines the probability that $A_H(t)$ will exceed $\hat{A}_H(t)$. However, *Figure 15.6* $Y(\tau)$ (middle curve) and fBm envelope processes for H = 0.50 (lower curve) and H = 0.83 (upper curve).

since the process exhibits LRD, if $A(t)$ exceeds $\hat{A}(t)$ at time t, it is possible that it will stay *above* it for a long period of time.

We should note that the source does not necessarily need to be self-similar in order to match this characterization, as long as it matches the behavior of the envelope process over the time-scale of interest. We investigate the accuracy of the fBm envelope process representation by inspecting how well it can model the worst-case behavior of real network traffic. Assume that the input traffic is characterized by a trace with N sample points, defined by $A(t)$, where $A(t)$ represents the cumulative number of cell arrivals up to time $t, t \in [1, 2, ..., N]$. We propose a very simple method for computing the fBm envelope process's parameters for this trace, by computing the trace's optimal envelope process. The advantage of this approach is that we do not need to accurately estimate the trace's Hurst parameter. The optimal envelope process (the worst-case sample path) for this trace is defined by $Y(t — s) = max_{s<t}(A(t) - A(s))$. We assume that the process is stationary so that $Y(\tau), \tau = t - s$ defines the maximum number of cell arrivals in an interval of size T. Therefore, we can choose the fBm envelope process's parameters $\hat{A}_H(.)$ so that it matches the behavior of $Y(.)$. We compare the envelope process representation to Bellcore's LAN trace. We compute the sample average arrival rate and sample variance for this trace and substitute for $\bar{a}$ and σ^2 in equation (15.8). We compute the optimal envelope process, *i.e,* $Y(.)$, and choose H so that $\hat{A}_H(.)$ matches the behavior of $Y(.)$. In Figure 15.6, the upper curve corresponds to the fBm envelope process with $\epsilon = 10^{-3}$. The lower curve represents the Brownian motion envelope process with the same ϵ. The middle curve corresponds to $Y(\tau)$. We can see that the fBm envelope process matches closely the behavior of the LRD trace. Moreover, we also note that the ordinary Brownian motion envelope process is unable to bound the behavior of the LRD source even if we choose ϵ large.

This representation of the input traffic has several major advantages:

- It is parsimonious, *i.e.*, only three parameters are required in order to completely characterize a source.
- It can represent SRD and LRD, *i.e.,* the source does not necessarily need to be LRD. We need only to choose the parameters for the fBm envelope process so that it matches the source's optimal envelope process over the appropriate time-scale.

- It provides very accurate delay bounds with minimal computational complexity [Mayor96d].

Nevertheless, we should note that it is extremely hard to estimate H accurately.

4.1 COMPUTING THE MAXIMUM BUSY PERIOD BY USING ENVELOPE PROCESSES

We use fBm envelope processes to compute a bound for the busy period and compare it to the previous result. Assume a deterministic service queueing system with service rate given by c. If $A_H(t)$ is an envelope process, Chang [Chang94] showed that the length of each busy period is bounded above by a constant d. Let

$$d \overset{\text{def}}{=} inf\{t \geq 1 : A_H(t) - ct \leq 0\} \tag{15.9}$$

Therefore, we compute $\hat{d}_H$ by substituting $\hat{A}_H(t)$ into equation (15.9). In this case, $\hat{A}_H(t) = ct$ so that we have

$$\overline{a}t + k\sigma t^H = ct$$

Therefore, $\hat{d}_H$ is given by

$$\hat{d}_H = (\frac{\sqrt{(-2\log\epsilon)}\sqrt{\overline{a}v}}{(c-\overline{a})})^{\frac{1}{1-H}}$$

We arrive at the same equation (15.6) for the busy period, computed in the previous section! It shows that the envelope process characterization leads to an accurate solution with little computational effort.

4.2 COMPUTING TAIL PROBABILITIES OF AN ATM QUEUEING SYSTEM

We derive a probabilistic bound for the maximum number of cells in an ATM queueing system by using a probabilistic envelope process. This approach has two main advantages over the traditional method of using large deviation theory: i) it can be used to compute end-to-end delay and ii) it handles multiplexing of heterogeneous sources. We show that the delay bound agrees with delay experienced by real network traffic. We believe that this framework is general and accurate so that it can be used in real ATM networks in order to provide QoS requirements.

Previous delay bounds are usually divided into two categories:

1. Strict delay bounds based on the output flow of a traffic regulator.
2. Probabilistic delay bounds computed by large deviation theory.

The first category includes the works of Cruz [Cruz91a] and Parekh [Parekh93]. Their main attraction derives from their simplicity so that they can be used in real networks since they can be computed in real-time. Their main disadvantages are that i) they give a *strict* delay bound, *i.e.*, it is not a probabilistic bound and ii) they are based on the output flow of an *inaccurate* traffic enforcement mechanism. Because of these reasons, those delay bounds are usually loose bounds so that it is impractical to use them in real networks. The second category includes the work of Chang [Chang94], Norros [Norros94], *etc.* Their main advantage comes from their probabilistic approach. They compute a probabilistic upper-bound delay so that the frequency that the bound is violated depends on a QoS parameter chosen by the application. Their main drawback is that they are usually computationally complex so that they cannot be derived in real-time. Moreover, they are not capable of providing end-to-end statistics or handling multiplexing of heterogeneous sources. Therefore, the main advantage of our delay bound computation over all previous works is that it combines the simplicity of the first delay bounds category with the desirable probabilistic approach of the second one.

Consider a continuous-time, work conserving ATM queueing system, with deterministic service rate given by c. Following Norros' work [Norros94], let $V(t)$ be the stationary stochastic process given by

$$V(t) = sup_{s \leq t}(A_H(t) - A_H(s) - c(t-s)), 0 \leq s \leq t \qquad (15.10)$$

Let $A_H(t)$ be the stationary stochastic input process defined as

$$A_H(t) = \bar{a}t + \sigma Z(t), t \geq 0$$

where $Z(t)$ is a normalized fBm process with Hurst parameter H. Equation (15.10) describes the amount of work in a queueing system with service rate c and cumulative work arrival process $A_H(t)$. We note that the mean, variance and Hurst parameter of the arrival process $A_H(t)$ are given by $\bar{a}$, σ^2, and H, respectively We assume that $\bar{a} < c$, so that we can write

$$V = \lim_{t \to \infty} V(t)$$

V is the distribution function of the amount of work in the queueing system at steady state. We want to find q_{max} such that

$$P(V > q_{max}) \approx \epsilon.$$

Norros and Duffield [Norros94, Norros95] showed that

$$\begin{aligned} P(V > q_{max}) &\approx max_{t \geq 0}(P(A_H(t) > ct + q_{max})) \\ &= max_{t \geq 0}(P(A_H(t) - ct > q_{max})) \end{aligned}$$

This approximation was shown to be logarithmically accurate for large q_{max} by Duffield [Duffield95] when $A_H(t)$ is a long-range dependent process. Let $Q(t) = A_H(t) - ct$, so that we can write

$$P(V > q_{max}) = max_{t \geq 0}(P(Q(t) > q_{max})). \tag{15.11}$$

Let $\hat{Z}(t)$ and $\hat{A}_H(t)$ be the envelope processes of $Z(t)$ and $A_H(t)$ respectively

$$P(A_H(t) > \hat{A}_H(t)) = P(Z(t) > \hat{Z}(t)) = \epsilon.$$

Moreover, let

$$\hat{Q}(t) = \hat{A}_H(t) - ct. \tag{15.12}$$

Therefore, at a given time t, we can write

$$P(Q(t) > \hat{Q}(t)) = P(A(t) > \hat{A}_H(t)) = \epsilon.$$

Since $\hat{Q}(t)$ is a deterministic process, we find the maximum of $\hat{Q}(t)$ occurring at time $t^* \geq 0$

$$q_{max} = max(\hat{Q}(t)).$$

Therefore, we can write

$$\epsilon = P(Q(t) > \hat{Q}(t)) \geq P(Q(t) > q_{max}). \tag{15.13}$$

Finally, by combining equations (15.11) and (15.13), we find the tail probabilities to be given by

$$P(V > q_{max}) \approx \epsilon.$$

We can say that *the probability that $Q(t)$ exceeds q_{max} is approximately given by ϵ.* In other words, whenever the arrival process does not exceed its envelope process, the maximum number of cells in the system does not exceed its estimate. Intuitively, we have changed the problem of finding the tail probabilities of a stochastic system into the easier problem of finding the maximum of a deterministic function.

4.2.1 Practical Use: Upper Bound Queue Size for an ATM System with fBm Arrivals . Intuitively, the explanation for our derivation is quite simple. By focusing our attention on a given busy period, if we limit the amount of work that enters the system during the busy period by using an envelope process, we can find a bound for the maximum queue size. In order to use this framework, we just need to i) substitute the formula of the envelope process in equation (15.12), and ii) find the maximum of $\hat{Q}(t)$. We use the fBm envelope process defined previously in order to find q_{max} for the case of fBm arrivals. By substituting the equation (15.8) into equation (15.12) we write

$$\hat{Q}(t) = \hat{A}_H(t) - ct = \bar{a}t + k\sigma t^H - ct. \tag{15.14}$$

Therefore, q_{max} occurs at time $t*$ such that $\frac{d\hat{Q}(t^*)}{dt} = 0$. In this case,

$$\frac{d\hat{A}_H(t^*)}{dt} = c. \tag{15.15}$$

We solve equation (15.15) in order to find $t*$

$$t^* = [\frac{k\sigma H}{(c-\bar{a})}]^{\frac{1}{1-H}}. \tag{15.16}$$

The time-scale of interest regarding queueing performance is defined by the time until the queue size reaches its *peak, i.e., t**. Therefore, we call $t*$ the maximum time-scale (MaxTS), and it defines the point in time where the unfinished work in the queueing system achieves its *maximum* in a probabilistic sense. It means that the average arrival rate has just dropped below the link capacity so that the queue size starts decreasing. The average arrival rate converges to the source's mean arrival rate by the law of large numbers. Therefore, we need to worry only about the time-scale for which the source's rate still exceeds the link capacity, in a probabilistic sense. In other words, after a period of time, the probability that the average arrival rate exceeds the link capacity is negligible, so that the arrival model does not need to reproduce the source's behavior for those time-scales. This is the most important time-scale in terms of traffic modeling. The main difference between LRD and SRD sources is that LRD sources can sustain *high* rates [Mayor96e] for very long periods of time, *i.e.,* they converge more slowly to their mean rate than SRD sources do. As a rule of thumb to choose the parameters of an input source in order to match the fBm envelope process, we need to i) find the MaxTS analytically, and to ii) choose the parameters of the fBm process so that it matches the source's optimal envelope process *at this MaxTS.* By doing so, we can also use the fBm characterization to model SRD sources. In other words, even though a SRD source does not have an envelope process that follows equation (15.8) throughout *all* time intervals, we can still use an fBm envelope process by just focusing on the MaxTS time-scale.

By substituting t^* back in equation (15.14), we compute q_{max}

$$q_{max} = \hat{A}_H(t^*) - ct^*. \tag{15.17}$$

Therefore, q_{max} is given by

$$q_{max} = (c - \bar{a})^{\frac{H}{H-1}} (k\sigma)^{\frac{1}{1-H}} H^{\frac{H}{1-H}} (1 - H). \tag{15.18}$$

Since the fBm arrival process only exceeds $\hat{A}_H(t)$ with probability $\epsilon << 1$, the maximum number of cells in the system will be bounded by q_{max} with the same probability. We find c' so that q_{max} is equal to K. In other words, a buffer of size K will overflow with probability ε if the link capacity is c'. Therefore, c is given by

$$c' = a + K^{\frac{H-1}{H}} (k\sigma)^{\frac{1}{H}} H (1 - H)^{\frac{1-H}{H}}.$$

This result was also obtained by Norros [Norros95]. For the special case *H=0.5,* equation (15.17) degenerates into a simple quadratic equation. Therefore, q_{max} is given by

$$q_{max} = \frac{k^2\sigma^2}{4(c - \bar{a})} = \frac{\sigma^2 \log \epsilon}{2(\bar{a} - c)}. \tag{15.19}$$

This result is exactly the same result given by the diffusion equation for a deterministic service system. On the other hand, if H = 0.90, q_{max} is given by

$$q_{max} = (c - \bar{a})^{-9} (k\sigma)^{10} 0.04.$$

We define β to be the ratio between q_{max} for H equal to 0.90 and 0.50 respectively. Therefore,

Figure 15.7 Maximum Queue Size for $\epsilon = 10^{-6}$.

$$\beta = \frac{{q_{max}}^{0.90}}{{q_{max}}^{0.50}} = 0.16[\frac{k\sigma}{(c-\bar{a})}]^{8} .$$

If the traffic source has *large* variance, β can also be very large. In fact, network traffic has also been shown to suffer from the infinite variance syndrome (IFV) [Mandelbrot69]. In this case, q_{max} can be much larger than the bounds computed by traditional Markov queueing models.

4.2.2 Example . We substitute the Bellcore LAN traffic parameters in equation (15.19). Figure 15.7 shows that if the link utilization is greater than 40%, *i.e.*, $c < 2.5\bar{a}$, the fBm queueing system can exhibit a queue size *100* times greater than the Brownian motion's maximum queue size. The over-optimistic queueing results of traditional models have been reported earlier in [Duffield95, Erramilli96, Mayor96a].

4.3 COMPUTING TAIL PROBABILITIES BY USING LARGE DEVIATION THEORY

Montgomery [Montgomery96] proposes a framework to investigate the time-scale in which cell losses are more likely to occur. We apply his framework to our queueing system driven by an fBm source. By using the same derivation of Section 3.1, the probability that over a time interval of length t the $A_H(t)$ source can overcome the potential service ct and further exceed a buffer level b is given by

$$\begin{aligned} P(A_H(t) > ct + b) = \overline{\Phi}(\frac{t(c-\bar{a})+b}{\sqrt{\bar{a}v}t^H}) &\approx exp(-\frac{1}{2}g(t)^2) \qquad (15.20) \\ &= exp(-\frac{1}{2}(\frac{t(c-\bar{a})+b}{\sqrt{\bar{a}v}t^H})^2) . \end{aligned}$$

Therefore, a minimizer $t^* \in arginf_{t>0}g(t)$ so that the overflow probability is maximized, would correspond to a likely time-scale on which overflow occurs in this system. Therefore, t^* is given by

$$ {t_H}^* = \frac{bH}{(c-\bar{a})(1-H)}. \qquad (15.21)$$

In this case, the probability of buffer overflow is given by

$$P(A_H(t^*) > ct^* + b) \approx exp(-\frac{1}{2}(\frac{t^*(c-\overline{a})+b}{\sqrt{\overline{a}v}t^{*H}})^2).$$

The time-scale of overflow events increases linearly with buffer size. For small buffers, cell losses are more likely to occur at a small time-scale, *i.e,* they are caused by the high-frequency component. Montgomery found that for a Brownian motion process, $t^* = b/(c-\overline{a})$. It is exactly the result developed here when H = 1/2. Therefore $t_H{}^* = \frac{H}{(1-H)}t^*$, *i.e.,* for a given buffer size, cell losses in LRD systems usually occur at larger time-scales than in traditional Markovian systems. Since the low-frequency component is associated with large time-scales, we can say that for large buffer systems, the low-frequency component (LRD) dominates bandwidth requirements. This result was predicted by S.Q. Li in [Li93, Li95].

This time-scale of cell losses, *i.e., t**, was also derived independently by Addie in [Addie95a] and Ryu in [Ryu96a]. Ryu called it the Critical Time Scale (CTS). If we substitute *b* by q_{max} in equation (15.21), we can see that the CTS and the MaxTS are exactly the same time-scale ! By using the large deviation framework Ryu claimed that i) LRD does not cause cell losses for realistic buffer scenarios and ii) traditional SRD models can compute accurate queueing statistics. We dispute the first claim by saying that as long as the input process behaves according to the fBm envelope process throughout the time-scale of interest, LRD is responsible for cell losses. Nevertheless, we agree with Ryu's claim that Markovian models might be able to reproduce the same level of cell losses achieved by LRD processes, by considering the *right* time-scale. But, in many cases, it is simpler to use an LRD process and to compute the tail probabilities, than to match the SRD process's parameters, in order to emulate LRD.

5. STATISTICAL MULTIPLEXING

5.1 HETEROGENEOUS SOURCES

We extend the previous framework to handle statistical multiplexing. In fact, we derive a new framework capable of computing bandwidth and buffer requirements for aggregate traffic composed of heterogeneous LRD sources [Mayor97b]. Assume that we have N independent sources $A^i_H(t)$ defined by the following parameters: mean $\bar{a}_i$, standard deviation $\bar{\sigma}_i$, and Hurst parameter H_i, for $i \in [1, N]$. The aggregate traffic is given by $A_H(t) = \sum_{i=1}^{N} A_H{}^i(t)$. The envelope process of each source is given *by* $\hat{A}^i_H(t)$. The envelope process of the aggregate traffic is given

Figure 15.8 Analytical (solid curve) and simulation (dashed curve) results for $P(Q > q_{max}) \approx 10^{-3}$ versus link utilization.

by $\hat{A}_H(t)$. We can compute q_{max} by finding $t*$ for the envelope process of the aggregate traffic.

The mean of the aggregate traffic is given by the sum of the mean of the individual sources. Similarly, since the sources are independent, the variance of the aggregate traffic is also given by the sum of the variance of the individual sources. Therefore, the envelope process of the aggregate traffic is defined by

$$\hat{A}_H = \sum_{i=1}^{N} a_i t + k(\sum_{i=1}^{N} \sigma_i{}^2 t^{2H_i})^{\frac{1}{2}},$$

where c is the deterministic service rate, *i.e,* the link capacity. Therefore, we can find $t*$ by solving the following equation

$$k\frac{1}{2}(\sum_{i=1}^{N} \sigma_i{}^2 t^{2H_i})^{-\frac{1}{2}} (\sum_{i=1}^{N} \sigma_i{}^2 2H_i t^{2H_i - 1}) = c - \sum_{i=1}^{N} a_i. \qquad (15.22)$$

In other words, we can solve equation (15.220 numerically in order to find $t*$ and substitute it in equation (15.17) in order to find q_{max}. We note that the largest H will dominate the standard deviation of the aggregate traffic for large time-scales.

5.1.1 Example: Heterogeneous Sources . We multiplex three different fBm sources described in Table 1. We compute q_{max} for different values of link utilization and compare to results obtained by simulation. Each fBm trace contains 10,000,000 points and was generated by the fft Algorithm[2] proposed by Paxson in [Paxson95]. In Figure 15.8 we see that the analytical results match closely the simulation results. In our results, the value of H used to define the envelope process is based on the measured (sample) values generated by the fft Algorithm.

5.2 HOMOGENEOUS SOURCES

For homogeneous sources, the envelope process is given by

$$\hat{A}^i_H(t) = \bar{a}t + k\bar{\sigma}t^H.$$

[2] This program is available at "http:/town.hall.org/Archives/pub/ITA/html/contrib/fft-fgn.html"

Table 15.1 Description of the fBm sources.

Source	Mean (cells/time-slot)	Variance	Hurst
1	10	100	0.90
2	15	225	0.70
3	5	25	0.85

Therefore $q^i{}_{max}$ is defined by the following equations:

$$t_i{}^* = [\frac{k\bar{\sigma}H}{(c-\bar{a})}]^{\frac{1}{1-H}}$$

$$q^i{}_{max} = \hat{A}^i_H(t_i{}^*) - ct_i{}^*. \qquad (15.23)$$

Next, we multiplex the sources together and compute $q_{max.}$ Since the sources are independent, the aggregate traffic $A_H(t)$ is defined by $N\bar{a}$, $\sqrt{N}\bar{\sigma}$, and H. The Hurst parameter is preserved under multiplexing of identical sources [Erramilli96]. We investigate q_{max} when the link capacity is Nc. In this case, we can write

$$\hat{A}_H(t) = N\bar{a}t + \sqrt{N}k\bar{\sigma}t^H.$$

In this case, equation (15.22) reduces to

$$\frac{k}{2}\frac{(N\sigma^2 2Ht^{2H-1})}{(\sqrt{N}\sigma t^H)} = N(c-\bar{a}).$$

Therefore,

$$t^* = [\frac{\sqrt{N}k\bar{\sigma}H}{N(c-\bar{a})}]^{\frac{1}{1-H}} = N^{\frac{1}{2(H-1)}}t_i{}^*.$$

Moreover,

$$q_{max} = N(\bar{a}-c)N^{\frac{1}{2(H-1)}}t_i{}^* + N^{\frac{H}{2(H-1)}}N^{\frac{1}{2}}k\bar{\sigma}(t_i{}^*)^H.$$

$$q_{max} = N^{\frac{(H-1/2)}{(H-1)}}q_{max}{}^i \qquad (15.24)$$

We should notice that equation (15.24) is very sensitive to H. There is a significant gain when we multiplex homogeneous sources. For example, if $H = 1/2$, $q_{max} = q_{max}{}^{i}$, *i.e.*, the maximum queue size in the aggregate queue is equal to the maximum queue size in each of the individual queues when we do not multiplex. This result is predicted by the diffusion equation and is the basis of the effective bandwidth approximation [Duffield95]. It is interesting to note that for LRD sources the *multiplexing gain* is even greater. For example, if $H = 0.90$, $q_{max} = \frac{q_{max}{}^{i}}{N^4}$. This is truly astonishing because it predicts a huge savings in terms of buffer requirements when we multiplex several sources. We show that this gain is also predicted by Norros and Duffield's tail probabilities [Duffield95]. Norros [Norros94] showed that the tail probabilities of an ATM queue driven by a fractional Brownian motion (fBm) process is given by

$$P(Q > q) \approx exp(-\frac{(c - \bar{a})^{2H} q^{2-2H}}{2\gamma^2 k^2 \sigma^2}). \tag{15.25}$$

where $\gamma = H^H(1 - H)^{1-H}$, and Q represents the number of cells in the system at steady state. $q^i{}_{max}$ is given by

$$q^i{}_{max} \overset{\text{def}}{=} inf\{q \geq 0 : P(Q > q) < \epsilon\}.$$

We use equation (15.25) in order to find

$$q^i{}_{max} = [-\frac{2 \log \epsilon k^2 \sigma^2 \gamma}{(c - \bar{a})^{2H}}]^{\frac{1}{2-2H}} = \beta$$

If we multiplex N homogeneous sources with link capacity N times larger, we have

$$q_{max} = [-\frac{2 \log \epsilon k^2 \sigma^2 \gamma N}{N^{2H}(c - \bar{a})^{2H}}]^{\frac{1}{2-2H}} = N^{\frac{(H-1/2)}{(H-1)}} \beta = N^{\frac{(H-1/2)}{(H-1)}} q_{max}{}^{i}.$$

Therefore, our result is the same as predicted by Duffield's large deviation equation.

5.2.1 Example: Homogeneous Sources . We substitute the LAN parameters in equation (15.23) and compute q_{max} for a single source with $\epsilon = 10^{-3}$. By using equation (15.24), we also compute q_{max} when four homogeneous sources, defined by the same LAN parameters, are multiplexed together. Figure 15.9 shows that the multiplexer's buffer, required in order to achieve a given cell loss probability, is more than one order of magnitude smaller than the buffer required in the case of a single source.

Figure 15.9 $P(Q > q_{max}) \approx 10^{-3}$ versus link utilization.

5.3 BANDWIDTH REQUIREMENTS

We also derive a framework for computing the bandwidth requirements for aggregate traffic to achieve a maximum probabilistic delay. The problem may be stated as:

Given a set of sources with mean $\bar{a}_i$, standard deviation σ_i, and Hurst parameter H_i, what is the link capacity needed so that the maximum queue size will be bounded by q_{max} with probability ϵ ?

Using equation (15.17) we write

$$c = \frac{\hat{A}_H(t^*) - q_{max}}{t^*}. \tag{15.26}$$

Moreover, we also know that

$$c = \frac{d\hat{A}(t^*)}{dt}. \tag{15.27}$$

By combining equations (15.26) and (15.27) we get

$$k\frac{1}{2}(\sum_{i=1}^{N} {\sigma_i}^2 t^{2H_i})^{-\frac{1}{2}} (\sum_{i=1}^{N} {\sigma_i}^2 2H_i t^{2H_i-1}) - \sum_{i=1}^{N} a_i - k(\sum_{i=1}^{N} {\sigma_i}^2 t^{2H_i-2})^{\frac{1}{2}} - \frac{q_{max}}{t} = 0. \tag{15.28}$$

Therefore, we can find t^* by solving equation (15.28) numerically and then we compute c by substituting t^* in equation (15.26) or in equation (15.27). We can clearly use this to derive an Admission Control Policy for ATM networks.

5.3.1 Example: Heterogeneous Sources.

We multiplex three different sources, two fBm processes and Bellcore's trace[3], see Table 2. We compute the link capacity so that the probability of the queue size exceeding 1,000 cells is 10^{-3}, *i.e.*, $q_{max} = 1,000$ and $\epsilon = 10^{-3}$, By using the framework proposed above, we find c to be equal to 40.12 cells per

[3] For simplicity, we consider each packet arrival in this trace to define the number of cell arrivals within one-time slot in the simulation. Therefore, the resulting arrival process exhibits also LRD.

Table 15.2 Description of the arrival sources.

Source	Mean (cells per time-slot)	Variance	Hurst
Bellcore	8.82	81.79	0.83
fBm process	6.45	3.31	0.70
fBm process	10.33	43.65	0.90

Table 15.3 Delay bounds versus link capacity.

Link Capacity (c)	q_{max}	Actual Maximum Queue Size
40.12	480	1687
36.10	4444	5590

time-slot. We compare it to results driven by a simulation where each source contains 1,000,000 sample points. Table 3 shows that q_{max} given by $P(Q > q_{max}) \approx 10^{-3}$ is indeed below 1,000 cells when $c = 40.12$. Moreover, the actual maximum queue size observed is above 1,000 cells. We repeat the same simulation by underestimating the link capacity by 10%, *i.e.*, c is 90% of its original value. It is interesting to notice that there is a huge increase in the delay experienced by the cells even though the utilization is *only* 71%. In fact, q_{max} exceeds the 1,000 cell delay bound required by the system. Therefore, we believe that this framework is sufficiently accurate to compute the bandwidth required to satisfy the QoS requirements needed by real traffic in ATM networks.

6. NETWORK MANAGEMENT

Previously, we derived a framework for computing bandwidth and buffer requirements for an ATM queueing system driven by LRD sources. We believe that this framework can be used to implement a Connection Admission Control Policy for ATM networks. In this section, we summarize the impact of LRD on flow control and congestion detection protocols [Mayor97a].

6.1 FLOW CONTROL

The Leaky Bucket (LB) is the standard flow control mechanism adopted by the ATM Forum. In order to choose the parameters of an LB mechanism so that it achieves a desirable small violation probability, we have to solve a G/D/l/k queueing system. Therefore, depending on the characteristics of the arrival source, choosing the LB's parameters can be an extremely complex problem. We propose a very simple framework that can be used to set-up the LB parameter. Our calculus, based on the fBm envelope process, is quite general and requires much less computational effort than traditional methods, so that it can be applied in realtime. The main idea is quite simple: choosing the LB parameters based on the envelope process of the arrival source instead of solving the *G*/*D*/1/*k* queueing system.

The LB can be seen as a traffic regulator [Schwartz96, Cruz91a] with output given by the process *L*(*t*) so that

$$L(t) \leq Rt + S.$$

In an interval of length *t* it *accepts* up to (*Rt* + *S*) cells. *L*(*t*) can also be seen as a deterministic envelope process, *i.e.,* it defines the maximum number of cells that a source can send in any time interval. If the arrival process behaves according to its traffic descriptor, the LB should *accept* all incoming cells. Otherwise, if the source *misbehaves*, the LB should mark incoming cells. It is very hard to choose the LB parameters for a *bursty* source so that it only marks cells when the process is misbehaving. In other words, given a traffic descriptor defining the source's behavior, we have to choose the LB parameters so that it accepts all incoming cells, as long as the source behaves according to this descriptor. By using an envelope process, we develop a framework that can be effectively used to set up the parameters, without having to solve a queueing system. Let $A_H(t)$ and $\hat{A}_H(t)$ define the cumulative arrival process and its probabilistic envelope process at time *t*, respectively. In other words, $A_H(t)$ defines the cumulative number of cell arrivals up to time *t*. In order to minimize the probability of incorrectly dropping cells, we should have $L(t) \geq A_H(t), \forall t > 0$. In other words, as long as the source is not misbehaving, the LB mechanism should not mark any incoming cells. Moreover, if we assume that the probability of the source exceeding its probabilistic envelope process is negligible, we can write

$$A_H(t) \leq \hat{A}_H(t), \forall t > 0.$$

Moreover, we assume that

$$L(t) \geq \hat{A}_H(t), \forall t > 0, \tag{15.29}$$

since the probabilistic envelope process is a tighter bound than the deterministic *L*(*t*) envelope process. Therefore, by substituting the fBm envelope process formula into equation (15.29), we can write

$$\bar{a}t + k\sigma t^H \leq Rt + S. \tag{15.30}$$

Figure 15.10 The Leaky Bucket Parameters curve.

Moreover,

$$t(\bar{a} - R) + k\sigma t^{H} - S \leq 0. \qquad (15.31)$$

In order to choose the LB parameters, we find t^* that maximizes equation (15.31) as

$$t^* = [\frac{k\sigma H}{R - \bar{a}}]^{\frac{1}{1-H}}.$$

Substituting t^* back into equation (15.31) we get

$$(\bar{a} - R)[\frac{k\sigma H}{R - \bar{a}}]^{\frac{1}{1-H}} + k\sigma[\frac{k\sigma H}{R - \bar{a}}]^{\frac{H}{1-H}} - S \leq 0 \qquad (15.32)$$

Therefore, by using equation (15.32) we can compute *R* given S, or vice versa. In the case of a Brownian motion process, *i.e.*, independent and identically distributed arrivals, equation (15.32) degenerates to a simple quadratic equation. For the general case, we can solve equation (15.32) numerically.

6.1.1 Example . We substitute the LAN traffic parameters into equation (15.32); see Figure 15.10. The dashed curve corresponds to the special case $H = 0.5$. We can see that even if *R* is close to the average rate, *S* is still relatively *small.* The solid curve corresponds to the case when $H = 0.90$. In this case, if *R* is close to the mean rate, *S* is prohibitively large. In fact, the inability of the LB to police the average rate whenever the source is *bursty* has already been reported previously [Mayor95].

6.2 CONGESTION DETECTION MECHANISMS

Congestion happens when the incoming arrival rate exceeds the output link capacity. Most of the congestion avoidance algorithms are based on feedback mechanisms, *i.e.*, reactive control mechanisms. Their philosophy is quite simple: whenever a node or a switch detects congestion it sends a feedback packet to the source signaling that it should lower its transmission rate in order to avoid cell losses. Most congestion detection algorithms are based on queue occupancy or average arrival rate. In the first case, whenever the number of cells in the ATM queue exceeds a given threshold the feedback packet is sent. But is queue occupancy really an accurate congestion indication. Moreover, is it possible to define a threshold so that there is a high probability that congestion will occur whenever the queue size exceeds this threshold. On the other hand, rate-based congestion indication mechanisms compute the average arrival rate over a fixed time interval in order to detect congestion. If this rate exceeds a rate threshold, they assume that congestion is imminent and the feedback packet is sent to the source. But, is the estimate of the incoming arrival rate a better congestion detection measure than queue occupancy. In order to answer these questions, we compared the efficacy of both congestion detection mechanisms driven by realistic LRD traffic [Mayor96c], We also compared the results to a queueing system driven by a Markov Modulated Arrival Process. Our results were obtained by simulation and the main findings are:

- Queue occupancy is a good estimator for the high-frequency component of the arrival process.
- The average rate reflects fluctuations in the low-frequency component of the arrival process [Li94, Li95].

Therefore, we arrived at the following conclusions:

- In Markov Modulated queueing systems, cell losses are caused by the high-frequency component. In this case, queueing occupancy is an accurate and reliable congestion detector mechanism.
- For LRD sources, queueing occupancy is an accurate mechanism when small buffers are used. In this case, cell losses are also caused by the high-frequency component of the arrival process so that the system behaves similar to a traditional Markovian queueing system [Mayor96e]. On the other hand, for large buffer systems, the average rate is better able to accurately detect congestion than queueing occupancy when the source exhibits LRD.

We believe that an accurate congestion detection mechanism should estimate both the low-frequency and the high-frequency component of the arrival process, before making a decision whether or not congestion is imminent.

7. CONCLUSION

We showed that a self-similar traffic source can transmit at high rates for very long periods of time, possibly generating very long busy (idle) periods. This pervasive behavior leads to the buffer inefficacy phenomenon which has been pointed out by numerous researchers and been verified on real ATM networks. It is our belief that network links might be lightly loaded or heavyly loaded for very long time intervals, so that we should develop dynamic bandwidth allocation mechanisms in order to improve network utilization and avoid congestion.

We showed performance evaluation tools capable of predicting queueing statistics for an ATM queueing network with self-similar arrivals. We also showed that the delay estimates predicted by the fBm envelope process framework agree with large deviation theory estimates but require minimal computational complexity. It also agrees closely with results obtained by trace-driven simulations using Bellcore's LAN traffic. We showed that contrary to our initial intuition, there is a significant *multiplexing gain* when we aggregate several LRD sources. We also studied the efficacy of ATM protocols based on the assumption of self-similar arrivals:

- We showed that the Leaky Bucket mechanism is not a good traffic regulator when the source has long-range dependence.
- We studied the efficacy of congestion detection algorithms. We showed that their efficacy depends primarily on link utilization and buffer size.

References

[Addie95a] R. Addie *et al.,* "Fractal Traffic: Measurements, Modeling and Performance Evaluation", *Proc. of IEEE Infocom'95,* pages 977-984, April 1995.

[Addie95b] R. Addie *et al.,* "Performance of a Single Server Queue with Self-Similar Input", *Proc. of IEEE ICC'95* , pages 461-465, June 1995.

[Beran94] J. Beran, *Statistics for Long-Memory Processes*, New York: Chapman & Hall, 1994.

[Chang94] C. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks",*IEEE Transactions on Automatic Control*, 39(5):943-953, May 1994.

[Chi73] M. Chi, E. Neal and G. Young, "Practical Application of Fractional Brownian Motion and Noise to Synthetic Hydrology". *Water Resources Research*, 9:1523-1533, December 1973.

[Cruz91a] R. Cruz, "A Calculus for Network Delay, Part I: Elements in Isolation", *IEEE Transaction on Information Theory*, 37(1):114-131, January 1991.

[Duffield95] N. Duffield, J. Lewis and N. O'Connel, "Predicting Quality of Service for Traffic with Long-Range Fluctuations", *Proc. of IEEE ICC'95*, pages 473-477, June 1995.

[Erramilli96] A. Erramilli, O. Narayan and W. Willinger, "Experimental Queueing Analysis with Long-Range Dependence Packet Traffic". *IEEE/ACM Transactions on Networking*, 4(2):209-223, April 1996.

[Garret94] M. Garrett and W. Willinger, "Analysis Modeling and Generation of Self-Similar VBR Video Traffic", *Proc. of ACM SIGCOMM'94*, pages 269-279, September 1994.

[Hosking84] J. R. Hosking, "Modeling Persistence in Hydrological Time Series Using Fractional Differencing", *Water Resources Research,* 20(12):1898-1908, 1984.

[Huang95a] C. Huang *et al.*,"Fast Simulation for Self-Similar Traffic in ATM Networks", *Proc. of IEEE ICC'95*, pages 438-444, June 1995.

[Huang95b] C. Huang et al., "Modeling and Simulation of Self-Similar Variable Bit Rate Compressed Video: A Unified Approach", *Proc. of ACM SIGCOMM'95*, pages 114-125, September 1995.

[Huebner] F. Huebner, "On the Accuracy of Approximating Loss Probabilities in Finite Queue by Probabilities to Exceed Queue Levels in Infinite Queues", to be published.

[Jain91b] R. Jain, *"The Art of Computer Systems Performance Analysis"*, John Wiley & Sons, Inc, 1991.

[Lau95] W. Lau et al.,"Self-Similar Traffic Generation: The Random Midpoint Displacement Algorithm and Its Properties", Proc. *of IEEE ICC'95*, pages 466-472, June 1995.

[Leland94] W. Leland, M. Taqqu, W. Willinger and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)*", IEEE/ACM Transactions on Networking,* 2(1):1-15, February 1994.

[Li93] S.Q.Li and C.L. Hwang, "Queue Response to Input Correlation Functions: discrete spectral analysis", *IEEE/ACM Transaction on Networking*, 1(5):522-533, October 1993.

[Li94] H.D. Sheng and S.Q.Li, "Spectral Analysis of Packet Loss Rate at a Statistical Multiplexer for Multimedia Services", *IEEE/ACM Transactions on Networking*, 2(1):53-65, January 1994.

[Li95] S.Q. Li et al., "Link Capacity Allocation and Network Control by Filtered Input Rate in High-Speed Networks", *IEEE/ACM Transactions on Networking,* 3:678-692, February 1995. pages 738-748, April 1996.

[Lin96] Y. Lin, W. Su and C. Lo, “Virtual Path Management in ATM Networks”,*Proc. of IEEE ICC’96*, pages 642-652, June 1996.

[Likhanov95] N. Likhanov and B. Tsybakov, “Analysis of an ATM Buffer with Self-Similar (“Fractal”) Input Traffic,*Proc. of IEEE ICC’95* , pages 985-992, June 1995.

[McLeod78] A. I. McLeod and K. W. Hipel,“Preservation of the Rescaled Adjusted Range: 1. A Reassessment of the Hurst Phenomenon”, *Water Resources Research*, 14(3):491-508, 1978.

[Mandelbrot68] B. Mandelbrot and J. Ness, “Fractional Brownian Motions, Fractional Noises and Applications”, *SIAM Review,* pages 423-437, October 1968.

[Mandelbrot69] B. Mandelbrot, “Long-run Linearity, Locally Gaussian Processes, H-spectra and Infinite Variances”, *International Economic Review*, 10(1)82-106, February 1969.

[Mandelbrot71] B. Mandelbrot,“A Fast Fractional Gaussian Noise Generator”, *Water Resources Research*, 7(1): 543-553, 1971.

[Mayor95] G. Mayor and J. Silvester, “The Multi-level Leaky Bucket Mechanism”,*Proc. of IEEE ICCC’N 95,* September 1995.

[Mayor96a] G. Mayor and J. Silvester, “An ATM Queueing System with a Fractional Brownian Noise Arrival Process”,*Proc. of IEEE ICC’96,* June 1996.

[Mayor96b] G. Mayor and J. Silvester, “A Trace-Driven Simulation of an ATM Queueing System with Real Network Traffic”, *Proc. of IEEE ICCC’N 96,* September 1996.

[Mayor96c] G. Mayor and J. Silvester, “A Comparative Study of Congestion Detection Mechanisms”, *Proc. of IEEE ITS*, pages 229-233, October 1996.

[Mayor96d] “ Time Scale Analysis of an ATM Queueing System with Long-Range Dependent Traffic”, G. Mayor and J. Silvester, to appear in *Proc. of IEEE Infocom,* 1997.

[Mayor96e] G. Mayor and J. Silvester, “An ATM Queueing System with Long-Range Dependent Traffic: Providing QoS Guarantees”, submitted to *ACM/IEEE Transactions on Networking.* Also available as USC Technical Report CENG 96-18.

[Mayor97a] G. Mayor, “Performance Modeling and Network Management for Self-Similar Traffic”, USC Ph.D Thesis, Department of Computer Engineering, 1997.

[Mayor97b] G. Mayor, J. Silvester and N. Fonseca, “Providing QoS for Long-Range Dependent Traffic”, submitted to *IEEE Journal on Se-*

lected Areas in Communications . Also available as USC Technical Report.

[Montgomery96] M. Montgomery and G. de Veciana, "On the Relevance of Time Scales in Performance Oriented Traffic Characterizations",*Proc. IEEE Infocom'96*, pages 513-520, April 1996.

[Narayan] O. Narayan, "Exact Asymptotic Queue Lenght Distribution for Fractional Brownian Traffic", to be published.

[Norros94] I. Norros, "A Storage Model with Self-Similar Input", *Queueing Systems* 16, pages 387-396, 1994.

[Norros95] I. Norros, " The Management of Large Flows of Connectionless Traffic on the Basis of Self-Similar Modeling", *Proc. of IEEE ICC'95*, pages 451-455, June 1995.

[Parekh93] A. Parekh and R. Gallager," A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case", *IEEE/ACM Transactions on Networking,* 2(2):137-150, February 1993.

[Parulekar96] M. Parulekar and A. Makowski, "Tail Probabilities for a Multiplexer with Self-Similar Traffic", *Proc. of IEEE Infocom'96* , pages 1452-1459, April 1997.

[Paxson95] V. Paxson, "Fast Approximation of Self-Similar Network Traffic", UC Berkeley Technical Report, LBL3675C, 1995.

[Pruthi95] P. Pruthi and A. Erramilli,"Heavy-Tailed ON/OFF Source Behavior and Self-Similar Traffic", *Proc. of IEEE ICC'95,* pages 445-450, June 1995.

[Rathgeb91] E. Rathgeb,"Modeling and Performance Comparison of Policing Mechanisms for ATM Networks", *IEEE JSAC,* 9(3):325-334, April 1991.

[Ryu96a] B. Ryu and A. Elwalid,"The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities", *Proc. of ACM Sigcomm'96*, pages 3-14, September 1996.

[Samorodnitsky94] G. Samorodnitsky and M. Taqqu, *"Stable Non-Gaussian Random Processes",* Chapman Hall, 1994.

[Schwartz96] M. Schwartz, *"Broadband Integrated Networks",* Prentice Hall, 1996.

[Taqqu86] M. Taqqu and J. Levy, "Using Renewal Processes to Generate Long Range Dependence and High Variability", *Dependence in Probability and Statistics*, Boston, MA, 1986.

[Willinger95] W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson , "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level", *IEEE Transactions on Networking*, 5(1):71-86, February 1997.

Chapter 16

DISCRETE-TIME ATM QUEUES WITH INDEPENDENT AND CORRELATED ARRIVAL STREAMS

Sabine Wittevrongel and Herwig Bruneel
SMACS Research Group
Department of Telecommunications and Information Processing
Ghent University
Sint-Pietersnieuwstraat 41
B-9000 Gent, Belgium
sw,hb@telin.rug.ac.be

Abstract In this tutorial paper, we present a set of fundamental techniques of analysis for discrete-time queueing models with either independent or correlated arrival streams, deterministic service-time distribution and one or more equivalent servers. The main characteristic of our approach is that it is almost entirely analytical (except for a few minor numerical calculations) and that an extensive use of probability generating functions is being made. The analysis leads to simple and accurate (exact or approximate) formulas for a wide variety of performance measures of practical importance, such as mean and variance of buffer occupancies and cell delays, cell loss probabilities, delay jitter, etc. The theory developed in the paper is also applied to the detailed performance evaluation of ATM multiplexers and ATM switching elements with dedicated-buffer output queueing arrangements.

Keywords: Asynchronous transfer mode, discrete-time queueing theory, analytical techniques, generating-functions approach

1. INTRODUCTION AND PRELIMINARIES

In digital communication networks, buffers are used for the temporary storage of information units which cannot be transmitted instantaneously to their destination when they are generated or when they arrive at a given point in the network. Thus buffers are encountered in multiplexers, concentrators, traffic shapers, switching elements, ..., or, more generally, in any subsystem of a network where some form of competition exists between the various information units, with respect to the use of the available resources. A deep understanding of the behavior of these buffers and the mechanisms that lead to this behavior is of crucial importance to network designers, because the performance of the network may be very closely related to it. For instance, information units may get lost whenever a buffer is fully occupied at the time of their arrival to this buffer, they may experience undesirable delays or delay variations in buffers, and so on.

Usually, information streams are transmitted in digital communication networks according to the synchronous transmission mode. This means that the digital information is chopped in "packets" (or, in an ATM-context, "cells") of fixed length, that time is divided in "slots" of constant length (such that one slot suffices for the transmission of one packet or cell), and that packet or cell transmissions must necessarily begin (and end) at slot boundaries, that is, at a discrete sequence of time points. Therefore, discrete-time queueing models are very natural tools for the statistical analysis of the behavior of buffers in digital communication networks.

In these models, the arrival stream of digital information into a buffer is commonly characterized by specifying the numbers of arriving packets or cells during the consecutive slots. In basic models, these numbers of arrivals are assumed to be independent and identically distributed (i.i.d.) discrete random variables, and the corresponding arrival process is referred to as an independent or uncorrelated arrival process. In some applications, however, such descriptions are not sufficient to characterize the possibly very complicated arrival streams that may occur (for instance, in ATM-based integrated-services digital networks which carry a great variety of information types and services). Therefore, more advanced models will allow the numbers of arrivals during consecutive slots to be nonindependent or correlated. Also, the arrival stream of packets or cells in a buffer need not necessarily be described in a global manner; it can also be characterized on a more microscopic level by specifying the packet (or cell) generating properties of each individual traffic source connected to the input of the buffer, by means of an appropriate uncorrelated or correlated arrival model.

In any case, it is clear that discrete-time queueing models can be conveniently described by a shorthand notation of the form "A-B-c", as introduced in [3], where the A-descriptor refers in some sense to the probability distribution of the number of arrivals per slot, the B-descriptor characterizes the distribution of the "service times" of the information units in the buffer, and the c-descriptor indicates the number of "servers", i.e., the number of output channels of the buffer. Note that this shorthand notation differs from the classical Kendall-notation (of form "A/B/c") used for continuous-time systems (see, for instance, [10]), where the first descriptor A refers to the interarrival-time distribution. Also note that in the present discussion, all information units are assumed to be of fixed length, which implies they have constant transmission times, i.e., the service-time distribution is deterministic, or, the descriptor B always equals D in our case.

In this tutorial paper, we present a set of fundamental techniques for the analysis of discrete-time queueing models which can be used for the performance evaluation of ATM multiplexers and ATM switches. The basic information unit will henceforth be the "ATM cell" (or "cell", for short).

2. MULTIPLEXING AND SWITCHING

An ATM network may be viewed as a collection of network nodes connected by a set of transmission links, whereby the so-called store-and-forward principle is used to transmit ATM cells from a given source, over the network, to a given destination (Figure 16.1). At the entrance point to the network there will typically be some kind of multiplexer which concentrates digital information of various types (such as voice, data, video, ...) and possibly of various users, onto one common link in order to increase the utilization of this link. The ATM cells are then transmitted to a nearby network node, where they are temporarily stored until a transmission channel is available to forward them to a following network node; the same process is then used from node to node until, finally, the cells reach their destination upon going through a demultiplexing stage. It is clear that the (internal) nodes of the ATM network may thus receive cells from a multitude of sources destined for a multitude of destinations. Hence, switching elements are required in each node to route the incoming cells to their correct destinations via an appropriate output link of the node. As the incoming cells may, in general, arrive quite irregularly and their requested destinations may be identical, multiplexers and switches must dispose of buffer space to store the cells which cannot be forwarded immediately. The purpose of this paper is to study the queueing performance of ATM multiplexers and ATM switches in terms of the usual performance characteristics such as buffer occupancy (in cells) and cell delays. Specifically, in this paper, we first use GI-D-*c* type (multiserver) queueing models (where GI stands for "general independent") to analyze the behavior of ATM switches with dedicated-buffer output queueing arrangements. Secondly, we analyze the behavior of ATM multiplexers with nonindependent sources of the on/off-type.

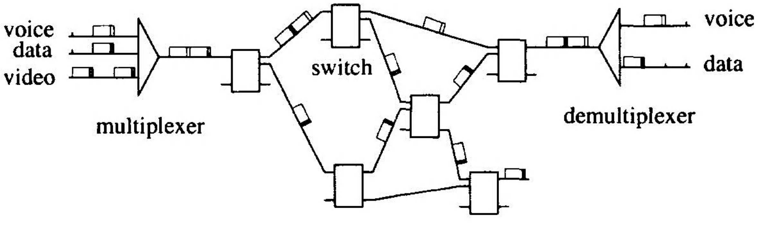


Figure 16.1. Scheme of an ATM network.

3. ATM SWITCHING ELEMENTS WITH OUTPUT QUEUEING AND UNCORRELATED ARRIVALS

An $N \times N$ ATM switching element is a device with N inlets and N outlets. Its task consists of routing incoming ATM cells from a given inlet to a requested outlet (or group of outlets, if several outlets lead to the same destination). Each inlet or outlet can carry exactly one cell during each slot. If during a given slot more cells arrive for a given group of outlets ("destination group") than this group can handle, an output conflict occurs and some cells have to be temporarily stored. Although buffer space can be provided almost anywhere in the switching element, i.e., at the inlets, at the outlets, at some intermediate level, or even a combination of these, we will assume here that all buffering occurs at the output side of the switching element ("output queueing"). For the sake of generality, we assume that the output links are organized in N/c different destination groups, each group containing exactly $c \geq 1$ outlets, and that one separate output buffer is provided for each destination group. A module with this organization will be referred to as an $N \times N(c)$ switching element (Figure 16.2).

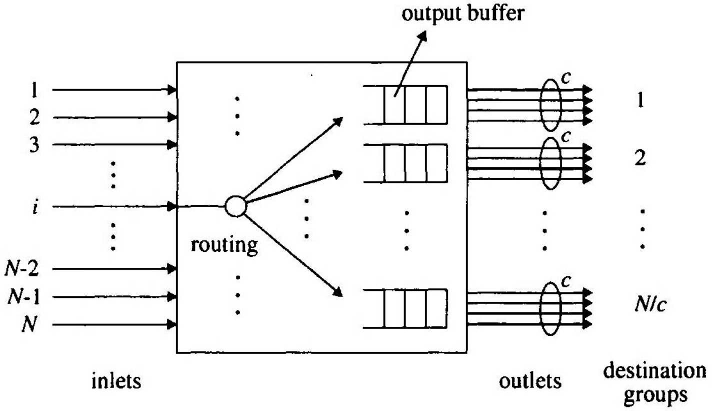


Figure 16.2. $N \times N(c)$ switching element.

Each output buffer thus stores all the cells which cannot be delivered to a given destination immediately because of previous arrivals with the same destination. Note that an output buffer should be primarily considered as a logical concept; in actual implementations, it is usually divided into two separate parts : the actual buffer space, i.e., a memory in which cells wait for transmission, and c hardware registers, which contain up to c cells whose transmission is in progress. In the sequel we will refer to the total number of cells in the (logical) output buffer as the "system contents", and to the number of cells in the actual buffer memory as the "queue contents" of the buffer. It is clear that, in terms of queueing theory, each output buffer can be viewed as a discrete-time queueing system with c servers (i.e., the hardware registers) and up to N arrivals per slot, the details of the cell arrival process being dependent on the characteristics of the sources connected to the inlets of the switch and the distribution of the desired destinations of the cells.

We will assume here that cells enter the system via the inlets according to independent Bernoulli arrival streams, i.e., at most one cell arrival may occur on any inlet, during any slot, independently of the arrivals on other inlets or during other slots. Let ρ denote the probability of a cell arrival. We further assume that the incoming cells are routed independently and uniformly to one of the N/c destination groups, i.e., the probability that a cell is destined for any destination group is equal to c/N, independently of other cells' destinations. Let us concentrate on a tagged output buffer. As the arrival processes on different inlets are independent, it is clear that the total number of cell arrivals in this buffer during each slot has a binomial distribution with parameters N and $\rho c/N$, i.e., with mean value ρc. Also, the total numbers of cell arrivals in the buffer during different slots are independent random variables. In terms of the notation introduced in Section 1, we can thus model each output buffer as a Binom-D-c queueing system. Note that, if the switch size N goes to infinity, the binomial distribution transforms into a Poisson distribution (with mean value ρc) and the queueing model becomes Pois-D-c. The corresponding probability generating functions (pgf's) of the arrival process, which we denote by $E(z)$, are then given by $E(z) = e^{\rho c\,(z-1)}$ for the Poisson case, and by $E(z) = [1-\rho c/N + \rho c\; z/N]^N$ for the binomial case.

The above discussion reveals that the buffer requirements of switching modules with dedicated-buffer output queueing can be investigated through an analysis of the Binom-D-c and the Pois-D-c discrete-time queueing models. We therefore proceed with an analysis of the GI-D-c model which contains both models as a special case.

4. ANALYSIS OF THE GI-D-*C* MODEL

4.1 MODELING ASSUMPTIONS

We briefly summarize the assumptions of the GI-D-c queueing model.

- The buffer has an unlimited storage capacity and $c > 0$ output links.
- Time is divided into fixed-length intervals ("slots"), such that one slot suffices for the transmission of one cell. The transmission of a cell via an output channel of the buffer starts at the beginning of a slot and ends at the end of this slot. This means that cells cannot leave the buffer at the end of the slot during which they have arrived in the buffer.
- New cells enter the buffer according to a general independent arrival process, i.e., the numbers of cells arriving in the buffer during the consecutive slots are modeled as i.i.d. random variables, with a general

probability distribution, characterized by the pgf $E(z)$. Cells may arrive in the buffer at any time instant during a slot. The exact location of the arrival instants within a slot is irrelevant for the analysis.

With the above assumptions, it is clear that a steady state only exists if $E'(1) < c$, i.e., if the mean number of cell arrivals per slot is strictly less than the number of servers. In the sequel, we assume this condition fulfilled.

4.2 SYSTEM CONTENTS AND QUEUE CONTENTS

Let us define s_k as the system contents, i.e., the total number of cells stored in the buffer including those in transmission, at the beginning of slot k, and let $S_k(z)$ denote the pgf of s_k. Also, let s and $S(z)$ denote the steady-state versions (i.e., the limits for $k \rightarrow \infty$) of s_k and $S_k(z)$ respectively. It is easily seen that the evolution of the system contents is governed by the following system equation :

$$s_{k+1} = \left(s_k - c\right)^+ + e_k \quad , \tag{16.1}$$

where $(.)^+$ denotes max(., 0) and e_k represents the total number of cell arrivals during slot k. Using standard z-transform techniques, this system equation can be translated into the z-domain, as follows :

$$S_{k+1}(z) = E(z) \cdot E\left[z^{(s_k - c)^+}\right] = E(z) z^{-c} \cdot \left\{ \sum_{j=0}^{c-1} \left(z^c - z^j\right) \text{Prob}[s_k = j] + S_k(z) \right\}.$$

Taking limits for $k \rightarrow \infty$ and solving the resulting equation for $S(z)$, we then obtain the following expression for $S(z)$:

$$S(z) = \frac{E(z) \sum_{j=0}^{c-1} \left(z^c - z^j\right) \text{Prob}[s = j]}{z^c - E(z)} \quad . \tag{16.2}$$

Equation (16.2) contains the c unknown constants Prob[$s = j$] for $0 \leq j \leq c-1$. These can be determined by invoking the analyticity of the pgf $S(z)$ inside the unit disk $\{z : |z| \leq 1\}$ of the complex z-plane, which implies that any zero of the denominator of (16.2) in this area must necessarily also be a zero of the numerator. Any such zero thus yields one linear equation for the unknowns appearing in the numerator. Now, by use of **Rouché's** theorem (see, for instance, [3]) it can be shown that the denominator of (16.2) has exactly c zeros inside the unit disk, one of which occurs at $z = 1$. Note that the zero at $z = 1$ does not yield any information on the unknowns because the numerator of (16.2) vanishes at $z = 1$ regardless of these unknowns. A c-th linear equation can be obtained from the normalization condition of the system-contents distribution. Using this procedure, it can be shown that

$$S(z) = \left(c - E'(1)\right) \frac{(z-1)E(z)}{z^c - E(z)} \prod_{j=1}^{c-1} \frac{z - z_j}{1 - z_j} \ , \tag{16.3}$$

where the z_j, $1 \leq j \leq c-1$, (and $z = 1$) are the complex zeros of z^c - $E(z)$ inside the unit disk of the complex z-plane. Note that the z_j's can be found easily by numerical means, e.g., by using the Newton-Raphson iteration scheme, in view of the fact that the equation $z^c = E(z)$ can be replaced by an equivalent set of c simpler equations of the form $z = \tilde{E}_j(z)$, for $0 \leq j \leq c-1$, each having exactly one root inside the unit disk of the complex z-plane (one of which is $z = 1$), where the $\tilde{E}_j(z)$'s are the c complex c-th order roots of $E(z)$.

Next, let q and $Q(z)$ denote the queue contents, i.e., the number of cells actually waiting in the buffer (excluding the ones in the servers) at the start of an arbitrary slot in the steady state, and its pgf, respectively. Then, clearly, q is given by $q = (s\text{-}c)^+$ and equation (16.1) implies that $S(z) = E(z).Q(z)$.

The mean system contents and the mean queue contents in the steady state can be found by evaluating the first derivatives of $S(z)$ and $Q(z)$ at $z = 1$, yielding $E[s] = S'(1) = E'(1) + E[q]$ and

$$E[q] = Q'(1) = \frac{1}{2} \sum_{j=1}^{c-1} \frac{1 + z_j}{1 - z_j} + \frac{E''(1) - (c-1)E'(1)}{2\left(c - E'(1)\right)} \ . \tag{16.4}$$

4.3 TAIL DISTRIBUTION OF SYSTEM CONTENTS

In this section, we concern ourselves with the determination of the probability mass function (pmf) Prob$[s = n]$ of the system contents for the GI-D-c model. Specifically, we use an approximation technique to derive explicit expressions for the tail probabilities of the system contents.

The probabilities Prob$[s = n]$ can be determined, in principle, by applying the inversion formula for z-transforms and Cauchy's residue theorem from complex analysis (see e.g. [10]) to the pgf $S(z)$, given in equation (16.3). As a result, Prob$[s = n]$ is then obtained as the negative sum of the residues of $S(z).z^{-1-n}$ in the poles of S(z). It is not difficult to see, however, that this sum of residues will be dominated, for large values of n, by the term associated to the pole (or poles) of $S(z)$ with the smallest absolute value. In the particular case of equation (16.3), the poles of $S(z)$ are the roots of the equation

$$z^c = E(z) \tag{16.5}$$

outside the unit disk of the complex z-plane. This equation can be shown to have the following properties, if $E'(1) < c$.

1. Equation (16.5) has exactly one real positive root, say z_0, larger than 1.
2. The multiplicity of z_0 is 1.
3. Unless $E(z)$ is a function of z^M for some integer M larger than 1, such that M and c have a common divisor larger than 1, equation (16.5) has no other roots with the same absolute value as z_0.
4. Equation (16.5) has no roots outside the unit disk whose absolute value is lower than z_0.

We do not prove the above properties here, but simply mention that the aforementioned theorem of Rouché plays an important role in the proof. The interested reader is referred to [1], [6], [7] and [22] for more details.

We thus conclude that the dominant term in the expression for Prob[$s = n$] is the (negative) residue of $S(z).z^{-1-n}$ in the pole z_0. In view of property 1 above, z_0 can be very easily determined numerically, e.g. by means of the Newton-Raphson algorithm. As its multiplicity is one, the residue formula is also quite simple; as a result, we obtain the following approximation for the tail probabilities of the system contents :

$$\text{Prob}[s = n] \cong -\theta\, z_0^{-1-n} \quad , \quad \text{for large } n \ , \tag{16.6}$$

where

$$\theta \triangleq \lim_{z \to z_0} (z - z_0)\, S(z) = (c - E'(1))\, \frac{(z_0 - 1)\, z_0^{c}}{c\, z_0^{c-1} - E'(z_0)} \prod_{j=1}^{c-1} \frac{z_0 - z_j}{1 - z_j} . \tag{16.7}$$

A quantity of considerable practical interest is the probability that the system contents exceeds a given threshold S. From (16.6), we obtain

$$\text{Prob}[s > S] \cong -\frac{\theta\, z_0^{-S-1}}{z_0 - 1} \quad , \quad \text{for large } S \ . \tag{16.8}$$

In many (ATM) traffic studies the above probability is used as an approximation for the cell loss ratio (or, the overflow probability) (the fraction of cells that arrive at the buffer but cannot be accepted) of a multiserver queue (c servers) with *finite* capacity S and the same arrival statistics (see e.g. [13]). We comment on this approximation later on.

5. BUFFER REQUIREMENTS OF ATM SWITCHES WITH OUTPUT QUEUEING

In this section, we apply the results obtained for the GI-D-c model to the performance analysis of ATM switches with output queueing. Also, we discuss the accuracy of the approximate formulas for the tail probabilities.

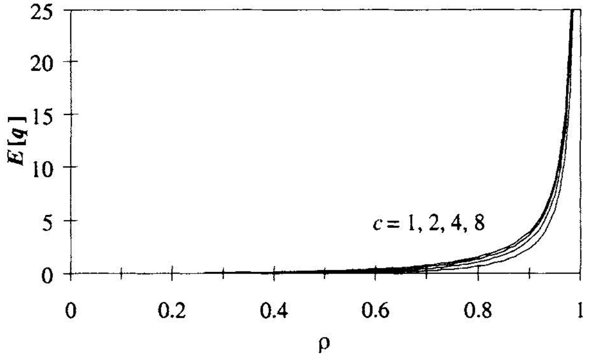


Figure 16.3. Mean queue contents $E[q]$ versus the load ρ, for $N = 32$ and various c.

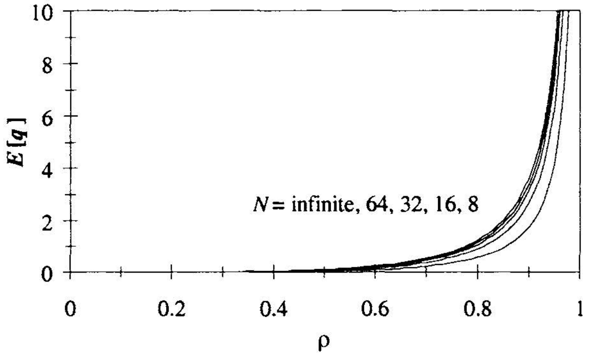


Figure 16.4. Mean queue contents $E[q]$ versus the load ρ, for $c = 4$ and various N.

Specifically, we concentrate on the storage requirements of an output buffer, as characterized, in a global sense, by the mean queue contents $E[q]$, or, in a more specific manner, by the required queue size in order to attain a prescribed value of the cell loss ratio. As an example, let us consider an ATM switch with $N = 32$ inlets and outlets. In Figure 16.3, the mean queue contents $E[q]$ is plotted versus the load ρ, for various values of the number of outputs per destination group c. The figure reveals that, for a given switch size N, the (mean) output queue contents (at the start of a slot) is fairly insensitive to the value of c. As the number of output queues in an Nx$N(c)$ switch is equal to N/c, this implies that the total mean buffer occupancy (at the start of a slot), of all the output queues together, is more or less inversely proportional to c. The influence of the switch size N, for a given destination group size c, is illustrated in Figure 16.4, where we have plotted the mean queue contents $E[q]$ versus the load ρ, for $c = 4$ and various values of N. We observe that, on the average, more buffer space is occupied as N gets larger, but the influence of N becomes negligible as soon as N is sufficiently high. This phenomenon can be intuitively explained by considering the variance of the total number of cell arrivals during an arbitrary slot $\mathrm{var}[e] = \rho c\ (1-\rho c/N)$, which clearly shows that the arrival variance and, hence, also the congestion in the output buffer, increases with N, the increase being most important for high values of c and ρ and becoming negligible for high N.

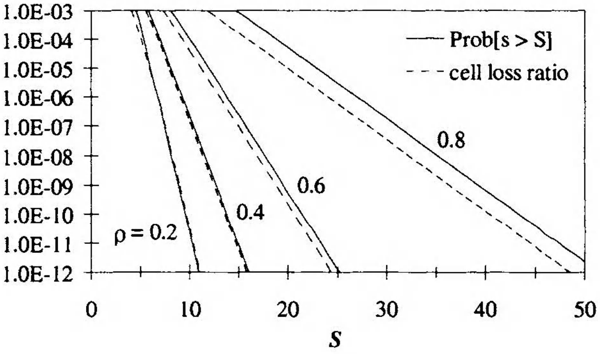


Figure 16.5. Prob[$s > S$] and cell loss ratio versus S, for $N = 16$, $c = 4$ and various ρ.

Let us now turn to the tail probabilities of the system contents. Let us fix the switch size N to 16. The probability of having a system contents greater than S cells in the buffer at the start of a slot is plotted (in solid line) versus the value of S in Figure 16.5, for $c = 4$ and various values of ρ. It has been verified by direct numerical calculation that these "approximate" results, obtained from equation (16.8), are nearly identical to the "exact" results. For instance, for $S = 8$, differences occur only in the sixth significant decimal digit, whereas the deviations decrease further as S gets larger. In Figure 16.5, we have also indicated the corresponding values of the actual cell loss ratio when the queue has a finite capacity equal to S cells (which implies that no more than S cells can be present in the system at the beginning of a slot). These values are represented by the dashed lines in the figure and were also obtained numerically (by solving one set of balance equations for each value of S !). The plots reveal an acceptable agreement between the cell loss ratio of the finite-capacity model and the tail probabilities of the infinite-capacity model for intermediate values of the load. For high values of the load, the infinite-capacity model yields greater values, whereas the inverse implication holds for low traffic. Similar phenomena can be observed for continuous-time models, e.g. in a comparison between the M/M/1 and M/M/1/K models, see e.g. [10]. The buffer size required to attain a prescribed cell loss ratio of e.g. 10^{-10} can be easily obtained from these graphs : for instance, if $c = 4$, the required buffer size (i.e., S) is given by 21 and 41 for $\rho = 0.6$ and $\rho = 0.8$ respectively. Note that, if the tail probabilities

Prob[$s > S$] were used instead of the cell loss ratios, the results would have been about 22 and 44 respectively, i.e., a little higher than the actual values. As, in practice, loads higher than 0.5 are more likely to occur than lower ones, it is usually safe to use tail probabilities instead of cell loss ratios, whose calculation (by numerical means) is much more involved.

It is clear that the cell loss ratio of any given output buffer of an $N\times N(c)$ switching module is also the cell loss ratio for the whole module, owing to the statistical equivalence of the output queues. However, the required total buffer space (for all the output buffers together) to attain such a cell loss ratio is N/c times higher than for an individual output buffer, if we assume that dedicated output buffers are used, i.e., that no buffer sharing is applied (between output buffers). In Figure 16.6 we have plotted the cell loss ratio of an $N\times N(c)$ switching element with a capacity of S cells per output queue (as approximated by the probability Prob[$s > S$] for one output queue) versus the total required buffer space, i.e., the quantity $(N/c)\ S$, for $N = 16, \rho = 0.8$, and various values of c. The figure confirms the results obtained for the total mean buffer occupancy, that the use of multiserver output queues seriously reduces the buffer requirements of a switching module. Specifically, it can be observed from Figure 16.6 that the required (total) buffer size to attain a given cell loss level is roughly inversely proportional to c, in agreement with our earlier observations for the (total) mean buffer occupancy.

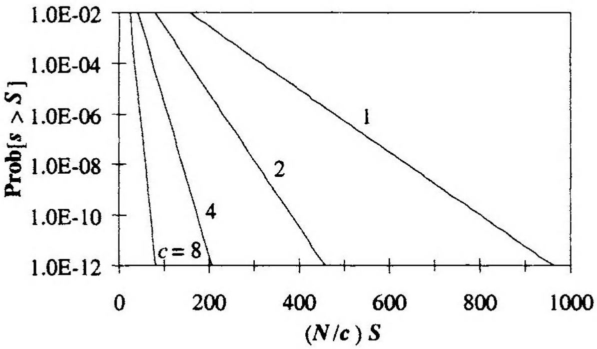


Figure 16.6. Prob[$s > S$] versus $(N/c)\ S$, for $N = 16$, $\rho = 0.8$ and various c.

6. ATM STATISTICAL MULTIPLEXERS WITH CORRELATED ARRIVALS

An ATM statistical multiplexer is a device with N inlets and one outlet, whose function is to provide the sharing of one common communication channel (i.e. the outlet) by a multitude of users or traffic sources connected to the inlets. For this purpose, the multiplexer collects the ATM cells coming from the different inlets in one common buffer and then transmits the cells on the outlet at the rate of one per slot as long as the buffer is nonempty. A realistic queueing model for a multiplexer essentially implies a statistical description of the traffic sources that generate the cells to be transmitted.

In some studies, the arrival streams of cells in the multiplexer buffer are described as being independent from slot to slot. The corresponding queueing model is the GI-D-1 model, which is a special case of the model analyzed in Section 4. However, it has been observed that the traffic streams generated by typical ATM sources tend to be of a correlated nature. For the design of ATM networks it is therefore essential to model a certain degree of slot-to-slot dependency in the cell arrival streams on the multiplexer inlets.

Very popular in this respect is the so-called on/off source model, where each source alternates between on-periods and off-periods, because of its analytical tractability. Discrete-time models with geometric distributions for the on-periods and the off-periods are discussed, for instance, in [2], [16] and [20]. A somewhat related discrete-time model in which the on-periods consist of a geometrically distributed number of constant-length intervals is considered in [25], and the case of a mixture of two geometric distributions for the on-periods is treated in [18]. We consider here an even more general source model with an *arbitrary* distribution for the on-periods. In the next section, we present the analysis of the corresponding queueing model with correlated arrivals, which we denote as the COR-D-1 model. With appropriate modifications, the analysis method to be developed can also be used to analyze the buffer behavior for other types of correlated arrivals.

7. ANALYSIS OF THE COR-D-1 MODEL

7.1 MODELING ASSUMPTIONS

The assumptions of the COR-D-1 queueing model are as follows.

- The buffer system consists of one single server and an infinite-capacity waiting room for cells.
- Time is slotted. The transmission of a cell takes exactly one slot and can start or end at slot boundaries only.
- Cells are generated by a finite number N of independent sources of the on/off type. Each source alternates between on-periods and off-periods. During an on-period, a source generates one cell per slot. No cells are generated during an off-period. We assume that the (lengths of the) off-periods are geometrically distributed with parameter β, i.e., Prob[off-period = i slots] = $(1-\beta)\,\beta^{i-1}$, $i \geq 1$. Furthermore, the (lengths of the) on-periods are assumed to be i.i.d. random variables with pgf $A(z)$ and pmf $a(i)$. Finally, it is assumed that the on-periods and the off-periods are independent.

In the sequel, the average number of cell arrivals in the buffer per slot is assumed to be strictly less than 1, so that the buffer system can reach a steady state.

7.2 SYSTEM EQUATIONS

As mentioned before, we assume that each source will alternately be off (state B), or on. A source is called in state A_n, $n \geq 1$, if it is in the nth slot of an on-period. Hence, each source can be characterized by an infinite-dimensional Markov chain with states B and A_n, $n \geq 1$, and transition probabilities as shown in Figure 16.7, where $p_a(n\text{-}1)$ is the probability of having an on-period of at least n slots, given that this on-period consists of at least n-1 slots, i.e.,

$$p_a(n-1)=\left(\sum_{i=n}^{\infty} a(i)\right)\Big/\left(\sum_{i=n-1}^{\infty} a(i)\right) \quad , \qquad n>1 \ . \tag{16.9}$$

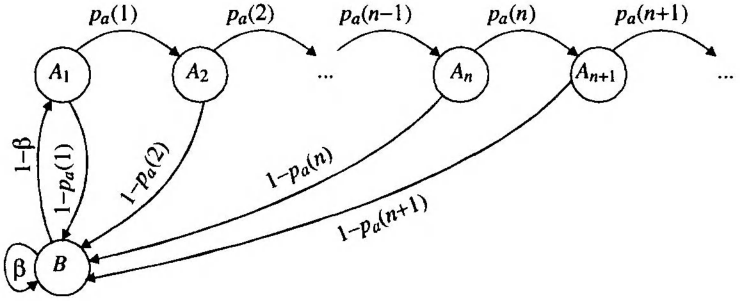


Figure 16.7. State transition diagram of a multiplexer inlet.

Let us define the random variables $g_{n,k}$ $(n \geq 1)$ as the number of sources in the nth slot of an on-period during slot k. Also, let g_n denote the steady-state version of $g_{n,k}$. Due to the infinite-state cell arrival model on each inlet described above, the joint pgf $N(x_1, x_2, ...)$ of g_n $(n \geq 1)$ is given by

$$N(x_1,x_2,\ldots) \triangleq E\left[\prod_{n=1}^{\infty} x_n^{g_n}\right]=\left[v_b+\sum_{n=1}^{\infty} v_a(n)x_n\right]^N \quad , \tag{16.10}$$

where $v_a(n)$ $(n \geq 1)$ and v_b denote the steady-state probabilities of finding an inlet in state A_n or state B respectively, during an arbitrary slot. These probabilities can be calculated from the balance equations for the Markov chain in Figure 16.7, together with the normalization equation. As a result, the joint pgf $N(x_1, x_2, ...)$ is obtained as

$$N(x_1,x_2,\ldots)=(1-\sigma)^N\left[1+(1-\beta)\sum_{n=1}^{\infty}\sum_{i=n}^{\infty}a(i)x_n\right]^N \quad , \qquad (16.11)$$

where σ is the load of one source, i.e., $\sigma=(1-\beta)A'(1)/[1+(1-\beta)A'(1)]$.

From Figure 16.7, we observe that exactly two transitions are possible from each state : transition to the same period, but one slot further or transition to the first slot of the other period. Consequently, $g_{n,k}$ $(n>1)$ contains one unity for each source which was in state A_{n-1} during slot k–1 and which changes to state A_n in the next slot. Similarly, $g_{1,k}$ contains one unity for each source which was in state B during slot k–1 and which changes to state A_1 in slot k. Therefore, the following relationships exist:

$$g_{1,k}=\sum_{i=1}^{N-\sum_{n=1}^{\infty}g_{n,k-1}}d_i \quad ; \qquad g_{n,k}=\sum_{i=1}^{g_{n-1,k-1}}c_{n-1,i} \quad , \qquad n>1 \ . \qquad (16.12)$$

Here the d_i's are i.i.d. Bernoulli random variables with pgf

$$D(z)\triangleq E\left[z^{d_i}\right]=\beta+(1-\beta)z \ . \qquad (16.13)$$

For given n, the $c_{n\text{-}1,i}$'s are i.i.d. Bernoulli random variables with pgf

$$C_{n-1}(z)\triangleq E\left[z^{c_{n-1,i}}\right]=1-p_a(n-1)+p_a(n-1)z \quad , \qquad n>1 \ . \qquad (16.14)$$

Moreover, the d_i's and the $c_{n\text{-}1,i}$'s are mutually independent.

As in Section 4.2, let the random variable s_k denote the system contents at the beginning of slot k. Then the system contents evolves according to the system equation :

$$s_{k+1}=\left(s_k-1\right)^+ +\sum_{n=1}^{\infty}g_{n,k} \quad , \qquad (16.15)$$

where the sum represents the total number of cell arrivals during slot k.

The above equations (16.12)-(16.15) make clear that the set of random variables $\{(g_{n,k-1}\ (n\geq 1),\ s_k)\}$ forms an infinite-dimensional Markov chain. In other words, the random vector $(g_{n,k-1}\ (n\geq 1),\ s_k)$ completely describes the state of the queueing system at the beginning of slot k.

7.3 FUNCTIONAL EQUATION

In order to analyze the buffer behavior, we define the joint pgf $P_k(x_1,x_2,\ \ldots,\ z)$ of the state vector $(g_{n,k-1}\ (n\geq 1),\ s_k)$ as

$$P_k(x_1,x_2,\dots,z) \triangleq E\left[\left(\prod_{n=1}^{\infty} x_n^{g_{n,k-1}}\right) z^{s_k}\right] . \tag{16.16}$$

Using the system equations (16.12)-(16.15) and standard z-transform techniques, the pgf $P_{k+1}(x_1, x_2,\dots, z)$ can then be expressed as

$$P_{k+1}(x_1,x_2,\dots,z) = [D(x_1 z)]^N \, E\left[\left(\prod_{n=1}^{\infty}\left(\frac{C_n(x_{n+1}z)}{D(x_1 z)}\right)^{g_{n,k-1}}\right) z^{(s_k-1)^+}\right] ,$$

where the expectation is over the joint distribution of $(g_{n,k-1}\ (n \geq 1),\ s_k)$. Since the transmission of cells is synchronized to the slot boundaries, each cell that arrives during a slot is still in the buffer system at the start of the next slot. Stated otherwise, having an empty system at the beginning of slot k, i.e. $s_k = 0$, implies that no cells have arrived during slot k–1, and hence also that $g_{n,k-1} = 0\ (n \geq 1)$. In view of this property, we get the following functional equation for the steady-state version $P(x_1, x_2, \dots, z)$ of $P_k(x_1, x_2, \dots, z)$:

$$P(x_1,x_2,\dots,z) = \frac{[D(x_1 z)]^N}{z}\left\{P\left(\frac{C_1(x_2 z)}{D(x_1 z)}, \frac{C_2(x_3 z)}{D(x_1 z)},\dots,z\right) + (z-1)p_0\right\} , \tag{16.17}$$

where the quantity p_0 indicates the probability of having an empty buffer at the beginning of an arbitrary slot in the steady state.

Unfortunately, we are not able to derive from (16.17) an explicit expression for $P(x_1, x_2, \dots, z)$ or not even for the pgf $S(z)$ of the steady-state system contents s. However, as shown in the following, all the relevant information concerning the system contents can be extracted from this functional equation, if we consider in (16.17) only those values of $x_n\ (n \geq 1)$ and z for which the arguments of the P-functions on both sides of (16.17) are equal to each other, i.e., $x_n = C_n(x_{n+1} z)/D(x_1 z)$, $n \geq 1$. From this equation, x_n $(n \geq 1)$ can be solved in terms of z. It turns out that for a given z, there may be more than one set of solutions. Here, we only choose the set of solutions which has the additional property that $x_n = 1\ (n \geq 1)$ for $z = 1$. Denoting this set of solutions by $\chi_n(z)$, we can show that

$$z\,\chi_n(z) = \left(\sum_{i=n}^{\infty} a(i)\left(D(\chi_1(z)z)/z\right)^{n-1-i}\right) \Big/ \left(1 - \sum_{i=1}^{n-1} a(i)\right) , \quad n \geq 1 . \tag{16.18}$$

Note in particular that

$$z\chi_1(z) = A\left(\frac{z}{\beta + (1-\beta)\chi_1(z)z}\right) \quad . \tag{16.19}$$

Choosing $x_n = \chi_n(z)$, $n \geq 1$, in (16.17), we then get a linear equation for the function $P(\chi_1(z), \chi_2(z), ..., z)$, which has the following normalized solution :

$$P(\chi_1(z), \chi_2(z), \ldots, z) = \frac{(z-1)(1-\rho)F(z)}{z - F(z)} \quad . \tag{16.20}$$

Here $\rho = N\sigma$ is the total multiplexer load and

$$F(z) \doteq [D(\chi_1(z)z)]^N = [\beta + (1-\beta)\chi_1(z)z]^N \quad . \tag{16.21}$$

In the next sections, we describe a technique to derive from equation (16.20) closed-form expressions for the mean value and the tail distribution of the system contents.

7.4 MEAN SYSTEM CONTENTS

Before calculating the mean system contents, we introduce some interesting source characteristics in terms of which we will express our results. First, it is not so difficult to see that the mean lengths of the on-periods and the off-periods are given by

$$A'(1) = \frac{K}{1-\sigma} \qquad \text{and} \qquad \frac{1}{1-\beta} = \frac{K}{\sigma} \quad , \tag{16.22}$$

for some constant K, which will be referred to as the burstiness factor. Note that the quantity K equals the ratio of the mean length of an on-period (or an off-period) in our model, to the mean length of the corresponding quantity in case of a Bernoulli arrival process. It is clear that the load σ describes the ratio of the mean lengths of the on-periods and the off-periods, whereas the burstiness factor K is a measure for the absolute lengths of these periods for a given load. Also we define the variance factor L_a of the source as the ratio of the variance of the on-period length in our model, to the variance of a geometrically distributed on-period with the same mean length, i.e.,

$$L_a = \frac{A''(1) + A'(1) - [A'(1)]^2}{A'(1)\,[A'(1) - 1]} \quad . \tag{16.23}$$

The pgf $S(z)$ of s can now be expressed as $S(z) = P(1, 1, ..., z)$. In order to obtain an expression for the mean system contents $E[s]$, we evaluate the first derivative of equation (16.20) with respect to z at $z = 1$. After some algebra, we then get

$$E[s] = \rho + \frac{(N-1)\rho^2}{2N(1-\rho)} \left[K + L_a (K-1) + \frac{\rho}{N} (L_a - 1) \right] . \tag{16.24}$$

It has been checked that the above general result is in agreement with the results obtained in [2], [18] and [25]. The above formula clearly demonstrates that the multiplexer performance depends not only on the mean length of the on-periods, but also strongly on the variance of the on-periods. First, we observe that for a given total load, the mean length of the on-periods has a substantial influence on $E[s]$. The mean system contents namely linearly increases with the burstiness factor K of the sources. Next, for a given load and a given mean length of the on-periods (given K), the mean system contents linearly increases with L_a, i.e., $E[s]$ increases linearly with the variance of the on-periods. Higher-order moments of the on-period distribution have no impact on the mean system contents.

7.5 TAIL DISTRIBUTION OF SYSTEM CONTENTS

Another important performance measure is the tail distribution of the system contents. It has been observed in many cases (see e.g. Section 4.3) that the tail distribution of the system contents has a geometric form. In such a case, an approximation for the tail distribution of the system contents can be expressed as

$$\text{Prob}[s = n] \cong -\theta z_0^{-1-n} \quad , \quad \text{for large } n \ . \tag{16.25}$$

Here z_0 is the pole with the smallest modulus of $S(z)$, which must necessarily be real and positive in order to ensure that the tail distribution be nonnegative anywhere, and θ is the residue of $S(z)$ in the point $z = z_0$.

7.5.1 Calculation of z_0

As in [25], it can be argued that z_0 is also the pole with the smallest modulus of $P(\chi_1(z), \chi_2(z), ..., z)$. Hence, in view of (16.20) and (16.21), z_0 is a real root of $z - F(z) = 0$, or

$$z - \left[\beta + (1-\beta)\chi_1(z) z\right]^N = 0 \ . \tag{16.26}$$

This can even be rigourously proved. As all sources are statistically independent, $F(z)$ is the Perron-Frobenius eigenvalue related to the aggregated arrival process to the multiplexer [12]. Hence, the dominant pole z_0 is determined by $z - F(z) = 0$ [15]. It is obvious that $\chi_1(z) > 0$ for $z > 1$. From (16.19) and (16.26), we then obtain the following equation tor z_0 :

$$\frac{z^{1/N} - \beta}{1 - \beta} - A\left(\frac{z}{z^{1/N}}\right) = 0 \quad . \tag{16.27}$$

From (16.27), the pole z_0 can easily be calculated exactly by means of, for instance, the Newton-Raphson iteration scheme.

7.5.2 Calculation of θ

Let us consider the case where the number of cells stored in the buffer just after a given slot is sufficiently large ($\gg N$). Then we may think that the number of cell arrivals during this slot (which cannot be larger than N) has almost no impact on the total buffer contents. Consequently, if j is sufficiently large ($j > J$), we may assume that the conditional probabilities $\text{Prob}[g_n = i_n\ (n \geq 1) \mid s = j]$ are almost independent of j, and approach to some limiting values for $j \to \infty$, denoted by $\omega(i_1, i_2, ...)$, i.e.,

$$\text{Prob}\left[g_n = i_n\ (n \geq 1) \mid s = j\right] \cong \omega\left(i_1, i_2, \ldots\right) \quad , \qquad j > J \quad , \tag{16.28}$$

with corresponding joint pgf $\Omega(x_1, x_2, ...)$.

Using (16.28), we can now express the joint pgf $P(x_1, x_2, ..., z)$ as

$$\begin{aligned} P\left(x_1, x_2, \ldots, z\right) \cong \sum_{i_1} \sum_{i_2} \cdots \sum_{j=0}^{J} \text{Prob}\left[g_1 = i_1, g_2 = i_2, \ldots, s = j\right] \left(\prod_{n=1}^{\infty} x_n^{\,i_n}\right) z^j \\ + \Omega\left(x_1, x_2, \ldots\right) \left(S(z) - \sum_{j=0}^{J} \text{Prob}\left[s = j\right] z^j \right) \quad . \end{aligned}$$

Setting $x_n = \chi_n(z)$, $n \geq 1$, we know that z_0 is a pole of both the P-function and $S(z)$. As J is finite, multiplying by $(z - z_0)$ on both sides of the above equation and taking the $z \to z_0$ limit, we find

$$\theta = \frac{\left(z_0 - 1\right)\left(1 - \rho\right) z_0}{\left[1 - F'\left(z_0\right)\right] \Omega\left(\chi_1\left(z_0\right), \chi_2\left(z_0\right), \ldots\right)} \quad . \tag{16.29}$$

In order to derive the pgf $\Omega(x_1, x_2, ...)$, we let $\pi(i_1, i_2, ... \mid k_1, k_2, ...)$ denote the one-step transition probability that there are i_n ($n \geq 1$) sources in the nth slot of an on-period during a slot, given that there were k_l ($l \geq 1$) sources in the *l*th slot of an on-period in the previous slot. Then, we have (for large *j*)

$$\text{Prob}\left[g_n = i_n \ (n \geq 1),\ s = j\right]$$
$$= \sum_{k_1} \sum_{k_2} \cdots \pi(i_1, i_2, \ldots \mid k_1, k_2, \ldots) \ \text{Prob}\left[g_l = k_l \ (l \geq 1),\ s = j + 1 - \sum_{n=1}^{\infty} i_n \right] .$$

Taking limits for $j \to \infty$ and using (16.25) and (16.28), we obtain

$$z_0 \, \omega(i_1, i_2, \ldots) = (z_0)^{\sum_{n=1}^{\infty} i_n} \sum_{k_1} \sum_{k_2} \cdots \ \pi(i_1, i_2, \ldots \mid k_1, k_2, \ldots) \, \omega(k_1, k_2, \ldots) \ .$$

The following equation for the pgf $\Omega(x_1, x_2, \ldots)$ can then be derived :

$$z_0 \, \Omega(x_1, x_2, \ldots) = \left[D(x_1 z_0) \right]^N \Omega\left(\frac{C_1(x_2 z_0)}{D(x_1 z_0)}, \frac{C_2(x_3 z_0)}{D(x_1 z_0)}, \ldots \right) \ . \qquad (16.30)$$

As can be expected intuitively, it is possible to show that the solution $\Omega(x_1, x_2, \ldots)$ of (16.30) has the same form of expression as the pgf $N(x_1, x_2, \ldots)$ corresponding to the unconditional cell arrival process. Specifically, $\Omega(x_1, x_2, \ldots)$ can be expressed as

$$\Omega(x_1, x_2, \ldots) = \left(1 - \sum_{n=1}^{\infty} \mu_a(n) + \sum_{n=1}^{\infty} \mu_a(n) \, x_n \right)^N , \qquad (16.31)$$

where $\mu_a(n)$ $(n \geq 1)$ is the (conditional) probability of finding a source in the *n*th slot of an on-period, when the number of cells in the multiplexer buffer is extremely large. From equations (16.18), (16.26), (16.30) and (16.31), an explicit expression can be derived for $\Omega(\chi_1(z_0), \chi_2(z_0), \ldots)$. Also, from equations (16.19) and (16.21), we obtain a closed-form expression for $F'(z_0)$. Finally, we find the following explicit expression for the residue θ :

$$\theta = \frac{(1-\rho) z_0^{\,2} (z_0 - 1)^{N+1}}{\left[1 - (N-1) R(z_0) \right] \left[R(z_0) + 1 \right]^{N-1} \left(z_0 - z_0^{\,1/N} \right)^N} , \qquad (16.32)$$

where

$$R(z_0) = A'\left(\frac{z_0}{z_0^{\,1/N}} \right) \frac{(1-\beta) z_0}{z_0^{\,2/N}} \ . \qquad (16.33)$$

8. BUFFER REQUIREMENTS OF ATM MULTIPLEXERS WITH ON/OFF SOURCES

In this section, we will use the above analysis to investigate the influence of the distribution of the on-periods on the buffer behavior. We consider the following examples for the pgf $A(z)$:

$$A_1(z) = z^m \quad ; \qquad A_2(z) = \frac{(1-\lambda)^2 z}{(1-\lambda z)^2} \quad ; \qquad A_3(z) = \frac{(1-\alpha)\, z}{1-\alpha z} \quad ;$$

$$A_4(z) = \frac{p\,(1-\alpha_1)\, z}{1-\alpha_1 z} + \frac{(1-p)\,(1-\alpha_2)\, z}{1-\alpha_2 z} \quad , \qquad \alpha_1 > \alpha_2 \quad ,$$

i.e. constant-length on-periods, a negative binomial distribution, a geometric distribution and a mixture of two geometric distributions respectively. In order to study the impact of the variance of the on-periods on the "overflow probability" Prob[$s > S$], we choose the parameters of these distributions such that the mean length of the on-periods is equal to m in all cases. It can be shown that this corresponds to choosing

$$\lambda = \frac{m-1}{m+1} \quad ; \qquad \alpha = \frac{m-1}{m} \quad ; \qquad \alpha_2 = 1 - \frac{(1-p)(1-\alpha_1)}{m(1-\alpha_1)-p} \quad ;$$

$$\frac{1}{1-\alpha_1} = m + \sqrt{\frac{(1-p)\left[\mathrm{var}_4 - m(m-1)\right]}{2p}} \quad ; \quad .$$

Here var$_4$ is the variance of the on-period lengths for the mixed geometric distribution, whereas the on-period variances corresponding to the pgf' s $A_1(z)$, $A_2(z)$ and $A_3(z)$ are given by

$$\mathrm{var}_1 = 0 \quad ; \qquad \mathrm{var}_2 = \tfrac{1}{2}(m-1)(m+1) \quad ; \qquad \mathrm{var}_3 = m\,(m-1) \quad .$$

In Figure 16.8, Prob[$s > S$] is plotted versus S, for $N = 16$, $\rho = 0.8$, $K = 5$ and the above four distributions for the lengths of the on-periods. The corresponding variances of the on-period lengths are then var$_1$ = 0, var$_2$ = 13.35, var$_3$ = 22.44 and var$_4$ = 56.09 respectively, and p = 0.4. The variance factors are given by $L_{a,1} = 0$, $L_{a,2} = 0.595$, $L_{a,3} = 1$ and $L_{a,4} = 2.5$. It is clear that for given values of the load and the mean lengths of the on-periods, the variance of the on-periods has a strong impact on the performance. We observe that the performance degrades with increasing variance of the on-period lengths.

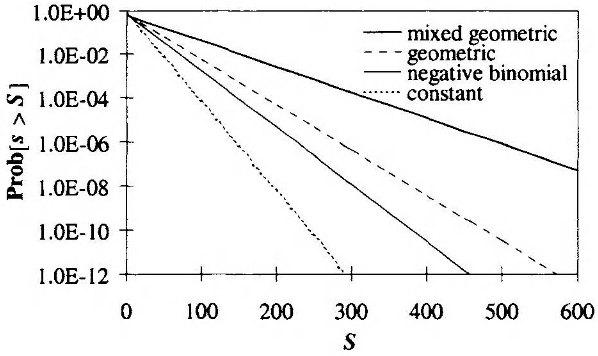


Figure 16.8. Prob[$s > S$] versus S, for $N = 16$, $\rho = 0.8$, $K = 5$, and various distributions of the on-period lengths.

In Figure 16.9, we consider a mixture of two geometric distributions for the lengths of the on-periods, and we have plotted the buffer overflow probability Prob[$s > S$] in terms of S, for $N = 8$, $\rho = 0.8$, $K = 5$, $L_a = 2$ and $p = 0.5$, 0.05, 0.01, 0.005, 0.001. The corresponding values of the third moment M_a are $M_a = 2038.68$, 3154.02, 4698.91, 5829.63, 10566.88. The figure clearly shows that the performance gets worse as the third moment M_a increases.

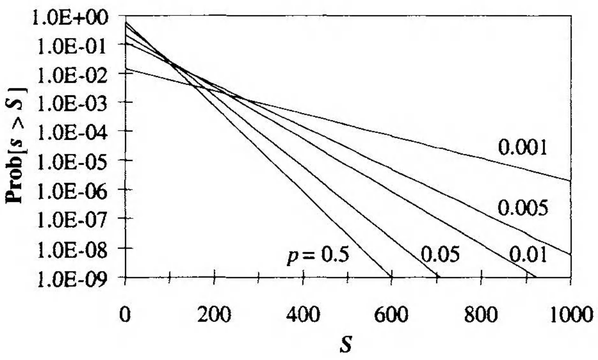


Figure 16.9. Prob[$s > S$] versus S, for $N = 8$, $\rho = 0.8$, a mixture of two geometric distributions for the lengths of the on-periods, $K = 5$, $L_a = 2$ and various values of p.

9. CELL DELAY FOR THE G-D-C MODEL

In the previous sections, we have focused our attention on performance measures related to the distribution of the system contents. In this section, we will study the delay a cell experiences in the buffer, under a first-come-first-served queueing discipline. We consider the G-D-*c* queueing model, for which the modeling assumptions are the following :

- The buffer has an infinite storage capacity and *c* servers.
- The transmission time of a cell is exactly one slot.
- Cells arrive in the buffer system according to a general, possibly correlated, arrival process which is not further specified.

Note that the G-D-*c* model contains both the GI-D-*c* model and the COR-D-1 model as special cases.

The delay of a cell is defined as the number of slots between the end of the arrival slot of the cell, and the end of the slot during which the cell is transmitted from the buffer. Let the random variable *u* indicate the delay of an arbitrary ("tagged") cell in the steady state, and let *U*(*z*) denote the corresponding pgf. In [26], the following relationship was established between *U*(*z*) and the pgf *S*(*z*) of the system contents :

$$U\left(z^c\right)=\frac{1}{cE'(1)}\sum_{j=0}^{c-1}\frac{1-z^{-c}}{1-\left(a^jz\right)^{-1}}\left\{\frac{1-z^c}{1-a^jz}S\left(a^jz\right)-\sum_{i=0}^{c-1}\frac{\left(a^jz\right)^i-z^c}{1-a^jz}\ \mathrm{Prob}[s=i]\right\}\ , \tag{16.34}$$

in terms of the *c* probabilities $\mathrm{Prob}[s=i]$, $0\le i\le c-1$, the mean number of cell arrivals $E'(1)$ during an arbitrary slot, and the *c* complex roots a^j, $0\le j\le c-1$, of the equation $z^c = 1$ ($a=e^{2\pi\iota/c}$ where ι is the imaginary unit).

Several parameters are of interest with respect to cell delays : the mean cell delay *E*[*u*] gives a global characterization of the "speed" of the queueing system, while the variance var[*u*] and the tail probabilities Prob[*u* > *U*] can be used to estimate the so-called delay jitter, i.e., to estimate to what degree the queueing system introduces variability in the cell interdeparture times, for cells belonging to such services as voice or video. Curves showing the probability that the delay exceeds some given threshold *U* versus *U* can be used, for instance, to characterize the delay jitter of an ATM multiplexer or an ATM switching element, in terms of the 10^{-k} quantile of the delay, for some integer value of *k*, i.e., the value *U** such that Prob[delay > *U**] = 10^{-k}. From the relationship (16.34), all the important delay characteristics can be derived in terms of characteristics of the system contents.

10. FURTHER READING

The analysis of the GI-D-*c* model and its application in the derivation of buffer requirements of ATM switches with dedicated-buffer output queueing, in Sections 4 to 5, have been discussed largely along the lines of [5] and [6]. The case of output buffer sharing, where the separate output buffers described in Section 3 are replaced by one common buffer for all the destination groups together, is investigated in [4]. An approximate analysis of the end-to-end cell delay through a multistage switching network is presented in [17]. The approach taken is to approximate the arrival processes on the inlets of the consecutive stages as independent Bernoulli processes and to approximate the cell delays in these stages as independent random variables.

The analysis of the COR-D-1 model in Section 7 has been mainly based on [21]. Arbitrarily distributed on-periods have also been considered in [8], [14] and [15]. In [15], a geometric approximation is derived for the tail distribution of the system contents, where the coefficient of the geometric form is approximated by the multiplexer load. In [8], the queueing system is analyzed by numerically solving a set of balance equations. Finally, in [14], a heuristic approximation is derived for the distribution of the system contents.

The material presented here basically deals with the analysis of discrete-time single queues. An overview of recent work on closed discrete-time queueing network models with a product form equilibrium queue-length distribution and restricted batch size movement is given in [24]. The restricted batch sizes allow to model communication networks taking into account the restricted link capacities, finite capacities of switches and the finite number of switch outlets. In [11], a review is presented of cost-effective methodologies for the analysis of complex queueing network models of integrated networks. The methods are based on the information theoretic principle of maximum entropy, queueing theoretic concepts and batch renewal processes.

Some of the material presented in this paper was also included in the book by Bruneel and Kim on discrete-time queueing models ([3]). In addition, this book also contains a discussion of various ATM multiplexer models and extended treatments of (single-server) discrete-time queueing systems with general service-time distributions, with multiple customer classes, with nonindependent arrivals, or with random server interruptions. Comprehensive lists of references to the literature on various aspects of discrete-time queueing analysis (mainly in the area of digital communication systems and networks), along with short descriptions of the models treated in those references, can also be found there.

Although most of the classical books on queueing analysis available in the scientific literature today are mainly concerned with continuous-time models, it is worth mentioning here that recently a few new books have appeared on discrete-time queues. Apart from the aforementioned [3], we think that [19] and [23] are the most relevant ones. Also, the 1983 book by Hunter on discrete-time Markov chains and their applications ([9]) contains an extensive treatment of various discrete-time queueing models.

ACKNOWLEDGMENTS

The first author is postdoctoral fellow of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.).

REFERENCES

[1] P. Brown, S. Simonian, Perturbation of a periodic flow in asynchronous server, *Proc. PERFORMANCE '87*, Brussels, December 1987, pp. 89-112.

[2] H. Bruneel, Queueing behavior of statistical multiplexers with correlated inputs, *IEEE Transactions on Communications* **36** (1988) 1339-1341.

[3] H. Bruneel, B. G. Kim, *Discrete-Time Models for Communication Systems Including ATM* (Kluwer Academic Publishers, Boston, 1993).

[4] H. Bruneel, B. Steyaert, Buffer requirements for ATM switches with multiserver output queues, *Electronics Letters* **27** (1991) 671-672.

[5] H. Bruneel, B. Steyaert, E. Desmet, G. H. Petit, An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queues, *International Journal of Digital and Analog Communication Systems* **5** (1992) 193-201.

[6] H. Bruneel, B. Steyaert, E. Desmet, G. H. Petit, Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues, *European Journal of Operational Research* **76** (1994) 563-572.

[7] A. M. Eikeboom, H. C. Tijms, Waiting-time percentiles in the multi-server $M^X/G/c$ queue with batch arrivals, *Prob. in the Engin. and Inform. Sciences* **1** (1987) 75-96.

[8] K. Elsayed, On the superposition of discrete-time Markov renewal processes and application to statistical multiplexing of bursty traffic sources, *Proc. IEEE GLOBECOM '94*, San Francisco, November/December 1994, pp. 1113-1117.

[9] J. J. Hunter, *Mathematical Techniques of Applied Probability, Volume* 2, *Discrete Time Models : Techniques and Applications* (Academic Press, New York, 1983).

[10] L. Kleinrock, *Queueing Systems, Volume I: Theory* (Wiley, New York, 1975).

[11] D. Kouvatsos, Information theoretic methodologies for QNMs of ATM switch architectures, in : *Performance Evaluation and Application of ATM Networks* (Kluwer Academic Publishers, Boston, 2000), pp. 413-448.

[12] M. Neuts, *Structured Stochastic Matrices of M/G/1 Type and their Applications* (New York, Marcel Dekker Inc., 1989).

[13] M. Schwartz, *Computer-Communication Network Design and Analysis* (Prentice Hall, Englewood Cliffs, New Jersey, 1977).

[14] A. Simonian, J. Guibert, Large deviations approximation for fluid queues fed by a large number of on/off sources, *Proc. ITC 14,* Antibes Juan-les-Pins, June 1994, pp. 1013-1022.

[15] K. Sohraby, On the theory of general ON-OFF sources with applications in high-speed networks, *Proc. IEEE INFOCOM '93*, San Francisco, March 1993, pp. 401-410.

[16] B. Steyaert, H. Bruneel, An effective algorithm to calculate the distribution of the buffer contents and the packet delay in a multiplexer with bursty sources, *Proc. GLOBECOM '91*, Phoenix, December 1991, pp. 471-475.

[17] B. Steyaert, H. Bruneel, G. H. Petit, E. Desmet, End-to-end delays in multistage ATM switching networks : approximate analytic derivation of tail probabilities, *Computer Networks and ISDN Systems* **25** (1993) 1227-1241.

[18] B. Steyaert, H. Bruneel, On the performance of multiplexers with three-state bursty sources : analytical results, *IEEE Transactions on Communications* **43** (1995) 1299-1303.

[19] H. Takagi, *Queueing Analysis, A Foundation of Performance Evaluation, Volume 3 : Discrete-Time Systems* (North-Holland, Amsterdam, 1993).

[20] A. M. Viterbi, Approximate analysis of time-synchronous packet networks, *IEEE Journal on Selected Areas in Communications* **SAC-4** (1986) 879-890.

[21] S. Wittevrongel, H. Bruneel, Effect of the on-period distribution on the performance of an ATM multiplexer fed by on/off sources : an analytical study, *Proc. PCN '95*, Istanbul, October 1995, pp. 33-47.

[22] C. M. Woodside, E. D. S. Ho, Engineering calculation of overflow probabilities in buffers with Markov-interrupted service, *IEEE Transactions on Communications* **COM-35** (1987) 1272-1277.

[23] M. E. Woodward, *Communication and Computer Networks : Modelling with Discrete-Time Queues* (Pentech Press, London, 1993).

[24] M. E. Woodward, Discrete-time queueing networks and models for high speed communication networks, in : *Tutorial Papers of the Fifth IFIP Workshop on Performance Modelling and Evaluation of ATM Networks* (UK Performance Engineering Workshop Publishers, Ilkley, 1997) (ISBN : 0 9524027 4 2).

[25] Y. Xiong, H. Bruneel, Performance of statistical multiplexers with finite number of inputs and train arrivals, *Proc. of IEEE INFOCOM* '92, Firenze, May 1992, pp. 2036-2044.

[26] Y. Xiong, H. Bruneel, B. Steyaert, Deriving delay characteristics from queue length statistics in discrete-time queues with multiple servers, *Performance Evaluation* **24** (1996) 189-204.

BIOGRAPHIES

Sabine Wittevrongel was born in Gent, Belgium, in 1969. She received the M.S. degree in Electrical Engineering and the Ph.D. degree in Applied Sciences from Ghent University, Belgium, in 1992 and 1998, respectively. Since September 1992, she has been with the SMACS Research Group, Department of Telecommunications and Information Processing, Ghent University, first in the framework of various projects, and since October 1994, as a researcher of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.). Her main research interests include discrete-time queueing theory, performance evaluation of ATM and IP networks, and the study of traffic control mechanisms.

Herwig Bruneel was born in Zottegem, Belgium, in 1954. He received the M.S. degree in Electrical Engineering, the degree of Licentiate in Computer Science, and the Ph.D. degree in Computer Science in 1978, 1979 and 1984 respectively, all from Ghent University, Belgium. From 1979 to 1998, he has been a researcher of the Fund for Scientific Research - Flanders (Belgium) (F.W.O.) at Ghent University. He has also been a part time Professor in the Faculty of Applied Sciences at the same university from 1987 to 1998. Currently he is full time Professor and the head of the Department of Telecommunications and Information Processing. He also leads the SMACS Research Group within this department. His main research interests include stochastic modeling of digital communication systems, discrete-time queueing theory, and the study of ARQ protocols. He has published more than 150 papers on these subjects and is coauthor of the book *H. Bruneel and B. G. Kim, "Discrete-Time Modelsfor Communication Systems Including ATM"* (Kluwer Academic Publishers, Boston, 1993).

Chapter 17

AN INFORMATION THEORETIC METHODOLOGY FOR QNMs OF ATM SWITCH ARCHITECTURES*

Demetres Kouvatsos
Computer and Communication Systems
Modelling Research Group,
University of Bradford,
Bradford BD7 1DP,
West Yorkshire,
UK.

Abstract The performance modelling and quantitative analysis of Asynchronous Transfer Mode (ATM) switch architectures constitute a rapidly growing application area due to the their ever expanding usage and the multiplicity of their component parts together with the complexity of their functioning. However, there are inherent difficullties and open issues associated with the cost-effective evaluation of these systems before a global integrated broadband network infrastructure can be established. This is due to the need for derived performance metrics such as queue length and response time distributions, the complexity of traffic characterisation and congestion control schemes and the existence of multiple packet classes with space and time priorities under various blocking mechanisms and buffer management schemes.

Queueing network models (QNMs) are widely recognised as powerful and realistic tools for the performance monitoring and prediction of packet-switched computer communication systems. However, analytic solutions for QNMs are often hindered by the generation of large state spaces requiring further approximations and a considerable (or, even prohibitive) amount of computation. This tutorial paper highlights a cost-effective methodology for the exact and/or approximate analysis of some complex QNMs of ATM networks consisting of multi-

* Supported by the Engineering and Physical Sciences Research Council (EPSRC), UK, under grant GR/K/67809.

buffered, shared buffer, shared medium and space division switch architectures. The methodology has its roots on the information theoretic principle of Maximum Entropy (ME), queueing theoretic concepts and batch renewal traffic processes. Comments on further research work are included.

Keywords: ATM Switch Architectures, Queueing Network Models (QNMs), Repetitive Service Blocking with Random Destination (RS-RD), Maximum Entropy (ME), Partial Buffer Sharing (PBS), Head-of-Line (HoL), Batch Renewal Process, GGeo, GE, sGGeo.

1. INTRODUCTION

Over the recent years a considerable amount of effort has been devoted towards the design and development of Asynchronous Transfer Mode (ATM) switch architectures, the preferred packet-oriented solution of a new generation of high-speed communication systems for multimedia applications, both for public information highways to support Broadband Integrated Services Digital Networks (B-ISDNs) and for local and wide area private networks. The performance modelling and quantitative analysis of Asynchronous Transfer Mode (ATM) switch architectures constitute worldwide a rapidly growing application area due to the their ever expanding usage and the multiplicity of their component parts together with the complexity of their functioning. However, there are inherent difficullties and open issues associated with the cost-effective evaluation of these systems before a global integrated broadband network infrastructure can be established. This is due to the need for derived performance metrics such as queue length and response time distributions, the complexity of traffic characterisation and congestion control schemes and the existence of multiple packet classes with space and time priorities under various blocking mechanisms and buffer management schemes.

Traffic in B-ISDN is essentially discrete and basic operational parameters are known via measurements obtained at discretised points of time. Under this framework, the time axis is segmented into a sequence of time intervals (or slots) of unit duration corresponding to the elementary unit of time in the system. Arrivals and departures are allowed to occur at the boundary epochs of a slot, whilst during a slot no cells enter or leave the system. Performance results obtained in the discrete-time domain, however, should be compared with corresponding mixed and/or continuous-time statistics which are expected to provide good approximations and may be more appropriate in cases with large numbers of cells in each slot (c.f., fluid-flow approach [1]).

Emerging architectural designs for ATM switches have R input and R output ports and can be broadly classified into four main architectures, namely multi-buffered, shared buffer, space division and shared medium switches (c.f., [2]). Simple multi-buffered switches supply each output port with a dedicated memory. Shared buffer switches incorporate a single memory shared by all input and output ports. Space division switches are based on multistage interconnection networks (MINs) whose switching elements may or may not be buffered at their input and/or output ports. Shared medium switches have a common high speed medium such as a parallel bus, in which all arriving cells are synchronously multiplexed and then de-multiplexed into individual streams, one for each buffered output port. In all buffered ATM switch architectures, an arriving cell will be either lost or blocked, as appropriate, if it finds a full input/output buffer. Typical performance measures such as cell-loss and state probabilities, delay distribution, throughput and mean queue length (MQL), can be used to assess ATM switch performance.

Traditionally, queueing network models (QNMs) are widely recognised as powerful and realistic tools for the performance monitoring and prediction of packet-switched computer communication systems. However, earlier proposed QNMs for ATM switches and networks are not in general analytically tractable except in special cases. Usually it is necessary to resort to either simulation or numerical methods; simulation is time consuming and cannot easily yield the great precision needed for some rare events, such as cell loss, whilst numerical methods are often hindered by the generation of large state spaces requiring a considerable (or even prohibitive) amount of computation as the system size increases. Thus there has been a great need to consider alternative analytic methodologies for QNMs leading to both credible and cost-effective approximations for the performance prediction and optimisation of ATM switches and networks.

This tutorial paper highlights an information theoretic based methodology for the exact and/or approximate analysis of complex queues and QNMs and their applications into the performance modelling and evaluation of some ATM switch architectures with bursty and/or short range depedence (SRD) correlated traffic. The methodology has its roots on the principle of Maximum Entropy (ME) [3], queueing theoretic concepts and batch renewal traffic processes [4]. Note that the principle of ME has been used as probability method of inference, in conjuction with queueing theoretic mean value constraints, for the approximate analysis of single queueing systems and the queue-by-queue decomposition of arbitrary QNMs in both continuous-time and discrete-time domains [5-7]. Moreover, batch renewal processes provide the means of defining the effects of correlated traffic in queueing systems and its behaviour as it traverses the network, quite free of the need to commit to arbitrary assumptions on burst structure. Central to the tractability of the analysis, however, is the knowledge of the circumstances under which a simpler traffic process can be used to approximate, with a torerable accuracy, a more complex batch renewal process and, thus, facilitate understanding of how the superposition of arrival processes is shaped deep into the network.

The paper is divided into seven sections. The ME formalism is introduced in Section 2. Section 3 defines a batch renewal process and describes the shifted Generalised Geometric (sGGeo) batch renewal process as a model of external SRD traffic exhibiting geometrically declined count and interval covariances. This section also devises the approximation of the sGGeo process by an ordinary Generalised Geometric (GGeo) process, based on the matching of the first two GGeo moments of counts to those of sGGeo. Moreover, it presents the GGeo-type two moment flow approximation formulae for corresponding merging, splitting and departing streams within the network and suggests the Generalised Exponential (GE) distribution as an alternative to GGeo process within a mixed (discrete/continuous) time domain. Section 4 reviews a ME product-form approximation for the performance evaluation of a simple multi-buffered ATM switch architecture with output port queueing together with a related queue-by-queue decomposition algorithm for arbitrary discrete-time open QNMs with external sGGeo-type traffic, single deterministic (D) servers, departures first (DF) or arrival first (AF) buffer management simultaneity policies and repetitive service blocking with random destination (RS-RD) mechanism. Section 5 presents an extended ME solution for a stable queue with C (C>1) priority classes, head-of-line (HoL) scheduling discipline and partial buffer sharing (PBS) scheme under AF and/or DF buffer management simultaneity policies. Moreover, it highlights its applicability, as a efficient building-block within a queue-by-queue decomposition algorithm, for the priority congestion control of an arbitrary multiple class QNM of a multi-buffered ATM switching network with bursty and/or SRD external arrivals and space/time priorities. Section 6 gives brief accounts of entropy maximisation and performance modelling aspects of shared buffer, space-division and shared medium ATM switch architectures with bursty and/or SRD external traffic and RS-RD blocking, as appropriate. Concluding comments follow in Section 7.

Remarks: i) Buffer management policies for discrete time queues stipulate how a buffer is filled or emptied in the case of simultaneous bulk arrivals and departures at a boundary epoch of a slot. In such cases, according to DF policy, departures take precedence over arrivals, while under arrivals first (AF) policy the opposite effect is observed (see Fig. 17.1). Buffer management policies can play a significant role in the determination of blocking probabilities in discrete time finite capacity queues.

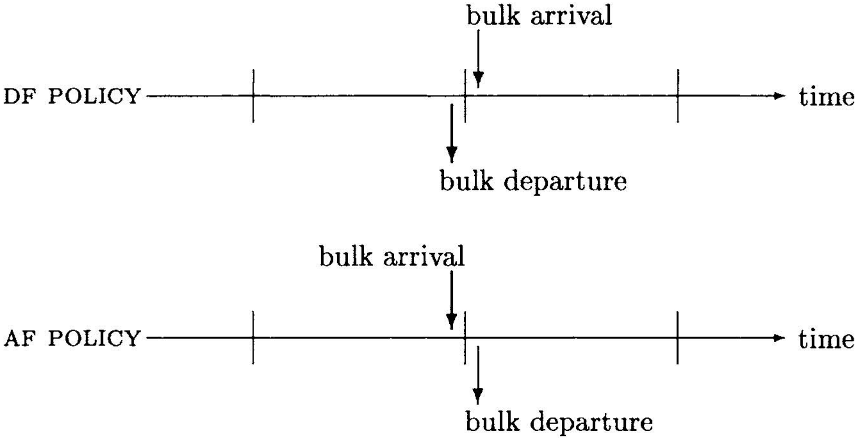


Figure 17.1 Effects of AF and DF buffer management policies at slot boundary epoch

ii) One of the most important blocking mechanisms with many applications to telecommunication systems is that of the repetitive service blocking with either random (RS-RD) or fixed (RS-FD) destination4. This type of blocking occurs when a job upon service completion at a queue i attempts to join a destination queue j, whose capacity is full. As a result, the job is rejected by queue j and immediately receives another service at queue i. This process is repeated until the job completes service at queue i at the moment where the destination queue is not full. Under the RS-RD blocking mechanism, each time the job completes service at queue i, a destination queue is selected independently of the previously chosen destination queue j . Under the RS-FD blocking mechanism, each time the job completes service at queue i, the same destination queue j is chosen.

2. ME FORMALISM

Consider a system Q that has a set of possible discrete states $\mathbf{S}$ = $(S_0, S_1, S_2, ...)$ which may be finite or countable infinite and state S_n, n = 0,1,2,... may be specified arbitrarily. Suppose the available information about Q places a number of constraints on $P(S_n)$, the probability distribution that the system Q is in state S_n. Without loss of generality, it is assumed that these take the form of mean values of several suitable functions $\{f_1(S_n), f_2(S_n), \ldots, f_m(S_n)\}$, where m is less than the number of possible states. The principle of maximum entropy (ME) [3,5] states that, of all distributions satisfying the constraints supplied by the given information, the minimally prejudiced distribution $P(S_n)$ is the one that maximises the system's entropy function:

$$H(P) = -\sum_{S_n \in \mathbf{S}} P(S_n) \log P(S_n) \qquad (17.1)$$

subject to the constraints:

$$\sum_{S_n \in \mathbf{S}} P(S_n) \; = \; 1 \qquad (17.2)$$

$$\sum_{S_n \in \mathbf{S}} f_k(S_n) P(S_n) \; = \; \langle f_k \rangle \, , \; \forall k = 1, 2, \ldots, m \qquad (17.3)$$

where $\{\langle f_k \rangle\}$ are the prescribed mean values defined on the set of functions $\{f_k(S_n)\}$, $k = 1,2,\ldots,m$, where m is less than the number of states in S. The maximisation of (17.1), subject to the constraints (17.2) and (17.3), can be carried out using Lagrange's method of undermined multipliers and leads to the solution

$$P(S_n) = \frac{1}{Z} \exp \left\{ -\sum_{k=1}^{m} \beta_k f_k(S_n) \right\} \qquad (17.4)$$

where $\{\beta_k, \, k = 1, 2, \ldots, m\}$ are the Lagrangian multipliers determined from the set of constraints (17.3) and $Z = exp(\beta_0)$, is the nomalising constant with β_0 being the Lagrangian multiplier determined by the normalisation constraint (17.2).

Note that if Q has a finite number of discrete states and only the normalisation constraint (17.2) is known, then the ME solution reduces to a uniform distribution.

In an information theoretic context, the ME solution corresponds to the maximum disorder of system states, and thus is considered to be the least biased distribution estimate of all solutions that satisfy the system's constraints. In sampling terms, if the prior information includes all constraints actually operative during a random experiment, the distribution predicted by the ME can be realised in overwhelmingly more ways than by any other distribution (c.f., [3]). In formal terms, the principle of ME may be seen as an information operator "o" which takes two arguments, a prior uniform distribution q and a new constraint information I of the form (17.1) and (17.2), yielding a posterior ME distribution p, i.e.,

$$p = qoI \qquad (17.5)$$

The maximisation of $H(p)$ uniquely characterises distribution p, satisfying four consistency inference criteria [8]. In particular, it has been shown that the ME solution is a uniquely correct distribution and that any other functional used to implement operator "o" will produce the same distribution as the entropy functional, otherwise it will be in conflict with the consistency criteria. In the field of systems modelling, expected values of various performance distributions of interest, such as the number of jobs in each resource queue concerned, are often known, or may be explicitly derived, in terms of moments of interarrival and service time distributions. Hence, the method of entropy maximisation may be applied to characterise useful information theoretic approximations of performance distributions of queueing systems and networks.

3. THE GGeo-TYPE APPROXIMATION OF THE sGGeo BATCH RENEWAL PROCESS

An arbitrary (persistent) discrete time arrivals traffic process may be described by the two-dimensional process $\{\xi(s), \kappa(s) : s \in \mathbb{Z}\}$ in which $\kappa(s)$ is the number of arrivals in the s^{th} (non-empty) batch and $\xi(s)$ is the interval (i.e. number of slots) between the $(s\text{–}1)^{\text{th}}$ batch and the s^{th} batch. A wide sense stationary process $\{\xi(s), \kappa(s) : s \in \mathbb{Z}\}$ is called a *batch renewal process* if both the $\{\kappa(s) : s \in \mathbb{Z}\}$ are independent and identically distributed (*iid*) and the $\{\xi(s) : s \in \mathbb{Z}\}$ are *iid* [4,9,10] Thus, a batch renewal process is completely defined by two constituent distributions which, by convention, are written

$$P[\xi(s) = t] \quad = \quad a(t) \tag{17.6}$$
$$P[\kappa(s) = n] \quad = \quad b(n) \tag{17.7}$$

Alternative description of the traffic may be in terms of the counts $\{c(t) : t \in \mathbb{Z}\}$, in which $c(t) \in \mathbb{N}_0$ is the number of arrivals (zero or more) at the t^{th} epoch, or in terms of the intervals $\{x(n) : n \in \mathbb{Z}\}$ between individual arrivals, in which $x(n) \in \mathbb{N}_0$ is the number of slots (possibly zero, as when there be two or more arrivals at an epoch) between the $(n\text{–}1)^{\text{th}}$ and n^{th} individual arrivals.

It is known that, given only measures of the correlation of counts and of the correlation between intervals, the *least biased choice* of process is a batch renewal process [4,10]. Consequently, batch renewal processes enable unbiased investigation into effects of traffic correlation and provide the reference basis against which other types of models may be compared [11-13].

3.1 THE sGGeo TRAFFIC PROCESS

The sGGeo traffic process (c.f., [4,9,10]) is the simplest non-trivial batch renewal process whose

- both constituent distributions (i.e. of intervals between batches $a(.)$ and of batch sizes $b(.)$) have an sGGeo form, namely

$$\mathsf{P}\big[\xi(s)=t\big]=a(t)=\begin{cases}1-\sigma & t=1\\ \sigma\tau(1-\tau)^{t-2} & t=2,3,\ldots \qquad 0\le\sigma,\tau\le 1\\ 0 & \text{otherwise}\end{cases} \tag{17.8}$$

$$\mathsf{P}\big[\kappa(s)=n\big]=b(n)=\begin{cases}1-\eta & n=1\\ \eta\nu(1-\nu)^{n-2} & n=2,3,\ldots \qquad 0\le\eta,\nu\le 1\\ 0 & \text{otherwise}\end{cases} \tag{17.9}$$

- both count correlation and interval correlation decline geometrically with lag $\ell=1,2,\ldots$

$$\mathsf{Cov}\big[c(t),c(t+\ell)\big]=\lambda^2(a-1)\beta_a{}^\ell \tag{17.10}$$

$$\mathsf{Cov}\big[x(n),x(n+\ell)\big]={}^1\!/\!_{\lambda^2}(b-1)\beta_b{}^\ell \tag{17.11}$$

where

$$\beta_a=1-\sigma-\tau,\quad a=\mathsf{E}\big[\xi(s)\big]=1+{}^\sigma\!/\!_\tau \tag{17.12}$$

$$\beta_b=1-\eta-\nu,\quad b=\mathsf{E}\big[\kappa(s)\big]=1+{}^\eta\!/\!_\nu \tag{17.13}$$

$$\lambda=\mathsf{E}\big[c(t)\big]=\frac{1}{\mathsf{E}\big[x(n)\big]}=\frac{\mathsf{E}\big[\kappa(s)\big]}{\mathsf{E}\big[\xi(s)\big]}={}^b\!/\!_a \tag{17.14}$$

Remarks: The sGGeo process is completely defined by just four parameters (σ,τ,η,ν) or equivalently, in terms of the corresponding set of correlation functions (as might result from measurements of real traffic). Note that the sGGeo process is applicable as a model of a traffic source at an ATM switch of which traffic measurements (e.g., of a workstation) are given only in terms of the first two moments of message size and the first two moments of the intervals between initiation of messages. It is also the least biased choice of a traffic model characterised by measures of correlation alone, for which either the measured covariances are geometric (c.f., eq.s (17.10), (17.11) etc) or there may be so few measurements that the best procedure is to fit the straight line to the plot of the logarithms of measured covariances against lags.

3.2 MATCHING GGeo TO sGGeo ON COUNTS

The marginal distribution of counts of the sGGeo process is clearly given by

$$\mathsf{P}\big[c(t)=0\big] = 1 - \frac{1}{\mathsf{E}\big[\xi(s)\big]} = 1 - \frac{1}{a} \tag{17.15}$$

and, for $n = 1,2,\ldots$, by

$$\begin{aligned}\mathsf{P}\big[c(t)=n\big] &= \mathsf{P}\big[c(t)=n \mid c(t)>0\big]\cdot\mathsf{P}\big[c(t)>0\big]\\ &= \mathsf{P}\big[\kappa(s)=n\big]\cdot\frac{1}{a} = \frac{b(n)}{a}\end{aligned} \tag{17.16}$$

Therefore

$$\mathsf{E}\big[c(t)\big] = {}^{b}\!/_{a} = \lambda \tag{17.17}$$

$$\mathsf{E}\big[c(t)^2\big] = \mathsf{E}\big[\kappa(s)^2)\big]\cdot\frac{1}{a} = \lambda\left(2\frac{b-1}{1-\beta_b}+1\right) \tag{17.18}$$

For the trivial sGGeo process, in which there is no traffic correlation (neither of counts nor of intervals) so that $\beta_a = 0$ and $\beta_b = 0$, let

$$\begin{aligned}&\sigma \xrightarrow{become} 1-\tau'\\ &\tau \xrightarrow{become} \tau'\\ &\eta \xrightarrow{become} 1-\nu'\\ &\nu \xrightarrow{become} \nu'\end{aligned}$$

(where the primes are to distinguish the parameters from those of the nontrivial sGGeo). Then equations (17.8), (17.9) and (17.15) become

$$\mathsf{P}\big[\xi(s)=t\big] = \tau'(1-\tau')^{t-1} \qquad t=1,2,3,\ldots \tag{17.19}$$

$$\mathsf{P}\big[\kappa(s)=n\big] = \nu'(1-\nu')^{n-1} \qquad n=1,2,3,\ldots \tag{17.20}$$

$$\mathsf{P}\big[c(t)=n\big] = \begin{cases}1-\tau' & n=0\\ \tau'\nu'(1-\nu')^{n-1} & n=1,2,3,\ldots\end{cases} \tag{17.21}$$

where

$$\tau' = \lambda\nu' \tag{17.22}$$

The distribution of counts (17.21) is an ordinary GGeo and, because both counts are *iid* and intervals are *iid*, the interarrival distribution is given by

$$\mathsf{P}[x(n) = t] = \begin{cases} 1-\nu' & t = 0 \\ \tau'\nu'(1-\tau')^{t-1} & t = 1, 2, 3, \ldots \end{cases} \tag{17.23}$$

which is also GGeo (c.f., Fig. 17.2).

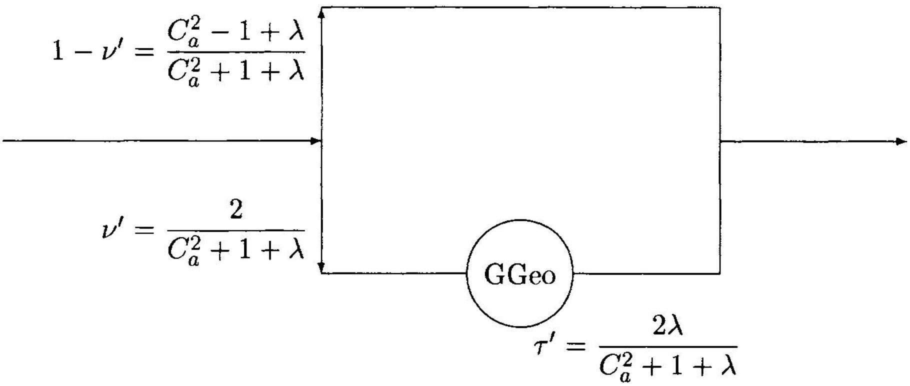

$1 - \nu' = \dfrac{C_a^2 - 1 + \lambda}{C_a^2 + 1 + \lambda}$
$\nu' = \dfrac{2}{C_a^2 + 1 + \lambda}$
GGeo
$\tau' = \dfrac{2\lambda}{C_a^2 + 1 + \lambda}$

Figure 17.2 The GGeo Distribution with parameters ν' and τ' $(0 < \nu', \tau' \leq 1)$

Clearly, for the GGeo process the following relationships hold:

$$a' = 1/\tau' \tag{17.24}$$
$$b' = 1/\nu' \tag{17.25}$$
$$\lambda' = \mathsf{E}[c(t)] = \tau'/\nu' \tag{17.26}$$
$$\mathsf{E}[c(t)^2] = \lambda' \, (2/\nu' - 1) \tag{17.27}$$

Therefore, the GGeo process which matches the shifted GGeo process on the first two moments of counts has

$$\lambda' = \lambda \tag{17.28}$$
$$\lambda' \, (2/\nu' - 1) = \lambda \left(2\frac{b-1}{1-\beta_b} + 1 \right) \tag{17.29}$$

giving

$$\nu' = \frac{1-\beta_b}{(b-1) + (1-\beta_b)}, \tag{17.30}$$

Note that, the interarrival time squared coefficient of variation (SCV) of the GGeo process (17.23), C_a^2, can be expressed as

$$C_a^2 = \frac{2 - \nu'(\lambda' + 1)}{\nu'} \qquad (17.31)$$

Remarks: The choice of the GGeo distribution is further motivated by the fact that measurements of actual traffic or service time may be generally limited and so only few parameters can be computed reliably. Typically, only the mean and variance may be relied upon. In this case, the choice of distribution which implies least biased (i.e., introduction of arbitrary and, therefore, false assumptions) within a discrete-time domain is that of GGeo type distribution. In an ATM environment, this model is directly applicable in cases of traffic with low level of correlation or where smoothing schemes are introduced at the adaptation level (e.g., for a stored video source) with the objective of minimising or even eliminating the problem of traffic correlation. For $\mu = 1$ and $C_s^2 = 0$, the GGeo distribution reduces to a proper D distribution. Note that in a mixed time domain, the GGeo process corresponds to the Generalised Exponential (GE) distribution (c.f., [5]) depicted in Fig. 17.3.

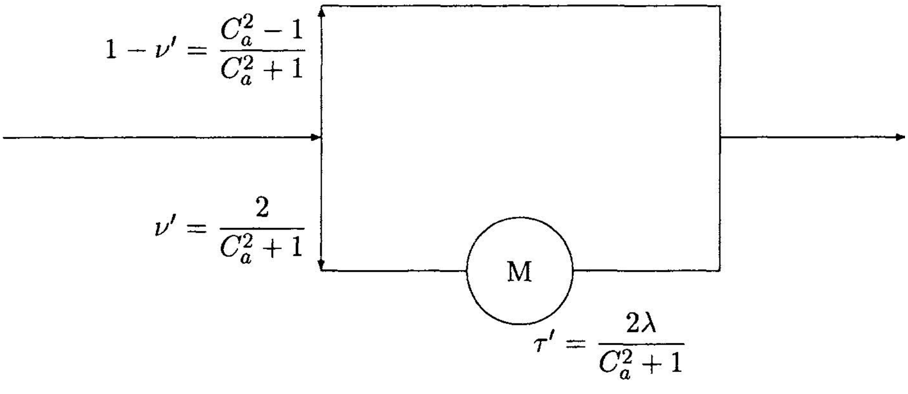

$1 - \nu' = \frac{C_a^2 - 1}{C_a^2 + 1}$
$\nu' = \frac{2}{C_a^2 + 1}$
M
$\tau' = \frac{2\lambda}{C_a^2 + 1}$

Figure 17.3 The GE Distribution with parameters ν' and τ' $(0 \leq \nu' \leq 1)$

3.3 GGeo-TYPE FLOW FORMULAE

This section presents the GGeo-type two moment flow approximation formulae for open discrete time QNMs with arbitrary configuration (c.f., [6]). The superposition process of M GGeo(λ_i, C_{ai}^2) interarrival times, $i = 1, 2, \ldots, M$, can be approximated by a GGeo(λ, C_a^2) process, whose rate and SCV are given by (c.f., Fig. 17.4)

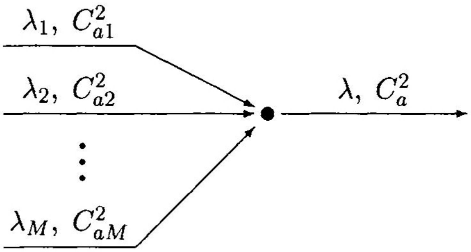


Figure 17.4 The merging process

$$\lambda = \sum_{i=1}^{M} \lambda_i \tag{17.32}$$

$$C_a^2 = \left\{ \sum_{i=1}^{M} \frac{\lambda_i}{\lambda} \left(C_{ai}^2 + \lambda_i + 1 \right)^{-1} \right\}^{-1} - \lambda - 1 \tag{17.33}$$

Furthermore, the mean rate, λ, and SCV, C_d^2, of the interdeparture time distribution of a stable GGeo/D/1 queue (with infinite capacity and $\mu = 1$) can be determined by (c.f., Fig. 17.5)

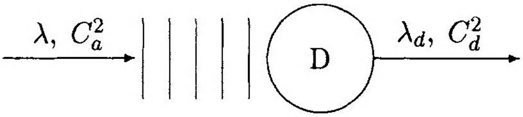


Figure 17.5 The interdeparture process

$$\lambda_d = \lambda \tag{17.34}$$

$$C_d^2 = C_a^2(1 - \lambda) - \lambda(1 - \lambda) \tag{17.35}$$

Finally, it is assumed that the interdeparture process of a stable GGeo/D/1 queue is approximated by a GGeo(λ, C_d^2) process which decomposes into M sub-processes with splitting probabilities $\{p_i\}$ and parameters $\{\lambda_i, C_{di}^2\}$, i = 1,2, ..., M. In this context, each split process clearly conforms exactly with a GGeo process with parameters (c.f., Fig. 17.6)

$$\lambda_i = p_i \lambda \tag{17.36}$$

$$C_{di}^2 = 1 + p_i \left(C_d^2 - 1 \right) \tag{17.37}$$

Note that similar GE-type two moment flow approximation formulae can be found in [5].

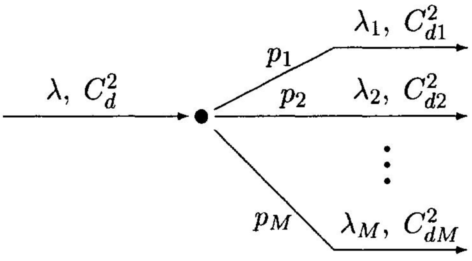


Figure 17.6 The split process

4. MULTI-BUFFERED ATM SWITCH ARCHITECTURES

This section focuses on a simple multi-buffered ATM switch architecture consisting of multiple output ports each of which has a dedicated buffer capacity of fixed size. A cell finding on arrival a full buffer will be lost. An example of this architecture is the NCX-1E6 ATM multiservice switch developed by ECI Telematics international Ltd [14].

4.1 A QUEUEING MODEL OF A MULTI-BUFFERED SWITCH

Consider a FCFS queueing model of a multi-buffered switch with external sGGeo-type arrivals and R× R (R > 1) input/output ports, D transmission times with rates $\{\mu_i = 1$ cell per slot, $i = 1, 2, \dots, R\}$, AF and/or DF policies and output port queueing depicted in Fig. 17.7. Matching a GGeo distribution to sGGeo traffic process on counts, the model is denoted by $\Pi_{R\times R}$(GGeo/D/1)/**K**, where **K** = (K$_1$,K$_2$,... ,K$_R$), K$_i$ is the finite capacity of output port queue $i, i = 1, 2, \dots, R$ and the superposition (or merging) of R GGeo-type interarrival streams at each of the R output ports is approximated by a GGeo distribution with overall parameters $(\lambda_i, C_{a\,i}^2)$, $i = 1, 2, \dots, R$.

Let at any given time the state of the system be represented by a vector $\mathbf{n} = (n_1,n_2,\dots,n_R)$, where n_i is the number of cells of each queue, i, $i = 1,2,\dots, R$ and S(**K**,R) be the set of states defined by

$$S(\mathbf{K}, R) = \{\mathbf{n} = (n_1, n_2, \dots, n_R)/0 \leq n_i \leq K_i,\ i = 1, 2, \dots, R\}$$

Moreover, let $p(\mathbf{n})$, $\mathbf{n} \in S(\mathbf{K}, R)$ and $p_i(n_i)$, $i = 1,2,\dots, R$, be the joint and marginal state probabilities (or queue length distributions-QLDs), respectively.

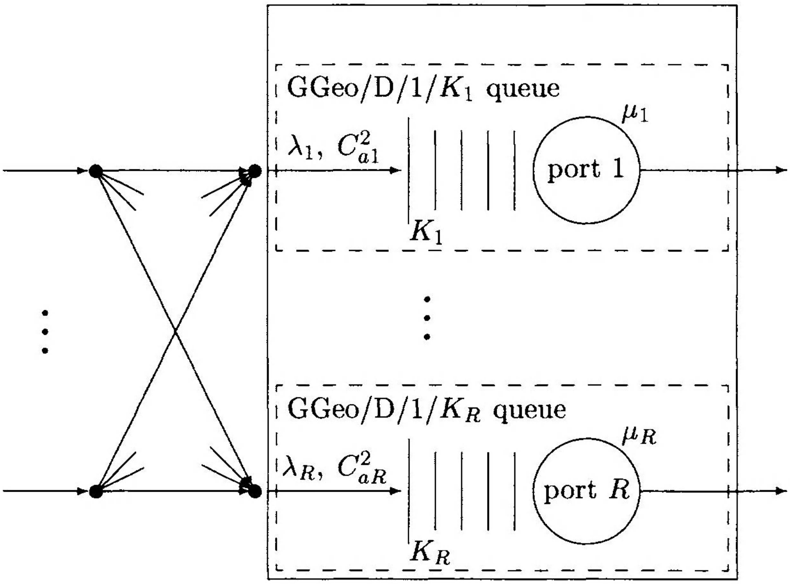


Figure 17.7 The $\Pi_{R\times R}$(GGeo/D/1)/**K** multi-buffered queueing model

The form of the ME solution $p(\mathbf{n})$ of a $\Pi_{R\times R}(\mathrm{GGeo/D/1})/\mathbf{K}$ queueing system can be characterised by maximizing the entropy functional $H(p) = -\sum_{\mathbf{n}} p(\mathbf{n}) \log p(\mathbf{n})$, subject to normalisation and the marginal constraints of server utilisation, $U_i\,(0 < U_i < 1)$, MQL, $L_i\,(U_i \leq L_i < K_i)$ and full buffer state probability, $\phi_i\,(0 < \phi_i < 1)$, and is given by

$$p(\mathbf{n}) = \prod_{i=1}^{R} \frac{1}{Z_i} g_i^{s_i(\mathbf{n})} x_i^{n_i} y_i^{f_i(\mathbf{n})}, \forall \mathbf{n} \in S(\mathbf{K}, R), \tag{17.38}$$

where $Z = \prod_i^R Z_i$ is the normalising constant, $\{s_i(\mathbf{n}), f_i(\mathbf{n})\}$ are suitable indicator functions and $\{g_i, x_i\, y_i,\ i = 1,2,\ldots, R\}$ are the Lagrangian coefficients corresponding to the constraints $\{U_i, L_i, \phi_i,\ i = 1, 2,\ \ldots, R\}$, respectively. Clearly, each of the terms of the product in (17.38) can be interpretted as the marginal ME solution $\{p_i(n_i),\ n_i = 0, 1, \ldots, K_i\}$ of a stable FCFS GGeo/D/1/K_i queueing model, i = 1,2, ..., R. Thus, (17.38) can be re-written as

$$p(\mathbf{n}) = \prod_{i=1}^{R} p_i(n_i) \tag{17.39}$$

where

$$p_i(n_i) = \begin{cases} \frac{1}{Z_i} & n_i = 0 \\ \frac{1}{Z_i} g_i x_i^{n_i}, & n_i = 0, 1, \ldots, K_i - 1 \\ \frac{1}{Z_i} g_i x_i^{K_i} y_i, & n_i = K_i \end{cases} \tag{17.40}$$

$$\tag{17.41}$$

It can be shown (c.f., [7]) that the Lagrangian coefficients $\{g_i, x_i,\ i = 1, 2, \ldots, R\}$ are invariant with respect to K_i and can be determined exactly by

$$g_i = \frac{\lambda_i(1 - x_i)}{x_i(1 - \lambda_i)}, x_i = \frac{L_{i,\infty} - \lambda_i}{L_{i,\infty}} \tag{17.42}$$

$$L_{i,\infty} = \frac{\lambda_i}{2}(1 + \frac{C_{a\,i}^2}{1 - \lambda_i}) \tag{17.43}$$

where $L_{i,\infty}$ can be interpretted, when $\lambda_i < 1$, as the exact MQL of the stable GGeo/D/1 queue with infinite capacity. Finally, the Lagrangian coefficients $\{y_i,\ i = 1, 2, \ldots, R\}$ can be computed by making use of the flow balance condition

$$\lambda_i(1 - \pi_i) = U_i,\ i = 1, 2, \ldots, R, \tag{17.44}$$

and are expressed by [7]

$$y_i = \begin{cases} 1 & \text{DF Policy} \\ \frac{1-\lambda_i}{1-x_i} + (1 - \nu_i')\left(\frac{\lambda_i}{1-\nu_i'-x_i} - \frac{1-\lambda_i}{1-x_i} + \right. & \\ \left. \frac{\lambda_i}{1-\nu_i'-x_i}\left(\frac{1-\nu_i'}{x_i}\right)^{K_i-1}\right) & \text{AF Policy} \end{cases} \tag{17.45}$$

where $\{\pi_i,\ i = 1, 2, \ldots, R\}$ are the cell loss (or blocking) probabilities that an arriving cell finds GGeo/D/1/K_i, i = 1,2, ...,R queue at full capacity and -by applying GGeo-type probabilistic arguments- they can be determined by

$$\pi_i = p_{a\,i}(0)(1 - \nu_i')^{K_i} + \sum_{n_i=1}^{K_i} ((1 - \nu_i')^{K_i - n_i} p_{a\,i}(n_i), \tag{17.46}$$

where

$$p_{a\,i}(n_i) = \begin{cases} p_i(0) + p_i(1) & n_i = 0 \\ p_i(n_i + 1) & n_i = 1, 2, \ldots, K_i - 1 \end{cases} \tag{17.47}$$

A similar analysis for a $\Pi_{R\times R}$(GE/D/1)/**K** queueing system, based on the generic form of ME solution (17.38) can be seen in [5].

4.2 NETWORKS OF MULTI-BUFFERED ATM SWITCHES

Consider at equilibrium an arbitrary open QNM with M (a multiple of R) single deterministic server queueing stations, sGGeo external interarrival times, FCFS scheduling discipline, AF and/or DF buffer management simultaneity policies and RS-RD blocking mechanism. This network configuration can be broadly considered as a QNM of [M/R] multibuffered ATM switches with output port queueing where each station k, $k = 1,2,\ldots,M$ represents an output port. Let $\{p_{km}, k,m = 1,\ldots,M\}$ be the transition probability (first order Markov chain) that a cell transmitted from station k attempts to join station m, $\{p_{k0}, k = 1,2,\ldots,M\}$ be the transition probability that a cell leaves the network upon finishing transmission at station k. By applying the GGeo-type approximation of an external sGGeo-type arrival proces with correlation parameters $\{\beta_{a0k}, a_{0k}, \beta_{b0k}, b_{0k}, k = 1,2,\ldots,M\}$, let λ_{0k} be the mean arrival rate and C^2_{a0k} be the SCV of the external GGeo-type interarrival process of cells at station k. At any given time the state of the entire network is represented by $\mathbf{n} = (n_1, n_2, \ldots, n_M)$, where n_i is the number of cells ar queue i, $i = 1,2,\ldots, M$.

The ME solution of the joint state probability, $p(\mathbf{n})$, of the QNM, subject to normalisation and the marginal constraints of utilisation, MQL and full buffer state probability, can be simply described by the product-form approximation (17.39)-(17.47) with R replaced by M. Clearly, the ME solution implies a queue-by-queue decomposition algorithm for the approximate analysis of the entire network.

A sketch of the ME algorithm is described below. The algorithm executes the matching of external GGeo to external sGGeo traffic on counts and assumes that the arrival process at each queueing station conforms with a GGeo distribution. The GGeo/D/1/K$_k$ queue, in conjunction with the GGeo-type flow formulae (17.32)-(17.37), plays the role of a cost-effective building block in the solution process. The algorithm incorporates the computational process of solving iteratively the nonlinear equations for $\{\pi_{0k}, \pi_{km}, k,m = 1,2,\ldots,M\}$ (c.f., (17.46) - (17.47)) where π_{0k} is the blocking probability that an external arrival is blocked by station k and π_{km} is the blocking probability that a cell following its transmission from station k will be blocked by station m. In particular, the probabilities $\{\pi_{0k}, \pi_{km}\}$ generally depend on the effective job flow balance equations for $\left\{\hat{\lambda}_{0k}, \hat{\lambda}_k\right\}$, effective flow transition probabilities $\{\hat{p}_{km}\}$, effective interarrival time SCV, $\hat{C}^2_{a\,k}$, effective transmission time parameters $\left\{\hat{\mu}_k, \hat{C}^2_{s\,k}\right\}$ and overall interarrival time parameters $\{\lambda_k, C^2_{a\,k}\}$.

Begin

Input Data

- M,
- For $k = 1, 2, \ldots, M$, $m = 0, 1, \ldots, M$

$$\left\{K_k, \lambda_{0k}, C^2_{a\,0k}, \mu_k, C^2_{s\,k}, a_{km}\right\}, \{\beta_{a\,0k}, a_{0k}, \beta_{b\,0k}, b_{0k}\},$$

Step 1 Matching GGeo to external sGGeo traffic on counts

$C^2_{a\,0k} = \frac{1-\beta_{b\,0k}}{b_{0k}-1+(1-\beta_{b\,0k})}$;

Step 2 Initialize π_{0k}, & π_{km} to any value in (0,1), $\forall\, k, m = 1, 2, \ldots, M$,

Step 3 Solve the system of nonlinear equations $\{\pi_{0k}, \pi_{km}, k, m = 1, 2, \ldots, M\}$ under AF and/or DF buffer management simultaneity policies, as appropriate;

Step 3.1 Calculate effective flow transition probabilities $\{\hat{p}_{km}\}$:

$$\hat{p}_{km} = p_{km}(1 - \pi_{km})/(1 - \pi_{ck}),$$
$$\hat{p}_{k0} = p_{k0}/(1 - \pi_{ck}),$$
$$\pi_{ck} = \sum_{k \neq m = 1}^{M} p_{km}\pi_{km};$$

where π_{ck} is the blocking probability that a departing cell from station k will be blocked by a downstream station.

Step 3.2 Calculate effective customer flow balance equations:

$$\hat{\lambda}_{0k} = \lambda_{0k}(1 - \pi_{0k}),$$
$$\hat{\lambda}_k = \lambda_{0k} + \sum_{m=1}^{M} \hat{p}_{mk}\hat{\lambda}_m;$$

Step 3.3 Calculate the effective transmission time parameters, $\left\{\hat{\mu}_k, \hat{C}^2_{s\,k}\right\}$,

$$\hat{\mu}_k = (1 - \pi_{ck}),$$
$$\hat{C}^2_{s\,k} = \pi_{ck};$$

(n.b., the effective transmission time can be modelled by a GGeo $(\hat{\mu}_k, \hat{C}^2_{s\,k})$ or GE $(\hat{\mu}_k, \hat{C}^2_{s\,k})$ distribution, as appropriate)

```
        Step 3.4 Calculate overall interarrival parameters, $\{\lambda_k, C^2_{a\,k}\}$
            $\lambda_k = \hat{\lambda}_k/(1-\pi_k)$,
            $C^2_{a\,k} = \frac{\hat{C}^2_{a\,k}-\pi_k}{1-\pi_k}$,
            $\pi_k = \frac{\lambda_{0k}\pi_{0k} + \sum_{m=1}^{M}\left(\hat{p}_{mk}\hat{\lambda}_{mk}\pi_{mk}/(1-\pi_{mk})\right)}{\lambda_{0k} + \sum_{m=1}^{M}\left(\hat{p}_{mk}\hat{\lambda}_{m}/(1-\pi_{mk})\right)}$,
            where $\pi_k$ is the blocking probability that an arriving cell is
                blocked by station $k$;
        Step 3.5 By applying the Newton Raphson method, obtain new
            values for blocking probabilities, $\{\pi_{0k}, \pi_{mk}, k, m = 1, 2, \dots, M\}$,
            based on the generic expressions (17.46)-(17.47);
    Step 4 For $k = 1, 2, \dots, M$
        Step4.1 Calculate interdeparture time parameters
            $\{\lambda_{dk}, C^2_{dk}, k = 1, 2, \dots, M\}$ (c.f., (17.34)-(17.35));
        Step 4.2 Calculate the splitting of the interdeparture time parameters
            $\{p_{km}\lambda_{dk}, C^2_{d\,km}, k = 1, 2, \dots, M, m = 0, 1, \dots, M\}$
            (c.f., (17.49)-(17.37));
        Step 4.3 Calculate new value for overall interarrival parameters
            $\{\lambda_k, C^2_{a\,k}, k = 1, 2, \dots, M\}$ (c.f., (17.32)-(17.33));
        Step 4.4 Return to Step 3 until convergence of $C^2_{a\,k}$;
    Step 5 For $i = 1, 2, \dots, M$, obtain performance metrics of interest by
        solving each queueing station of the network as a stable FCFS
        GGeo/GGeo/1/K$_k$ queue (c.f., [13]) with overall interarrival time
        parameters, $(\lambda_k, C^2_{a\,k})$ and effective transmission time parameters
        $(\hat{\mu}_k, \hat{C}^2_{s\,k})$, $k = 1, 2, \dots, M$;
End
```

The main computational cost of the proposed algorithm is of $O\{k\Omega^3\}$, where k is the number of iterations in step 3 and Ω^3 is the number of operations for inverting the associated Jacobian matrix of the system of nonlinear eq.s $\{\pi_{0k}, \pi_{km}\}$. However, if a quasi-Newton numerical method is employed, this cost can be reduced to be of $O\{k\Omega^2\}$. Moreover, the existence and unicity of the solution of the nonlinear system of Step 3.5 cannot be shown analytically due to the complexity of the expressions of the customer loss (or blocking) probabilities $\{\pi_{0k}, \pi_{km}\}$; nevertheless, numerical instabilities were never observed during extensive experimentations under any feasible set of initial values.

A comprehensive comparative study involving the numerical validation of the ME decomposition algorithm against simulation and an investigation into the effect of varying degree of correlation of external sGGeo-type traffic on network performance can be seen in [15,16]. Further studies involving the ME algorithm as applied to ordinary GGeo-type and also GE-type QNMs have been reported in [5] and [7], respectively.

5. ATM NETWORKS WITH PRIORITY CONGESTION CONTROL MECHANISM

Finite buffer queues and network models with time and space priorities are of great importance towards effective congestion control mechanism and quality of service (QoS) protection in Asynchronous Transfer Mode (ATM) networks.

An ATM cell can be either of high or low priority depending on whether the cell loss priority (CLP) bit in the cell's header has been set or not. A cell of high priority has by default its CLP bit set to zero. The CLP bit of low priority cells is set to one. It is the job of the priority mechanism to monitor the CLP bit of arriving cells and give preferential treatment to high priority cells. Priority mechanisms include time priorities and space priorities.

Time priority mechanisms such as Head-of-Line (HoL) take into account that some services may tolerate longer delays than others (e.g., data versus voice) and deal with the order with which cells are transmitted. Time priorities can be implicitly represented by using combinations of virtual path and channel identifiers (VPI/VCI).

Space prioritiy mechanisms control the allocation of buffer space to arriving cells at an input or output port queue of an ATM switch. Implicitly, they control traffic congestion by providing several grades of service through the selective discarding of low priority cells. This type of priority congestion control mechanism exploits the fact that certain cells generated by traffic sources are less important than others and may, therefore, be discarded without significantly affecting the QoS constraints. Space priority mechanisms aim to decrease the cell loss probability and delays for high priority cells in comparison with low priority cells. One of the main mechanisms for space priorities is the partial buffer sharing (PBS) scheme.

PBS works by setting a sequence of buffer capacity thresholds K_i, $i = 1, 2, \ldots, C$, corresponding to C priority classes (indexed from 1 to C in decreasing order of priority) of a single queue with overall finite capacity K_1. Highest priority cells of class $i = 1$can join the queue simply if there is space. However, lower priority cells of class i, $i = 2, \ldots, C$, can join the queue only if the total number of cells in the queue is less than the threshold value K_j. Once the number of cells waiting for service reaches K_i, then all lower priority cells of class j, $j = i + 1, \ldots, C$, will be lost on arrival but higher priority cells of class i, $i = 1, \ldots, j - 1$, will continue to join the queue until it reaches threshold value, K_i, $i = 1, \ldots, j - 1$ (c.f., for $R = 2$ classes, see Fig. 17.8). Once a cell of lower class is being transmitted, it cannot be lost. Different cell loss and QoS requirements under various load conditions can be met by adjusting the threshold value.

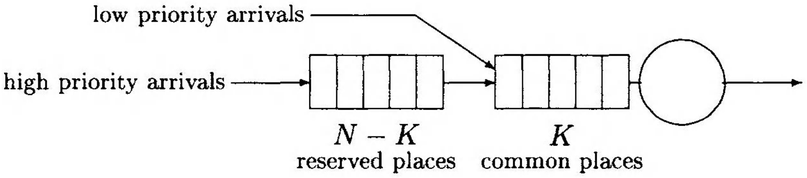


Figure 17.8 The partial buffer sharing space priority mechanism

In this section, an extended ME formalism is applied to characterise at equilibrium closed-form expressions for the state probabilities of a deterministic single server queueing model of a multi-buffered ATM switch with sGGeo-type arrivals and priority congestion control mechanism described by C ($C > 1$) priority classes under Head-of-Line (HoL) service discipline, AF and/or DF buffer management simultaneity policies and PBS scheme. This queueing model, in conjunction with two moment flow approximation formulae per class, plays the role of a cost-effective building block towards a queue-by-queue decomposition algorithm of a corresponding open queueing network model (QNM) of multi-buffered ATM switches under RS-RD blocking.

5.1 A QUEUEING MODEL FOR MULTIBUFFERED SWITCHES WITH PRIORITY CONGESTION CONTROL

Consider a deterministic single server queue at equilibrium with C ($C \geq 2$) HoL priority classes, external sGGeo-type arrivals with correlation parameters $\{\beta_{a0ki}, a_{0ki}, \beta_{b0ki}, b_{0ki}, k = 1, \ldots, M, i = 1, \ldots, C\}$, AF and/or DF buffer management policies and PBS scheme. Matching a GGeo distribution to sGGeo traffic process on counts, the model is denoted by ΣGGeo/D/1/K$_1$,...,K$_C$ such that ΣGGeo stands for a multiple of C class arrival streams, the total buffer capacity is K_1 and the PBS scheme is specified by the sequence of thresholds $\{K_1, \ldots, K_C : K_i < K_j, \forall i < j\}$. Cells are transmitted by a single deterministic server with mean rate $\mu_i = 1, i = 1, 2, \ldots, C$. Moreover, let λ_i be the mean arrival rate and C_{ai}^2 be the SCV of the GGeo-type interarrival time distribution per class i, $i = 1, 2, \ldots, C$.

Let at any given time the state of the system be described by a vector $\mathbf{S} \equiv (n_1, n_2, \ldots, n_C, \omega)$, where n_i, $i = 1, \ldots, C$, is the number of class i cells in the queue (waiting for or receiving service) and ω is the variable indicating the class of the current cell in service (n.b., for an idle queue $\omega = 0$). Let $\mathbf{Q}$ be the set of all feasible states $\{\mathbf{S}\}$ and $p(\mathbf{S})$ be at any given time the equilibrium probability that the ΣGGeo/D/1/K$_1, \ldots,$ K$_C$ priority queue is in state $\mathbf{S}$ and π_i be the cell loss or blocking probability that an arriving cell of class i, $i = 1, 2, \ldots,$ C, will find the queue occupied up to at least its corresponding threshold K_i, $i = 1, 2, \ldots, C$.

The form of the state probability distribution, $p(\mathbf{S}), \mathbf{S} \in \mathbf{Q}$, can be characterised by maximising the entropy functional $H(p) = -\sum_{\mathbf{S}} p(\mathbf{S}) \log p(\mathbf{S})$, subject to normalisation and marginal constraints of server utilisation, U_i $(0 < U_i < 1)$, busy state probability per class, θ_i $(0 < \theta_i < 1)$, MQL, L_i $(U_i \leq L_i < K_i)$ and full buffer state probability ϕ_i $(0 < \phi_i < 1)$ per class i, $i = 1, 2, \ldots, C$, satisfying the flow balance equation, namely

$$\lambda_i(1 - \pi_i) = \mu_i U_i, \quad i = 1, 2, \ldots, C \tag{17.48}$$

By employing Lagrange's method of undetermined multipliers, the ME solution is expressed by [17]

$$p(\mathbf{S}) = \frac{1}{Z} \prod_{i=1}^{C} g_i^{s_i(\mathbf{S})} \xi_i^{h_i(\mathbf{S})} x_i^{n_i(\mathbf{S})} y_i^{f_i(\mathbf{S})}, \ \forall \mathbf{S} \in \mathbf{Q}, \tag{17.49}$$

where Z is the normalising constant and $\{g_i, \xi_i, x_i, y_i, i = 1, 2, \ldots, C\}$ are the Lagrangian coefficients corresponding to constraints $\{U_i, \theta_i, L_i, \phi_i, i = 1, 2, \ldots, C\}$, respectively. Furthermore, aggregating (17.49) over all feasible states $\mathbf{S} \in \mathbf{Q}$, the joint ME queue length distribution $p(\mathbf{n})$ is given by:

$$p(\mathbf{n}) = \frac{1}{Z} \prod_{i=1}^{C} x_i^{n_i(\mathbf{S})} \xi_i^{h_i(\mathbf{S})} \sum_{s=1 \wedge n_s > 0}^{C} g_s y_s^{f_s(\mathbf{n})}, \ 0 \leq n_i \leq K_i, \ \& \sum_{i=1}^{C} n_i \leq K_1 \tag{17.50}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_C)$ and $p(\mathbf{0}) = 1/Z$. Note that the generic forms of ME solutions (17.49) and (17.50) are universal and are applicable to both GGeo-type and GE-type queueing systems under a PBS scheme.

By making asymptotic connection as $K_1 \to \infty$, the Lagrangian coefficients $\{g_i, \xi_i, x_i, i = 1, \ldots, C\}$ can be approximately determined in terms of input parameters. For example, for either a ΣGGeo/D/1/K$_1$,... ,K$_C$ or ΣGE/D/1/K$_1$,... ,K$_C$ queueing model, the Lagrangian coefficients $\{g_i, \xi_i, x_i, i = 1, \ldots, C\}$ are given by [18]

$$\begin{aligned} x_i &= \frac{L_{i,\infty} - \theta_{i,\infty}}{L_{i,\infty}}, \; g_i = \frac{\lambda_i}{(1-\lambda)} \frac{\lambda - \theta_{i,\infty}}{\theta_{i,\infty} - \lambda_i} \prod_{i \neq l = 1}^{C} \frac{\lambda - \theta_{l,\infty}}{\lambda - \lambda_l}, \\ \xi_i &= \frac{(\theta_{i,\infty} - \lambda_i)}{(\lambda - \theta_{i,\infty})} \frac{(1 - x_i)}{x_i}, \end{aligned} \qquad (17.51)$$

where $\lambda = \sum_{i=1}^{C} \lambda_i$ and, for $\lambda < 1$, constraints $\{L_{i,\infty}, \theta_{i,\infty}\}$,i = 1,2,..., C can be interpretted as the exact MQL and busy state probability per class i of the corresponding infinite capacity GGeo/D/1/HoL and GE/D/ 1/HoL queues, as appropriate, at equilibrium.

By applying the generating function approach (c.f., [19]), recursive expressions for [U_i, Z, L_i aggregate and marginal state probabilities and blocking (or cell loss) probabilities $\{\pi_i, i = 1, 2, \ldots, C\}$can be obtained. Consequently, Lagrangian coefficients, $\{y_i, i = 1, 2, \ldots, C\}$, can be determined recursively by making use of flow balance condition (17.48) and cell loss probabilities, $\{\pi_i, i = 1, 2, \ldots, C\}$.

5.2 NETWORKS OF MULTIBUFFERED SWITCHES WITH PRIORITY CONGESTION CONTROL

Consider at equilibrium an arbitrary open QNM with M a mutiple of R single server queueing stations, C distinct classes of cells, sGGeo external interarrival times, HoL scheduling discipline, AF and/or DF buffer management simultaneity policies, PBS scheme with thresholds $\{K_{ki}, i = 1, \ldots, C, \; k = 1, \ldots, M : K_{ki} < K_{kj}, \; \forall k_i < k_j\}$ and RS-RD blocking mechanism. This model may be used to represent a network of [M/R] multi-buffered ATM switches with space/time priorities and output port queueing.

Let p_{kimj} be the transition probability (first order Markov chain) that a class i cell transmitted from station k attempts to join station m as class j, p_{ki0} be the transition probability that a cell of class i leaves the network upon finishing transmission at station k. By applying the GGeo-type approximation of an external sGGeo-type arrival proces with correlation parameters $\{\beta_{a\,0ki}, a_{0ki}, \beta_{b\,0ki}, b_{0ki}, \; k = 1, 2, \ldots, M, i = 1, \ldots, C\}$, let λ_{0ki} be the mean arrival rate and $C^2_{a\,0ki}$ be the SCV of the external GGeo-type interarrival process of cells at station k. Let at any given time

n_{ik} be the number of cells of class *i* at queue *k*, $\mathbf{n}_k = (n_{k1}, n_{k2}, \ldots, n_{kR})$ be the state of queue *k*, and $\mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_M)$ be the state of the entire network.

The form of the ME solution $p(\mathbf{n})$, subject to normalisation and marginal constraints $\{U_{ki}, \theta_{ki}, L_{ki}, \phi_{ki}, k = 1, \ldots, M, i = 1, \ldots, C\}$ (c.f., Section 5.1) , can be clearly established in terms of the product-form approximation

$$p(\mathbf{n}) = \prod_{k=1}^{M} p_k(\mathbf{n}_k), \tag{17.52}$$

where $p(\mathbf{n}_k)$ is the marginal ME solution of queue *k*, given by (c.f., (17.49)

$$p(\mathbf{n}_k) = \frac{1}{Z_k} \prod_{i=1}^{C} x_i^{n_i(\mathbf{n}_k)} \xi_i^{h_i(\mathbf{n}_k)} \sum_{s=1 \wedge n_s > 0}^{C} g_s y_s^{f_s(\mathbf{n}_k)}, \tag{17.53}$$

where $n_i(\mathbf{n})$, $h_i(\mathbf{n})$ and $f_i(\mathbf{n})$ are suitable auxiliary functions.

The ME solution (17.52) implies a queue-by-queue decomposition algorithm for the approximate analysis of arbitrary open QNMs with single server queueing stations, $C(C > 1)$ HoL priority classes, AF and/or DF buffer management policies, PBS scheme and RS-RD blocking. A sketch of the algorithm is described below. It is an extension of the ME algorithm of Section 4.2 to the case of arbitrary open FCFS QNMs with multiple priority classes and RS-RD blocking mechanism. The algorithm incorporates the matching of GGeo to external sGGeo traffic on counts and assumes that the arrival process per class at each queue conforms with a GGeo distribution. Furthermore, the algorithm describes the computational process of solving iteratively the non-linear equations for blocking probabilities $\{\pi_{0ki}, \pi_{kimj}, k, m = 1, \ldots, M, i, j = 1, \ldots, C\}$ under the generic GGeo-type flow formulae for the first two moments of merging, spliting and departing streams as applied to each class of cells i, $i = 1, 2, \ldots, C$ (c.f., Section 3.3). Note that π_{0ki} is the blocking probability that an external arrival of class *i* is blocked by station *k* and π_{kimj} is the blocking probability that a cell of class *i* following its transmission from station *k* will be blocked by station *m*, as class *j* (i.e., class switching). In particular, the probabilities $\{\pi_{0ki}, \pi_{kimj}\}$ generally depend on the effective cell flow balance equations for $\left\{\hat{\lambda}_{0ki}, \hat{\lambda}_{ki}\right\}$, effective transition probabilities $\{\hat{p}_{0ki}, \hat{p}_{kimj}\}$, effective interarrival time SCV $\hat{C}^2_{a\,ki}$, effective transmission time parameters $\left\{\hat{\mu}_{ki}, \hat{C}^2_{s\,ki}\right\}$ and overall interarrival time parameters $\{\lambda_{ki}, C^2_{a\,ki}\}$

Begin

Input Data

- $M, C,$
- $\forall\, k, m,\ k = 1, 2, \ldots, M, m = 0, 1, \ldots, M\ \&\ \forall\, i, j = 1, 2, \ldots, C,$

  $\left\{K_k, \lambda_{0ki}, C^2_{a\,0ki}, \mu_{ki} = 1, p_{ki0}, p_{kimj}\right\}$, $\{\beta_{a\,0ki},\ a_{0ki}, \beta_{b\,0ki}, b_{0ki}\}$

Step 1 Matching GGeo to external sGGeo traffic on counts;

Step 2 Initialize π_{0ki}, & $\pi_{kimj} \leftarrow$ any value in (0,1), $\forall\, k, m = 1, 2, \ldots, M$ and $\forall\, i, j = 1, 2, \ldots, C,$

Step 3 Solve the system of non-linear equations $\{\pi_{0ki},\ \pi_{kimj},\ \forall\, i, j, k, m\}$ under AF and/or DF buffer management simultaneity policies, as appropriate;

Step 3.1 Calculate effective flow transition probabilities $\{\hat{p}_{kimj}\}$:

$$\hat{p}_{kimj} = p_{kimj}(1 - \pi_{kimj})/(1 - \pi_{cki})$$
$$\hat{p}_{ki0} = p_{ki0}/(1 - \pi_{cki})$$
$$\pi_{cki} = \sum_{k \neq m=1}^{M} \sum_{j=1}^{C} p_{kimj}\pi_{kimj};$$

where π_{cki} be the blocking probability that, following transmission, a cell of class *i* will be blocked by a downstream station.

Step 3.2 Calculate effective cell flow balance equations:

$$\hat{\lambda}_{0ki} = \lambda_{0ki}(1 - \pi_{0ki}),$$
$$\hat{\lambda}_{ki} = \hat{\lambda}_{0ki} + \sum_{m=1}^{M} \sum_{j=1}^{C} \hat{p}_{mjki}\hat{\lambda}_{mj};$$

Step 3.3 Calculate the effective transmission time parameters, $\left\{\hat{\mu}_{ki}, \hat{C}^2_{s\,ki}\right\}$

$$\hat{\mu}_{ki} = 1 - \pi_{cki};$$
$$\hat{C}^2_{s\,ki} = \pi_{cki};$$

(n.b., a GGeo or GE distribution with parameters $\hat{\mu}_{ki}$ and $\hat{C}^2_{s\,ki}$ can be used, as appropriate, to model the effective transmission time).

Step 3.4 Calculate overall interarrival parameters, $\left\{\lambda_{ki}, C^2_{a\,ki}\right\}$

$$\lambda_{ki} = \hat{\lambda}_{ki}/(1 - \pi_{ki});$$
$$C^2_{a\,ki} = \frac{\hat{C}^2_{a\,ki} - \pi_{ki}}{1 - \pi_{ki}};$$

$$\pi_{ki} = \frac{\lambda_{0ki}\pi_{0ki} + \sum_{m=1}^{M}\sum_{j=1}^{C}\left(\hat{p}_{mjki}\hat{\lambda}_{mj}\pi_{mjki}/(1-\pi_{mjki})\right)}{\lambda_{0ki} + \sum_{m=1}^{M}\sum_{j=1}^{C}\left(\hat{p}_{mjki}\hat{\lambda}_{mj}/(1-\pi_{mjki})\right)};$$

(n.b., π_{ki} is the blocking probability that a cell of class i, having just completed transmission via station m, $m \neq k$, is blocked by station k).

Step 3.5 Obtain new values for $\{\pi_{0ki}, \pi_{mjki}\}$, by applying Newton Raphson method,

Step 4 For all $k = 1, 2, \ldots, M$, $i = 1, 2, \ldots, C$

Step4.1 Calculate interdeparture time parameters $\{\lambda_{d\,ki}, C^2_{d\,ki}, k = 1, 2, \ldots, M, i = 1, \ldots, C\}$ (c.f., (17.34)-(17.35));

Step 4.2 Calculate the splitting of the interdeparture time parameters $\{p_{kimj}\lambda_{d\,ki}, C^2_{d\,km}, k = 1, 2, \ldots, M, m = 0, 1, \ldots, M\}$ (c.f., (17.49)-(17.37));

Step 4.3 Calculate new value for overall interaarival parameters $\{\lambda_{ki}, C^2_{a\,ki}, k = 1, 2, \ldots, M, i = 1, 2, \ldots, C\}$ (c.f., (17.32)-(17.33));

Step 4.4 Return to Step 3 until convergence of $C^2_{a\,ki}$;

Step 5 Obtain performance metrics of interest by solving each queueing station of the network as a stable ΣGGeo/GGeo/1/K$_1$,... ,K$_C$ priority queue with overall interarrival time parameters $(\lambda_{ki}, C^2_{a\,ki})$ and effective transmission parameters $(\hat{\mu}_{ki}, \hat{C}^2_{s\,ki})$, $k = 1, 2, \ldots, M$, $i = 1, 2, \ldots, C$;

End

The main computational cost of the proposed algorithm is of O{kR^{2}M^2}, where k is the number of iterations in step 3 via a quasi-Newton numerical method. The basic structure of the algorithm is also applicable to multiple class GE-type networks under PBS scheme and a related case study, involving the stable ΣGE/D/1/K$_1$, ... ,K$_C$ queue as a building block, can be seen in [17]. Note that numerical instabilities were never observed during extensive experimentations under any feasible set of initial values.

6. OTHER ATM SWITCH ARCHITECTURES

This section gives brief accounts of the ME applications into the performance modelling and analysis of shared buffer, space division Banyan Multistage Interconnection Networks (MINs) and shared medium ATM switch architectures with bursty and/or SRD traffic. Note that all corresponding ME algorithms can be readily extended in order to incorporate a priority congestion control mechanism (c.f., Section 4.1 and 4.2).

6.1 SHARED BUFFER ATM SWITCH ARCHITECTURES

Shared buffer ATM switch architectures incorporate a single memory of fixed size which is shared by all output ports. An incoming cell is stored in a shared buffer of finite capacity while its address is kept in the address buffer. Cells destined for the same output port can be linked by an address chain pointer or their addresses can be stored into a FCFS buffer which relates to a particular output port. A cell will be lost if on arrival it finds either the shared buffer or the address buffer full.

Consider a queueing model of a shared buffer switch with a multiple of R sGGeo type external arrivals and output port queueing depicted in Fig. 17.9. Following a GGeo matching to sGGeo on counts (c.f., Section 3.2), the model is denoted by$S_{R\times R}$(ΣGGeo/D/1)/K, where ΣGGeo stands for a multiple of R GGeo-type arrival processes at each output port, D indicates deterministic transmission times, K is the size of total shared buffer and R is the number of parallel single server queues. In the case of GE-type interarrival times, the model is denoted by $S_{R\times R}(\Sigma GE/D/1)/K$. Each server represents an output port and each queue corresponds to the address queue for the output port. There are Rx R bursty and heterogeneous interarrival streams of cells. Each traffic stream has a mean overall arrival rate, λ_{ji} , of cells and an overall interarrival time SCV, C^2_{aji},for stream (j,i), $i,j = 1, 2,.., R$. Each queue i has a cell transmission rate $\mu_i = 1$ cell per slot at port i, $i = 1,2,...,R$. A cell is lost if it arrives at a time when there is a total of K cells in the R queues.

Moreover, let the state of the system at any given time be represented by a vector $\mathbf{n} = (n_1, n_2, \ldots, \quad n_R)$, where n_i is the number of cells in queue i, $i = 1,2, ..., R$, S(K,R) be the set of all feasible states and $p(\mathbf{n}), \mathbf{n} \in$ S(K,R) be the joint state probability distribution. The generic form of the ME solution of an $S_{R\times R}(GGeo/D/1)/K$ queueing system, subject to normalisation and the constraints of server utilisation, U_i $(0 < U_i < 1)$, MQL, L_i $(U_i \leq L_i < K)$ and conditional aggregate full buffer probability, ϕ_i $(0 < \phi_i < 1)$, subject to $n_i > 0$, $i = 1, 2, \ldots, R$, is given by [20,21]

$$p(\mathbf{n}) = \frac{1}{Z}\prod_{i=1}^{R} g_i^{s_i(\mathbf{n})} x_i^{n_i} y_i^{f_i(\mathbf{n})}, \forall \mathbf{n} \in S(K,R), \qquad (17.54)$$

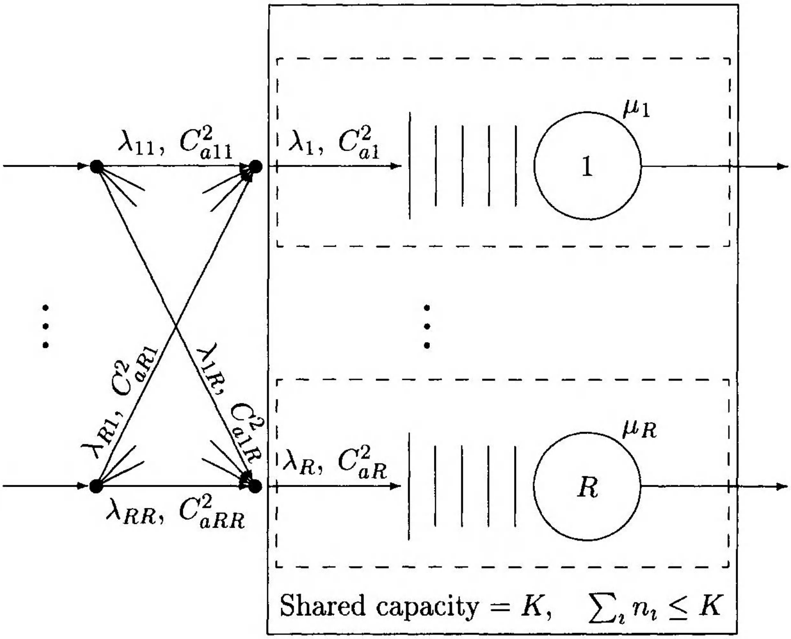

λ_{11}, C^2_{a11}
λ_1, C^2_{a1}
μ_1
1
λ_{R1}, C^2_{aR1}
λ_{1R}, C^2_{a1R}
λ_R, C^2_{aR}
μ_R
R
λ_{RR}, C^2_{aRR}
Shared capacity $= K, \quad \sum_\imath n_\imath \leq K$

Fıgure 17 9 The $S_{R\times R}(\Sigma GGeo/D/1)/K$ shared buffer queueıng model

where Z is the normalising constant, and $s_i(\mathbf{n})$ and $f_i(\mathbf{n})$ are auxiliary functions and $\{g_\imath, x_\imath, y_\imath, \imath = 1, 2, \ldots, R\}$ are the Largrangian coefficients corresponding to the constraints $\{U_\imath, L_\imath, \phi_\imath, \imath = 1, 2, \ldots, R\}$, respectively. Lagrangian coefficients $g_i, x_i, y_i, i = 1, 2, \ldots, R$ are assumed to be invariant to K and can be approximated by

$$g_\imath = \frac{\lambda_\imath(1 - x_\imath)}{x_\imath(1 - \lambda_\imath)}, \; x_\imath = \frac{L_\imath - \lambda_\imath}{L_\imath}, \; \imath = 1, 2, \ldots, R, \tag{17.55}$$

wherefor $\lambda_\imath (= \sum_{j=1}^{R} \lambda_{j\imath}) < 1, \imath = 1, 2, \ldots, R$ and L_i can be interpretted as the asymptotic MQL of the corresponding infinite capacity queue at equilibrium.

Moreover, Lagrangian coefficients $\{y_i, i = 1, 2, \ldots, R\}$ can be computed by

(i) applying the generating function approach to derive recursive expressions for Z, $\pi_{j\imath}$, and U_i, for $i,j = 1, 2, \ldots, R$, of an $S_{R\times R}(\Sigma GGeo/D/1)/K$ [21] (or $S_{R\times R}(\Sigma GE/D/1)/K$ [20]) shared buffer queue, where $\pi_{j\imath}$ is the cell loss (or blocking) probability at the output port queue i, and

(ii) using the Newton-Raphson algorithm to solve numerically the resultant non-linear simultaneous equations of the flow balance conditions

$$\sum_{j=1}^{R} \lambda_{ji}(1 - \pi_{ji}) = U_i\mu_i,\ i = 1, 2, \ldots, R \tag{17.56}$$

Note that because of the recursive nature of the z-transforms which are used in the computational implementation of the ME solution, the $\mathrm{S}_{R\times R}(\Sigma\mathrm{GGeo/D/1})/\mathrm{K}$ queueing model can be used as an effective building block in the analysis of networks of shared buffer switches such as the switch architecture Prelude proposed by CNET[22] with loss and also large Banyan MINs with RS-RD blocking.

The Prelude architecture is displayed in Fig. 17.10. Routing of cells through the network can be based upon the notion of the virtual circuit (VC). A VC has a fixed path through the network. All cells that belong to a particular VC flow along its path.

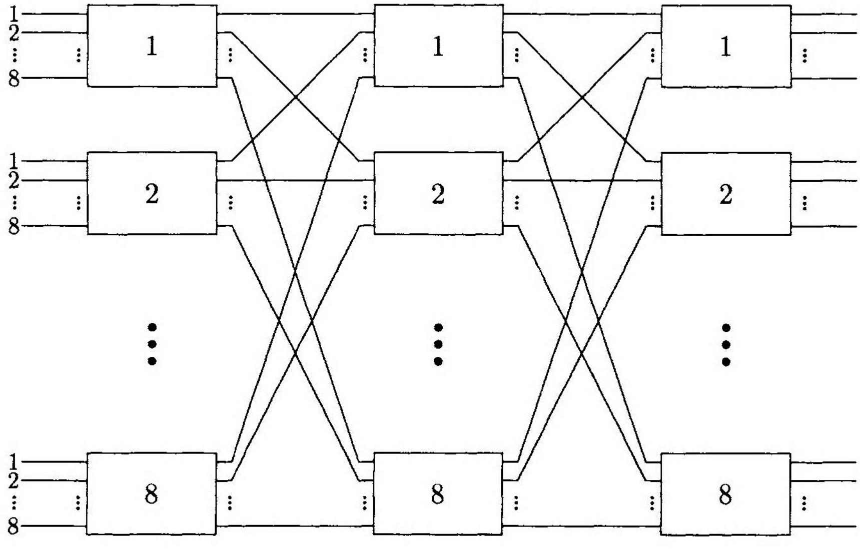


Figure 17.10 Prelude architecture: A network configuration of 8x8 shared buffer switches with loss

The first two moments of the external flow of each VC as it arrives at the network are known. Internal flows of cells belonging to VCs must be converted to flows through each switch/port and from one switch/port to another. Due to finite buffer sizes, cell loss will occur at switches and thus within a VC the flow of cells will reduce at each link comprising its path. Because cell flows are attenuated, it is not possible to calculate the flows *a priori.* However an iterative ME decomposition algorithm for arbitrary open QNMs into individual shared buffer switches can be developed based on

(i) The ME solution of the $S_{R\times R}(\sum GGeo/D/1)/K$ (or $S_{R\times R}(\sum GE/D/1)/K$) shared buffer queueing model,

(ii) The GGeo-type (or GE-type) flow formulae (c.f., Section 3.3) for calculating the overall mean and SCV of the interarrival and interdeparture times at each output port, and

(iii) The mean rate of VCs on each link of their paths.

The main computational cost of this algorithm is the calculation of cell loss probabilities at the output ports of the shared buffer switch, which must be obtained at each iteration. However, these computations can be performed in a few minutes on a SUN workstation.

The utility of the ME algorithm for GGeo-type networks of shared buffer switches can be seen in [23]. This algorithm captures the log-linear relationship between very small cell loss probabilities of a hot-spot output port and the optimum (minimum) buffer capacity, K. Moreover, the ME algorithm has been used to carry out cross performance comparisons involving typical multi-buffered and shared buffer switch architectures with the same targetted cell loss probabilities, input data (other than buffer capacity) and an increasing number of ports. It has been verified that, although these architectures have equivalent mean time delays, the percentage increase in buffer size requirements for the multi-buffered switch is substantially higher than those of the shared buffer switch (c.f., [23]).

6.2 BUFFERED BANYAN MIN ARCHITECTURES

Space division switches are primarily based on N×N MINs, where N is the total number of external input (or output) ports. MINs are composed of smaller switching elements represented by shared-buffer crossbars. Main features of a MIN include non-centralised switching control and multiple concurrent paths in tandem from input ports to output ports. Notably, the flow of cells through one switching element may be momentarily blocked (halted) if the downstream switching element has reached its buffer capacity. The ATM switch consists of L levels and M stages and employs, as basic building blocks, R-input and R-output shared buffer switching elements represented by shared buffer $S_{R\times R}(\Sigma GGeo/D/1)/K_{\ell m}$ (or $S_{RxR}(\ \Sigma GE/D/1)/\ K_{\ell m})$, $l =$ $0,1,\ldots,L-1$, $m = 0,1,\ldots,M-1$, queueing models (c.f., Fig. 17.9). The input/output ports of the MIN form an array of 'pins'. Each output pin is linked to a single down stream input pin at the next stage. These connections form the topology of the network and are represented in the forwards (FTM) and backwards (BTM) MxN topology matrices A typical finite buffered ATM switch with an 8 × 8 Banyan MIN based architecture is depicted in Fig. 17.11.

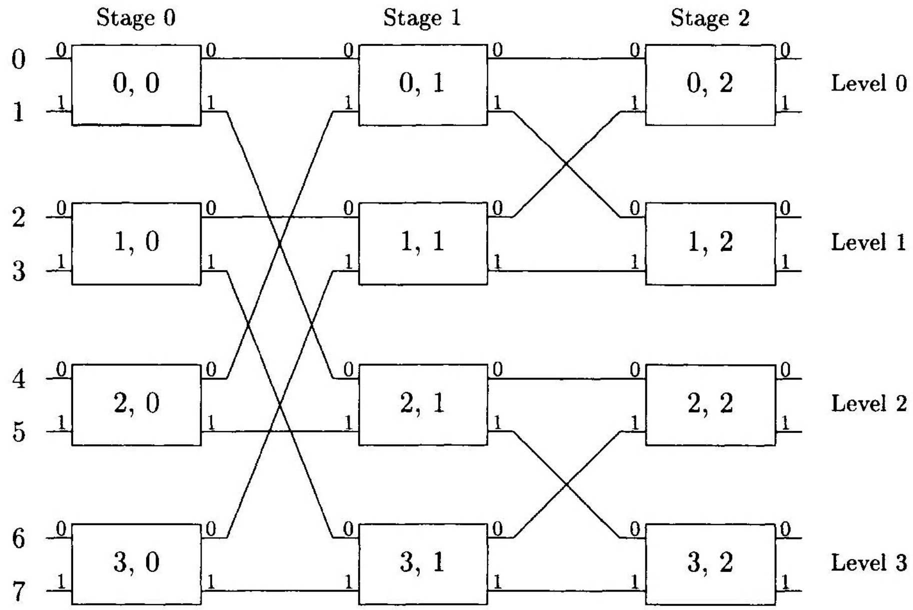


Figure 17.11 An 8×8 configuration of a regular Banyan Network

The flow to external input pin k can be parameterised by the overall mean arrival rate, λ_k, and the SCV of interarrival times, $C^2_{a\,k}$. Incoming cells traverse the network according to both topology matricies and $\{r_{ks}\}_{N\times N}$, the routing probability matrix, where is the probability that a cell originating at external input pin k has external output pin s as its destination. A cell is lost if on arrival at an external switching element $(i,0)$, $i = 0,1,2,3$, finds a full buffer. However, every cell that enters the MIN is guaranteed delivery to its destination. This constraint, along with the finite buffers of internal switching elements, implies that the cell will be blocked and thus, the MIN operates internally a blocking mechanism.

Entropy maximisation implies a decomposition of the Banyan MIN into individual shared buffer switching elements with modified arrival and service parameters reflecting the effective and overall flows through the switching elements. An iterative decomposition algorithm for arbitrary Banyan MINs with blocking can be determined (c.f., [24]) based on

(i) the ME solution of the $S_{R\times R}(\Sigma G\,Geo/GGeo/1)/K$ (or $S_{RxR}(\Sigma G\ E/D/1)/K$) shared buffer queues,

(ii) GE type flow formulae for the first two moments for the interarrival and interdeparture times at each output pin.

The main computational cost of the ME algorithm at every iteration is the calculation of blocking probabilities at the output pins of each switching element. These computations can be performed even for very large networks in a few minutes on a SUN workstation.

The ME algorithm for GE-type Banyan networks has been validated against simulation and moreover, has been utilised as a cost-effective tool, for investigating the trade-off between cell loss probabilities (or, equivalently, throughput) against end-to-end delay under different buffer capacity assignment policies across typical Banyan MINs (c.f., [24]).

6.3 SHARED MEDIUM ATM SWITCH ARCHITECTURES

Consider a shared medium ATM switch architecture represented by a polling model which consists of N input ports and R output ports depicted in Fig. 17.12.

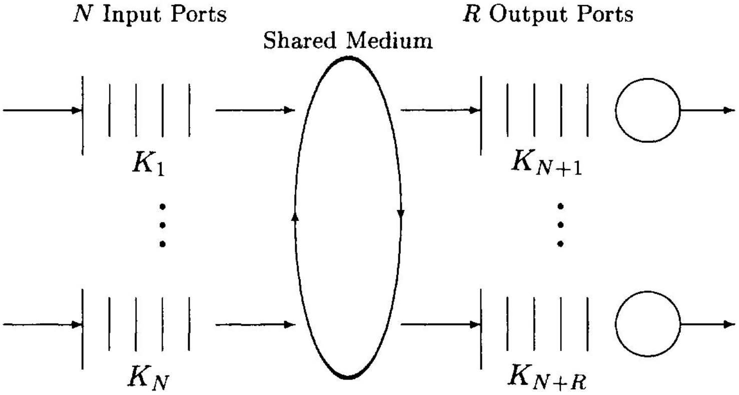


Figure 17.12 Queueing model of a shared medium switch

Each input port has finite buffer capacity and receives bursty traffic modelled by a GGeo (or GE) distribution. A shared medium such as a high speed bus plays the role of a server that forwards external input traffic towards the output ports in a cyclic fashion. Each output port has finite capacity and provides deterministic (D) type of service. An analytic ME algorithm can be developed [25], based on the concept of system decomposition whereby the polling system is partitioned into individual censored input and output port queues with or without server vacations, respectively and modified transmission times (see Fig. 17.13).

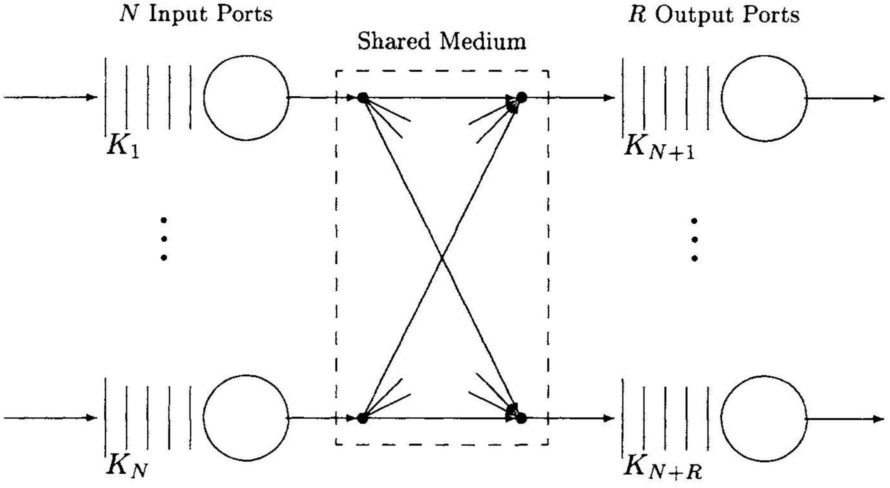


Figure 17.13 Decomposed queueing model of a shared medium switch

Subsequently, the shared medium with its associated N input link queues can be analysed as a separate system utilising a relationship between polling time and server vacation. Furthermore the output ports can be analysed as individual finite capacity queues. Finally, the analytical results from the different subsystems are combined together via an iterative process, based on the ME algorithm for arbitrary open queueing networks with RS-RD blocking at input port queues and appropriate two moment flow approximation formulae.

Analytic details and numerical results illustrating the credibility of the ME approximations against simulation for a GE-type queueing model of a shared medium switch can be seen in [25].

7. CONCLUSIONS

An exposition of an information theoretic methodology for the approximate analysis of complex QNMs, as applied to the performance evaluation and priority congestion control of ATM networks consisting of multi-buffered, shared buffer, space division Banyan MINs and shared medium, is carried out. The methodology leads to analytic solutions and cost-effective algorithms and it is based on the principle of ME, queueing theoretic concepts and the sGGeo batch renewal traffic process. Central to the tractability of the analysis, with a tolerable accuracy, is the matching of an ordinary GGeo distribution in a discrete-time domain to an sGGeo process, based on the first two moments of counts. Alternatively, the role of the corresponding GE distribution within a mixed-time domain is exposed. Numerical case studies utilising GGeo-type and/or GE-type queue-by-queue ME decomposition algorithms are referred, as appropriate, throughout the paper to earlier works.

The ME methodology provides telecommunication engineers with relatively simple but efficient means of accounting for the effects of external traffic with varying degrees of burstiness and SRD upon performance metrics at the edges and interior of ATM switch architectures and networks. Further research studies, based on the ME methodology, can examine the feasibility of analysing queueing performance when other types of non-trivial correlated arrival processes are replaced by simpler traffic models such as those based on the matching of their first two moments of counts as well as on the notion of equivalent QLDs under a common queueing reference system. Work of this kind is the subject of current studies (e.g., [26]).

References

[1] Mitra, D., "Stochastic Fluid Models", *Performance '87,* Brussels, 1987

[2] Tobagi, F., "Fast Packet Switch Architectures for Broadband Integrated Services Digital Networks", *Proc. of IEEE,* Vol. 78, No. 1, Jan. 1990

[3] Jaynes, E.T., "Information Theory and Statistical Mechanics", II *Phys. Rev.* 108, pp. 171-190, 1957

[4] Kouvatsos, D.D. and Fretwell, R.J., "Batch Renewal Process: Exact Model of Traffic Correlation", *High Speed Networking for Multimedia Application*, W. Effelsbery (ed.), Kluwer Academic Press, pp. 285-304, 1996

[5] Kouvatsos, D.D., "Entropy Maximisation and Queueing Network Models", *Annals of Operation Research*, Vol. 48, pp. 63-126, 1994

[6] Kouvatsos, D.D. and Tabel-Aouel, N.M., "GGeo-Type Approximations for General Discrete-Time Queueing Systems", Modelling and Performance Evaluation of ATM Technology, *IFIP Publication*, Perros, H. Pujolle, G. and Takahaghi, Y. (eds.), North-Holland, pp. 469-483, 1993

[7] Kouvatsos, D.D., Tabel-Aouel, N.M., and Denazis, S.G., "Approximate Analysis of Discrete-Time Networks with or without Blocking", *High Speed Networks and their Performance (C-21)*, Perros, H.G., and Viniotis, Y. (eds.), North-Holland, pp. 399-424, 1994

[8] Shore, J.E. and Johnson, R.W., "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy", *IEEE Trans. Inf. Theory* IT-26, pp. 26-27, 1980

[9] Kouvatsos, D.D. and Fretwell, R.J., "Discrete Time Batch Renewal Processes with Application to ATM Switch Performance", *Proc. 10th. UK Computer and Telecomms. Performance Eng. Workshop*, Jane Hillston *et al.* (eds.), Edinburgh University Press, pp. 187–192, Sept. 1994

[10] Kouvatsos, D.D. and Fretwell, R.J., "Closed Form Performance Distributions of a Discrete Time $GI^G/D/1/N$ Queue with Correlated Traffic", *Enabling High Speed Networks*, Fdida, S. and Onvural, R.O., (eds.), IFIP publication, Chapman and Hall, pp. 141–163, October, 1995

[11] Fretwell, R.J. and Kouvatsos, D.D., "Correlated Traffic Modelling: Batch Renewal and Markov Modulated Processes", *Performance Modelling and Evaluation of ATM Networks*, Volume 3, Kouvatsos, D.D. (ed.), IFIP publication, Chapman and Hall, pp. 20–43, Sept. 1997

[12] Laevens, K., "The Output Process of a Discrete-time GIG/D/1 Queue", *Proc. 6th. IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, Kouvatsos, D.D. (ed.), pp. 20/1-20/10, July 1998

[13] Molnár, S. and Miklós, G., "On Burst and Correlation Structure of Teletraffic Models", *Proc. 5th. IFIP Workshop on Performance Modelling and Evaluation ofATM Networks*, Kouvatsos, D.D. (ed.), pp. 22/1-22/10, July 1997

[14] Skliros, A., "Optimising Call Admission Control Schemes for NCX-1E6 ATM Multiservice Switches", *ECI Telematics International Ltd*, Private Communication, September 1998

[15] Kouvatsos, D.D. and Awan, I.U., “Arbitrary Discrete-Time Queueing Networks with Correlated Arrivals and Blocking”, *Proc. 6th. IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, UK Performance Eng. Workshop Publishers, Kouvatsos, D.D. (ed.), pp. 109/1-109/8, July 1998

[16] Kouvatsos, D.D., Awan, I.U., Fretwell, R.J., Dimakopoulos, G., “A Cost-Effective Approximation for SRD Traffic in Arbitrary Queueing Networks with Blocking”, *Research Report RS-01-00,* Computing Dept., Bradford University, Jan. 2000

[17] Awan, I.U. and Kouvatsos, D.D., “Approximate Analysis of Arbitrary QNMs with Space and Service Priorities”, *Performance Analysis of ATM Networks*, Kouvatsos, D.D. (ed.), Kluwer Academic Publishers, , pp. 497-521, 1999

[18] Kouvatsos, D.D.and Tabet-Aouel, N., “Product-Form Approximations for an Extended Class of General Closed Queueing Networks”, *Performance ’90*, King, P.J.B., Mitrani, I., and Pooley, R.J., (eds.), pp. 301-315, 1990

[19] Williams, A. C. and Bhandiwad, R. A. “A Generating Function Approach to Queueing Network Analysis of Multiprogrammed Computers”, *Networks* Vol. 6, pp. 1-22, 1976

[20] Kouvatsos, D.D. and Denazis, S.G., “A Universal Building Block for the Approximate Analysis of a Shared Buffer ATM Switch Architecture”, *Annals of OR*, Vol. 44, pp. 241-278, 1994

[21] Kouvatsos, D.D., Tabet-Aouel, N. and Denazis, S.G., “ME-Based Approximations for General Discrete-Time Queueing Models”, *Performance Evaluation*, Special Issue on Discrete-Time Models and Analysis Methods, Vol. 21, pp. 81-109, 1994

[22] Devault, M., Cochennec, J.Y. and Servel, M., “The Prelude ATD Experiment: Assessments and Future Prospects”, *IEEE JSAC*6(9), pp. 1528-1537, Dec. 1988

[23] Kouvatsos, D.D. and Wilkinson, J., “A Product-Form Approximation for Discrete-Time Arbitrary Networks of ATM Switch Architectures”, *Performance Modelling and Evaluation of ATM Networks*, IFIP Publications, Chapman and Hall, London, Vol. 1, pp. 365-383, 1995

[24] Kouvatsos, D.D. and Wilkinson, J., “Performance Analysis of Buffered Banyan ATM Switch Architectures”, *ATM Networks: Performance Modelling and Evaluation*, IFIP Publications, Chapman and Hall, London, Vol. 2, pp. 287-323, 1996

[25] Skianis, C.A. and Kouvatsos, D.D., “Performance Analysis of a Shared Medium ATM Switch Architecture”, *12th UK Computer and Telecommunications Performance Engineering Workshop*, UK Performance Eng. Workshop Publishers, Hillston, J. and Pooley, R. (eds.), The University of Edinburgh, pp. 33-48, Sept. 1996

[26] Fretwell, R.J., Dimakopoulos, G. and Kouvatsos, D.D., “Ignoring Count Correlation in SRD Traffic: sGGeo Process vs Batch Bernoulli Process”, *Proc. of 15th UK Performance Engineering Workshop*, Bradley, J.T.and Davies, N.J. (eds.), UK Performance Eng. Workshop Publishers, pp. 285-294, July 1999

INDEX OF CONTRIBUTORS

KEYWORD INDEX