INTEGRATED CONGESTION MANAGEMENT AT THE USER-NETWORK INTERFACE OF AN ATM/B-ISDN NETWORK

by

OLIVER T.W. YU

B.A.Sc., The University of British Columbia, 1981

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF ELECTRICAL ENGINEERING

We accept this thesis as conforming

to the required standard

The UNIVERSITY OF BRITISH COLUMBIA

September 1991

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of ELECTRICAL ENGINEERING

The University of British Columbia
Vancouver, Canada

Date OCTOBER 8, 1991

Abstract

This thesis presents an integrated congestion management platform of user traffic at the UNI of the ATM-based network considering the presence of signalling traffic. Integrated congestion management dictates that congestion control schemes are applied during the call access phase (call admission control scheme) and the information transfer phase (buffer control scheme) of user traffic source. The congestion control schemes are devised to meet the congestion performance requirements and to optimize the performance if possible.

UNI call admission and buffer controls developed for the conventional packet-switched network are not applicable to the ATM-based network because of the different input traffic characteristics. In most of the past investigations on the performance of conventional packet-switched networks, the individual input traffic is mostly computer-to-computer data; such individual and aggregate traffic are well-known to follow the Poisson process. On the other hand, ATM-based networks allow a variety of input traffic in addition to the Poisson-distributed traffic. In this thesis, individual user traffic process is modelled as a two-state Markov modulated Poisson process; the aggregate user traffic process is modeled as a batch Bernoulli renewal process under short-term condition and as a fluid process under long-term heavy traffic condition.

The signalling traffic at the UNI carries call control messages and network management messages originated from the user nodes. The signalling traffic must be serviced quickly since they directly affect call establishment and network efficiency. Up to now, all related congestion control researches only consider user traffic. Consequently, the primary objective for this thesis is to study the effect of the higher-priority signalling

traffic on the multiplexing of user traffic at the UNI. A novel modeling of user traffic multiplexing through the ATM statistical multiplexer at the UNI is proposed: it is characterized by a queueing model with random service disruptions due to the transport of higher priority signalling traffic.

The congestion performance requirements of the user traffic for the UNI are studied in terms of the stochastic cell loss requirement and the deterministic upper-bound cell delay requirement. However, in order to investigate the stochastic cell loss phenomenon due to buffer overflow, the stochastic queue behaviour must first be examined. Consequently, a novel algorithm to solve the stationary distribution of the queue length process under short-term heavy traffic and finite buffer capacity conditions is presented.

A novel UNI call admission control scheme is proposed, and its objective is to maintain the required network performance assigned to the UNI access-node by exerting call admission control in the call access phase of each user traffic source. It is analyzed using an input-limit static control model employing stochastic ordering between the cell loss ratio random variable and the desired threshold random variable as a criterion to decide if a new call should be admitted. The cell loss ratio random variable has been chosen as the performance objective rather than the long-term-time-averaged cell loss ratio, so as to take into account of the dynamic nature of bursty traffic sources.

A novel UNI intra-node buffer control scheme is proposed, and its objective is to optimize the network performance of the UNI access-node by exerting buffer control in the aggregate information transfer phase of the user traffic sources. It is analyzed by means of a sequential decision process model characterized by a stationary, Markovian and deterministic threshold control policy.

Table of Contents

List of Tables

List of Figures

Acknowledgment

I would like to express my sincere gratitude to my research supervisor, Dr. V.C.M. Leung for his valuable suggestions and critiques. I had to make a lot of adjustments to become a "poor" student again after working comfortably for few years in the industry, and I am grateful to my family for the constant encouragement and support. I would also like to thank Bell-Northern Research Inc. for granting me an educational leave of absence. Part of this research work is supported by the Natural Science and Engineering Research Council of Canada and the Electrical Engineering Department of the University of B.C.

Chapter 1 Introduction

In this introductory chapter, the evolution of N-ISDN into B-ISDN, the circumstances leading to the development of ATM-based network, and the concept of integrated congestion management are reviewed in the first section. The objectives and motivations of this research are explained in the second section. The approach that is employed to achieve the objectives is outlined in the third section. Finally, a roadmap of this thesis is given in the last section.

1.1 Background

The main feature of the original ISDN (Integrated Services Digital Network) [1–3] concept is the support of a wide range of voice and non-voice applications in the same network. The first generation of ISDN concept (1976–1988) is known as the narrowband ISDN (N-ISDN). It supports 64 Kbps basic access B channels, 16 or 64 Kbps signalling access D channels and 0.35–2.0 Mbps primary access H channels. The 64 Kbps basic access rate was chosen because it was the standard rate for switching and transmission of digitized voice in the telephony network,

The second generation of ISDN concept (initiated by CCITT in 1985) is known as the broadband ISDN (B-ISDN) [4–6] because it supports bit rate greater than 2 Mbps. It is generally agreed that 150 Mbps channels (to support high-resolution video) and 600 Mbps channels (to support multiple simultaneous high-resolution video) are required. Currently, the only widely available transmission facility supporting such data rates is the optical fiber.

While OSI (Open System Interconnection) is a reference model for generalized networks, ISDN is a reference model for specialized networks supporting integrated services (circuit-mode, packet-mode and etc.) access. The ISDN reference architecture is illustrated in Fig. 1 and it is composed of the following facilities:

- Backbone Transport Subnet
- Signalling Transport Facility
- User-Network Interface (UNI)

The CCITT recommendation for N-ISDN does not specify the implementation models for the above facilities. Some possible physical architectures of the backbone transport network are outlined as follows:

- A combination of circuit-switched and packet-switched networks.
- A single circuit-switched network supporting integrated transport of circuit-mode and packet-mode traffic.
- A single packet-switched network supporting integrated transport of circuit-mode and packet-mode traffic.
- A single hybrid transfer-mode network [7] supporting integrated transport and integrated switching of circuit-mode and packet-mode traffic.

Several physical architectures of the signalling transport network are possible:

- Inchannel Signalling
    - Distinct signalling transport network does not exist — signalling and user information traffic are transported over the same backbone transport network.

- Common Channel Signalling (CCS)

  a. Distinct signalling transport network exists — signalling traffic is transported over the signalling transport network and user traffic is transported over the backbone transport network.

  b. The architecture of the signalling transport network can be either the same as or different from the backbone transport network.

Previous research in the integrated transport of circuit-mode and packet-mode traffic over the N-ISDN has primarily focused on the strategies of using existing circuit-switched telephony network to support integrated transport. In [8–12], each connection of the circuit-switched network was characterized by a synchronous slot allocation in a Time Division Multiplexing (TDM) scheme in the link layer. In [13], each connection of the circuit-switched network was characterized by a frequency band allocation in a Frequency Division Multiplexing (FDM) scheme in the link layer.

With recent advances in fiber optics technology and the introduction of the B-ISDN concept, research on the reference model for the backbone transport network has begun to gain momentum. The reference model must account for the followings: (1) the requirement of supporting high bandwidth user services that demand small end-to-end delay (e.g. less than 1ms for voice telephony); (2) the requirement of supporting user services of varying data rates (e.g. 64Kbps for voice telephony, 768 Kbps for HiFi sound, 20→45Mbps for compressed normal-definition TV, 45→145Mbps for compressed extended-definition TV, 92→200Mbps for compressed HDTV); (3) the requirement of supporting integrated transport of circuit-mode traffic (e.g. voice and video telephony) and packet-mode traffic (e.g. data messaging).

Requirements (2) and (3) can be capably supported by a packet-switched network because it allows interconnection of devices of differing data rates; and it allows accommodation of bursty traffic and more efficient use of transmission resources.

However requirement (1) is not easily supported by a packet-switched network because the delay can be large and variable. This weakness in delay is due to the packet processing at each node and throughput bottlenecks caused by network congestion.

Overall, there is a general consensus that the B-ISDN backbone transport subnet is better served by a packet-switched network rather than by a circuit-switched network; and that the logical architecture of the packet-switched network should be modified to overcome its weakness. This has led to ongoing research in fast packet switching networks which overcome the delay weakness due to the packet processing at each node by means of the following modifications:

- Simplification of link layer protocol

  The link-by-link or node-to-node error and flow control functions are eliminated.
- Employment of internal virtual circuits

  Routing decision time is reduced once a virtual circuit is set up.
- Employment of hardware switching

  Routing function is implemented in hardware or firmware.

A fast packet switching network protocol known as the ATM (Asynchronous Transfer Mode) protocol is expected to be incorporated into future B-ISDN networks. The ATM protocol is a Transfer Mode Layer protocol in terms of the B-ISDN reference model illustrated in Fig. 2. The B-ISDN reference model incorporates the following layers:

1. Physical Media Dependent Layer

a. OSI equivalency: Physical Layer.

b. Emerging protocol standards: SONET (Synchronous Optical Network)

2. Transfer Mode Layer

a. OSI equivalency: Link Layer, Network Sublayer of SNACP (SubNetwork Access Convergence Protocol).

b. Emerging protocol standards: ATM (Asynchronous Transfer Mode)

3. Adaptation Layer

a. OSI equivalency: Network Sublayer of SNDCP (SubNetwork Dependent Convergence Protocol)

4. Bearer Service Layer

a. OSI equivalency: Network Sublayer of SNICP (SubNetwork Independent Convergence Protocol)

b. Emerging protocol standards:

- CO-CBR (Connection-oriented Constant Bit Rate) circuit-mode.
- CO-VBR (Connection-oriented Variable Bit Rate) packet-mode.
- CL-VBR (Connectionless Variable Bit Rate) packet-mode.

5. Teleservice Layer

a. OSI equivalency: Layers above and including Transport Layer.

b. Protocol standards: voice and video telephony, teletex, videotex, etc.

An ATM-based network is composed of an ATM-based subnet interconnecting ATM-based UNI's. The main features of the ATM protocol are as follows:

- Fixed-sized ATM protocol data units (cells) are employed within the network. Each cell is composed of 48 bytes of data and 5 bytes of header.
- Statistical multiplexing is employed within the network.
- Cell switching with virtual circuits are employed within the subnet.

1.2 Objectives and Motivations

The objectives for this thesis are stated as follows:

- To study the effect of the higher-priority signalling traffic on the multiplexing of user traffic at the UNI.
- To examine the instantaneous cell loss ratio by formal analysis, and to use it as a performance parameter for network access control at the UNI.
- To examine UNI intra-node buffer control and the appropriate conditions for this control to be applied.

The motivations for formulating the above objectives are discussed in the remainder of this section.

The goal of fast packet switching networks is to reduce the packet processing delays of conventional packet switching networks. As the implementation model of the fast packet switching network (e.g. ATM-based network) becomes more mature, it has stimulated more interest in researching effective congestion control techniques to further reduce packet delays due to network congestion.

In simple terms, if, for any time interval, the aggregate demand on a resource is more than its available capacity, the resource is said to be congested for that interval. In the case of fast packet switching networks, there is a large number of resources, such as buffers, transmission link bandwidths, processor times and so forth. In this thesis, the congestion issues associated with transmission links are examined. The network of links constitutes a distributed resource and congestion control schemes are designed to ensure that the aggregate demand at each link is less than its capacity. Congestion control techniques must minimize overhead since they are employed to reduce delay in the first place.

It is tempting to conclude that the tremendous capacity of fiber links solves the problem of link congestion automatically. However, past experience indicates that user demands always expand to fill the available capacity. One may also argue that the employment of virtual circuits and deterministic bandwidth allocation would eliminate the dynamic congestion problem. However, the preceding statement is true only if traffic generation processes are deterministic; or if traffic generation processes are stochastic and we are willing to sacrifice bandwidth utilization by allocating maximum stochastic limit.

To alleviate dynamic congestion, buffering is employed and consequently queueing delay is introduced. In essence, an ATM cell-switched network may be characterized as a network of queues. At each node, there is a queue of cells for each outgoing link. Under heavy traffic condition (cell arrival rate approaches transmission rate), queue length will grow dramatically. If the rate at which cells arrive and queue up exceeds the rate at which cells are transmitted, the queue size grows without bound and the average delay experienced by a cell goes to infinity.

The maximum queueing delay for each link can be bounded by limiting the buffer

size. When the buffer is full, additional incoming cells are discarded and lost. Consequently, the objective of congestion management is to optimize the trade-off between maximum queueing delay and cell lost rate while keeping overhead to a minimum.

Different types of traffic respond differently to the adversaries of cell delay and cell loss. For instance, the quality of voice traffic is more sensitive to cell delay than to cell loss; while the quality of data traffic is more sensitive to cell loss than to cell delay.

The possible congestion control techniques are outlined as follows:

- Subnet Congestion Control
  a. Node-to-node flow control (applicable to stream traffic, not appropriate for bursty traffic).
  b. End-to-End flow control (applicable to stream traffic, not appropriate for bursty traffic).
  c. Intra-node buffer control
- UNI Congestion Control
  a. Controls applied during call access phase of user traffic source: network access control.
  b. Controls applied during information transfer phase of user traffic source: source policing control, intra-node buffer control.

Since the reference architecture for the B-ISDN subnet is still in a state of flux with ongoing standardization activities, it is premature to study subnet congestion control at this point of time. On the other hand, the integrated access model of user traffic at the

B-ISDN UNI [5, 14, 15] is relatively stable in terms of multiplexing technique (ATM-cell statistical multiplexing) and user traffic characteristics at the UNI. Consequently, this thesis focuses on UNI congestion control; and a methodology of integrated congestion management at the UNI is proposed. Integrated congestion management means that congestion control schemes are applied during the call access phase and the information transfer phase of user traffic source.

The purpose of network access control is to maintain the required network performance assigned to the UNI access-node during the information transfer phase of user traffic operation by exerting admission control during the call access phase. References [16], [15] and [17] employed the concept of equivalent bandwidth of bursty traffic in their admission criterion: the sum of the equivalent bandwidths of connections on any given link should not exceed a suitable fraction of the link bandwidth so as to ensure an acceptable cell loss ratio. Hirano et al. [14] used long-term-time-averaged values of cell delay and cell loss ratio as criteria for admission control. Kamitake et al. [18] takes into account the dynamic nature of cell loss ratio in formulating the criterion for admission control.

Source policing control ensures that the source generates traffic according to the values negotiated during the call access phase. The policing control acts on each source before traffic from all sources are multiplexed. Rathgeb [19] compared various source policing mechanisms in terms of their dimensioning and effectiveness. Butto et al. [20] presents an analysis of the performance of the "Leaky Bucket" mechanism.

The purpose of intra-node buffer control is to optimize network performance by exerting control during the user information transfer phase. Up to now, there is no

significant research on this issue.

In this thesis, network access and intra-node buffer congestion control techniques appropriate for the ATM-based network are proposed and analyzed. UNI network access and buffer control schemes developed for the conventional packet-switched network are not applicable to the ATM-based network because of the different input traffic characteristics. In most previous investigations on the performance of conventional packet-switched networks, the individual input traffic is mostly computer-to-computer data; such individual and aggregate traffic are well-known to follow the Poisson process [21]. On the other hand, ATM-based network allows a variety of input traffic in addition to the Poisson-distributed data traffic, and the aggregate input traffic may no longer be described appropriately by the Poisson process.

The individual/aggregate traffic processes belong to the "counting" or "point" stochastic process class because these processes are integral-valued for each time interval and their sample paths are monotone nondecreasing functions of time. The subclasses of the "point" process class in order of ascending complexity are as follows: (1) Fluid Process; (2) Poisson Renewal Process; (3) General or Non-Poisson Renewal Process; (4) Doubly Stochastic Renewal Process (DSRP), e.g. Markov Modulated Poisson Process (MMPP) and Switched Batch Bernoulli Process (SBBP); (5) Stationary Point Process; and (6) Non-Stationary Point Process. In references 22–30, 14, 18, the researchers modelled the individual/aggregate input traffic processes by different "point" process subclasses ranging from Fluid Process to Doubly Stochastic Renewal Process.

The input traffic being multiplexed at the UNI may originate from the user application layer (user-to-user application traffic) or the transfer mode layer (user-to-network and

user-to-user signalling traffic). The signalling traffic must be serviced quickly since it directly affects call set up times and network efficiency. Up to now, all related researches including those mentioned above consider only the user traffic but ignored the signalling traffic. Consequently, the primary objective for this thesis is to study the effect of the higher-priority signalling traffic on the multiplexing of user traffic at the UNI.

In this thesis, individual user traffic process is modelled as a two-state MMPP [22, 23, 24, 26, 29]; the aggregate user-application traffic process is modeled as a Batch Bernoulli Renewal Process for light traffic condition and as a Fluid Process for heavy traffic condition. The individual and aggregate signalling traffic processes are assumed to be Poisson in nature. However, the analytical technique can be easily adapted to an aggregate signalling traffic process with general distribution. The modelling of user traffic is discussed in chapter 3.

As for the research on the technique of UNI network access control, most researchers [17, 14, 31] employ long-term-time-averaged value of cell loss ratio as a criterion to decide if a new call should be admitted. However, it may not be appropriate for an ATM-based network. When the input traffic is highly bursty in nature, the instantaneous cell loss ratio could be very high during congestion periods even with the long-term-time-averaged value of cell loss ratio being kept small. In [18], the instantaneous cell loss ratio was studied by simulation but a formal analysis is not available. Consequently, the second objective for this thesis is to examine the instantaneous cell loss ratio by formal analysis, and to use it as a performance parameter for network access control.

Previous investigations have generally ignored UNI intra-node buffer control. While the network access control ensures that the network performance of the UNI access-

node does not fall below the required network performance assigned to it, the intra-node buffer control is employed to optimize the network performance. Consequently, the third objective for this thesis is to examine UNI intra-node buffer control and the appropriate conditions for this control to be applied.

1.3 Approach

The approach to achieve the above research objectives is outlined as follows:

- Formulate an ATM-based network model encompassing the subnet (backbone transport network) and the UNI.
  a. Modeling subnet traffic transport
  b. Modeling UNI input traffic multiplexing
  c. Modeling the aggregate input traffic process resulting from various bearer services.
- Identify congestion related network performance parameters that can be monitored for network congestion across the subnet and the UNI.
- Identify state parameters at the UNI that can be controlled to counteract network congestion. Formulate a process model for each of the congestion related performance parameters in terms of the controllable state parameters at the UNI; and then analyze the process model.
- Develop congestion control models at the UNI to meet the basic performance requirements for the UNI; and to optimize performance if possible.

1.4 Structure of the Thesis

The remainder of this thesis is structured as follows. In Chapter 2, an ATM-based network model is formulated in terms of the subnet traffic transport model and the UNI traffic multiplexing service model. The servicing of user traffic at the UNI access-node under heavy traffic condition is modelled as a fluid process with random service disruptions due to the higher-priority signalling traffic. In Chapter 3, the characterizations of user traffic and signalling traffic at the UNI access-node are described. The arrival of user traffic at the UNI access-node under heavy traffic condition is modelled as a fluid process.

In this thesis, the congestion performance requirements for the UNI access-node are defined in terms of the stochastic cell loss requirement and the deterministic cell delay (upper-bound cell delay) requirement. The congestion control schemes are devised to meet the performance requirements and to optimize the performance if possible.

To investigate the stochastic cell loss phenomenon due to buffer overflow, the stochastic queue behaviour must first be examined. Therefore, in Chapter 4, the queue behaviour under heavy traffic condition is analyzed via the fluid model with random service disruptions. Consequently, in Chapter 5, the instantaneous cell loss phenomenon is defined and analyzed.

In Chapter 6, a UNI access-node congestion management model is defined, which integrates the call-level (network access) and cell-level (buffer) congestion control schemes. Call-level control is applied during the access phase of the user traffic source and it is devised to meet the performance requirements. On the other hand, cell-level control is applied during the information transfer phase of the user traffic source and it is devised to

optimize the performance. In Chapter 7, applications of the congestion control schemes to voice and data traffic, and the corresponding results are discussed. Chapter 8 concludes this thesis with a summary of research contributions and an outline of areas for further research.

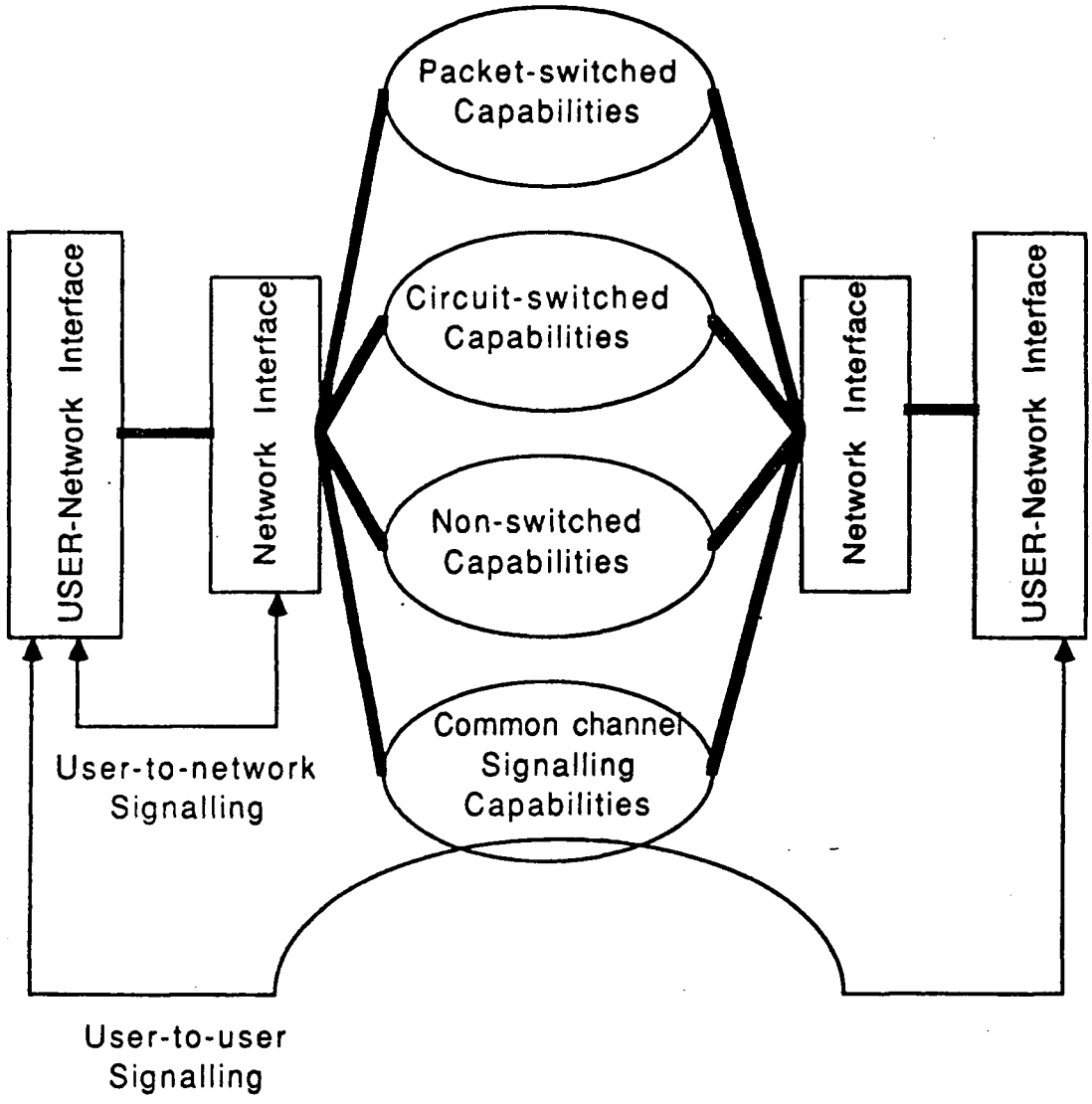
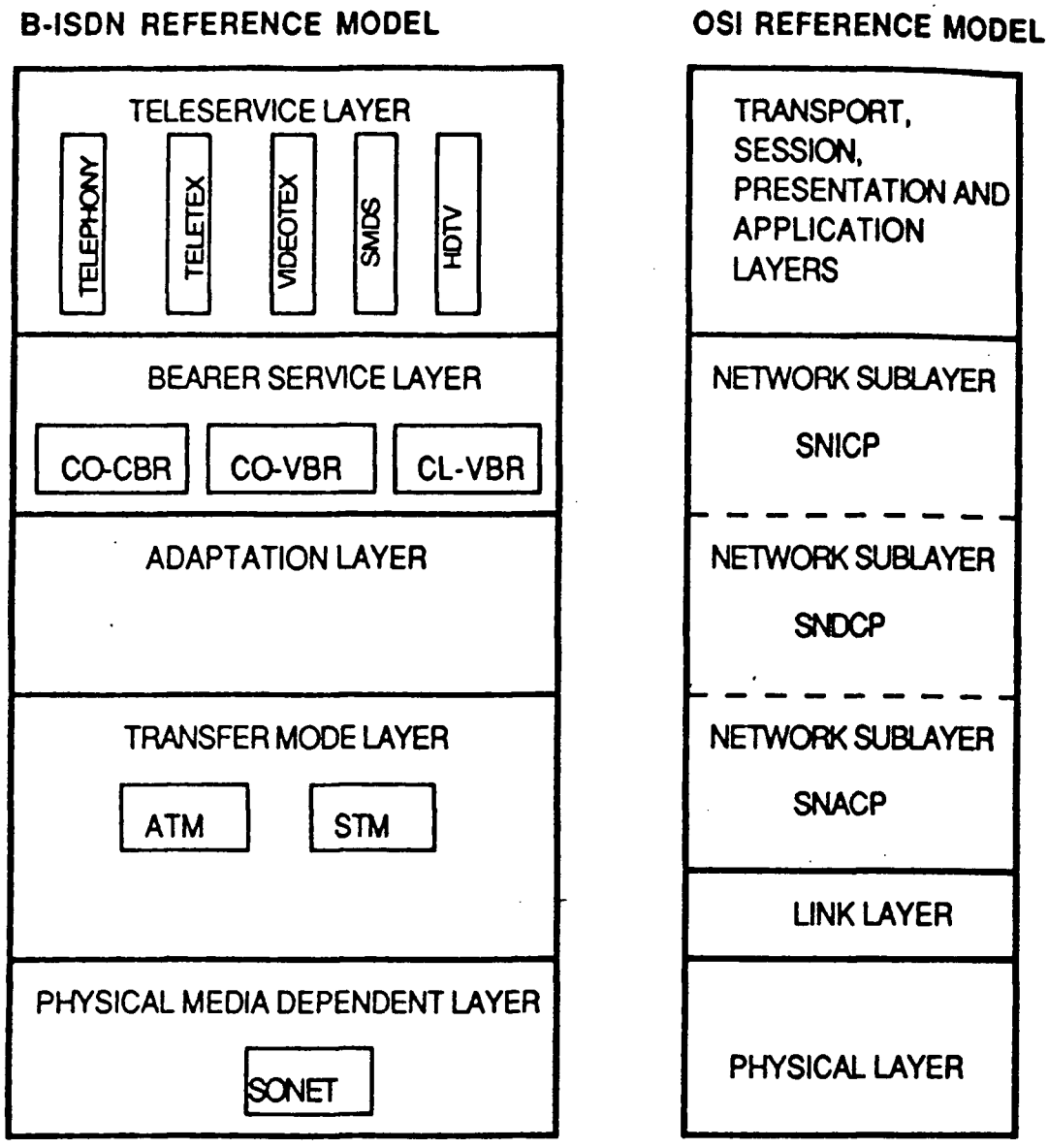


Fig. 1. ISDN Reference Architecture



SNICP: SubNetwork Independent Convergence Protocol
SNDCP: SubNetwork Dependent Convergence Protocol
SNACP: SubNetwork Access Convergence Protocol

CO-CBR: Connection-Oriented and Constant Bit Rate
CO-VBR: Connection-Oriented and Variable Bit Rate
CL-VBR: ConnectionLess-oriented and Variable Bit Rate

Fig. 2. B-ISDN Reference Model (with OSI Equivalency)

Chapter 2 ATM-Based Network Model

A B-ISDN complied network could be decomposed into layered models (bearer service layer, adaptation layer, transfer mode layer and the physical media dependent layer) of the subnet and the UNI. An ATM-based network is characterized by the employment of the ATM protocol in the transfer mode Layer of the B-ISDN protocol hierarchy. In this chapter, the subnet traffic transport model and the UNI traffic multiplexing model at the ATM layer are examined.

A hierarchical decomposition model of an ATM-based network is illustrated in Fig. 3. The transfer mode layer implemented by the ATM protocol is modelled as one of the hierarchical layer. In hierarchical model decomposition, the interaction of each layer with its neighbouring layers and the interaction of entities within the same layer are approximated using a few parameters. Each entity is modelled as an independent queue, Markov chain or deterministic entity. In the usual form of this approach, the interaction of an entity with other entities is specified by approximating its input and output processes.

If there is no end-to-end flow control, then either the bearer service layer or the adaptation layer can be modelled as an open queueing network with multiple chains of queues. The physical media dependent layer is modelled as a deterministic layer with fixed transmission rate and delay. The modeling of the ATM layer is described in the following sections in terms of the subnet and the UNI.

2.1 Subnet Traffic Transport Model

In this thesis, the cell-switched ATM subnet, consisting of ATM switching nodes interconnected by broadband channel facilities, is modelled as an open queueing network with multiple routing chains as illustrated in Fig. 4. The following assumptions are made:

- FCFS (First-Come-First-Served) servers are used to model communication channels.
- Channel propagation delay, transmission error and packet processing time in switching node are negligible due to the employments of fiber optic transmission facility and hardware switching.
- Kleinrocks's independence assumption [21] applies.
- Virtual circuits are established for each source-destination node pair before information transfer can occur.
- Each virtual circuit (without end-to-end window flow control) is modelled by an open chain of queues.

2.2 UNI Input Traffic Multiplexing Model

The model of input traffic multiplexing through the ATM statistical multiplexer at the UNI is illustrated in Fig. 5. The input traffic is classified into two priority groups: signalling traffic (higher priority) and user traffic (lower priority). The signalling traffic group carries call control and network management information. The user traffic group carries user originated information.

Each traffic group has its own queueing buffer of finite capacity. The two traffic queues are served by one cell output server (transmission link) of capacity μ cells per second. The input traffic is composed of cells of constant size (48 bytes data and 5 bytes header) and multiplexed onto the transmission link. Time is quantized in cell slot

duration which equals one cell transmission time. Cell multiplexing is synchronized with respect to cell slots.

While the signalling traffic queue is non-empty, cells from this queue are transmitted by the output server in a FIFO (First-In-First-Out) manner and the servicing of the user traffic queue is disabled. While the signalling traffic queue is empty, the servicing of the user traffic queue is enabled and user traffic cells are transmitted by the output server in a FIFO manner.

2.2.1 Cell Service Process of Signalling Traffic

The signalling traffic at the UNI access-node carries call control messages and network management messages originated from the user nodes. Call control messages must be transported quickly since they directly affect call establishment and network efficiency. Some types of network management messages can be urgent and asynchronous in nature if they are concerned with fault, reconfiguration and etc. Consequently, signalling traffic must have a higher transport priority than the user traffic at the UNI.

In this thesis, only call control signalling traffic is considered. For independent users, the aggregate call arrival is found to be Poisson in nature [21]. Each call arrival initiates the generation of a fixed length of signalling cells. Consequently, a M/D/1 discrete time queueing model (Markov chain) for the signalling traffic is employed. The queue is assumed to be stationary, i.e. in steady-state equilibrium.

Let I_{sig} be the random variable representing the stationary duration of the idle state of servicing signalling traffic, and $\overline{I}_{sig}$ be the average duration of the idle state. Then $P_{I_{sig}}(i)$ is the probability that the Markov chain visits any non-zero state for the first time in i cell slots given the initial condition that the queue is in the zero state. At any

given slot, it can be associated with either the idle state (not transporting a signalling cell) with probability p or the busy state (transporting a signalling cell) with probability $1-p$. Therefore, the resulting probability distribution of the stationary duration of the idle interval is geometrically distributed.

Let B_{sig} be the random variable representing the stationary busy interval of servicing signalling traffic, and $\overline{B}_{sig}$ be the average duration of the busy state. Then $P_{B_{sig}}(b)$ is the probability that the Markov chain visits the zero state for the first time in b cell slots given the initial condition that the queue is in a non-zero state.

2.2.2 Cell Service Process of User Traffic

The servicing of user traffic through the ATM statistical multiplexer at the UNI is a result of the non-disrupted user cell service being modulated by the *idle* and *busy* states of servicing signalling traffic; which correspond respectively to the *up* state and *down* state of the user cell service process.

During the *up* state, if the user traffic queue is non-empty, user traffic cells of uniform size are serviced at a uniform rate of μ cells per second. During the down state, user traffic cell service is suspended.

In this thesis, the servicing of user traffic under heavy traffic condition is modelled as a fluid process with random disruptions due to the servicing of the higher priority signalling traffic. The random disruptions are characterized by a stochastic sequence of up and down states, and the servicing is interrupted during the *down* state. The up or down state is described by an *i.i.d.* (independent and identically distributed) sequence of random variables, $\{(D_i,U_i), i \geq 1\}$, where D_i is the ith duration of the down state and U_i is the ith duration of the up state. Consequently, it is described by a stochastic cell service

process $S = \{S_i,\ i \geq 0\}$; where $S_{i=k}$ is a random variable denoting the number of user traffic cells (= 0 or 1) that can be serviced at cell slot k. The following indicator function

$$I(t) = \begin{cases} 0,\ T_i \leq t < T_i + D_{i+1} & ,\ i = 0, 1, .. \\ 1,\ T_i + D_{i+1} \leq t < T_{i+1} \end{cases} \tag{1}$$

where, T_i is the time at which the ith down duration starts.

specifies down state when $I(t) = 0$ and *up* state when $I(t) = 1$.

The observable parameters of the user cell service process result in the following statistics:

- Average duration of the stationary *down* state of user traffic servicing: $\overline{D}$.
  (It is equivalent to the average duration of the stationary *busy* state of signalling traffic servicing: $\overline{B}_{sig}$.)
- Average duration of the stationary *up* state of user traffic servicing: $\overline{U}$.
  (It is equivalent to the average duration of the stationary *idle* state of signalling traffic servicing: $\overline{I}_{sig}$.)

At steady-state condition, $S = \lim_{i \to \infty} S_i$. Then $P_S(s)$ is the probability that s cells are serviced at a slot where $0 \leq s \leq 1$, and it is given as follows:

$$P_S(0) = \frac{\overline{D}}{\overline{D} + \overline{U}} \quad \text{and} \quad P_S(1) = \frac{\overline{U}}{\overline{D} + \overline{U}} \tag{2}$$

Then $\overline{S}$ the average number of cells that are serviced at a slot is given as $\frac{\overline{U}}{\overline{D} + \overline{U}}$. On the other hand, $P_S(0)$ and $P_S(1)$ can also be interpreted as the steady state probabilities of the user cell service being *down* and *up* respectively.

2.3 Congestion Related Network Performance Parameters

To identify the network performance parameters that can be monitored for congestion across the subnet and the UNI, it is necessary to explain the concept of network performance and the QOS (Quality of Service) of teleservices and bearer services.

Bearer services provide the means to transport information (speech, data, video, etc.) between users in real time and without alteration of the content of the messages, and they can be considered as corresponding to the OSI layer 3. Teleservices combine the information transport and processing functions; and they can be considered to correspond to the layers above and including OSI layer 4.

The following outlines the differences between QOS and network performance:

1. User-oriented QOS of a teleservice is the acceptable quality of service from the user's point of view and includes subjective characteristics such as noise, error and delay.
2. Network-oriented QOS of a bearer service is the quality of bearer service that is necessary to fulfill the requested user-oriented QOS.
3. Network performance in supporting a bearer service is the combined performance requirements of all network elements supporting the bearer service. Network performance will often be much higher than necessary to fulfil the network-oriented QOS. Performance required for each individual network element (e.g. UNI access-node, network switching-node) can be derived from the network-oriented QOS and the network architecture.

The network performance of a circuit-switched network for circuit-mode bearer service can be characterized by the following parameters[32]: (1) during call access phase: dialing delay, call blocking; (2) during user information transfer phase: propagation

delay, idle noise, return echo loss, variable speech burst delay, cross talk, interruptions and transmission errors.

On the other hand, the network performance of a packet-switched network for packet-mode connection-oriented bearer service can be characterized by the following parameters: (1) during call access phase: call set-up delay and error; (2) during user information transfer phase: data packet transfer delay, throughput, residual error rate, reset probability and premature disconnect probability.

The network performance of a cell-switched ATM network for integrating both circuit-mode and packet-mode connection-oriented bearer services can be characterized by the following parameters [32]:

- Cell Loss Ratio (CLR) — It is the ratio of the number of lost cells to the total number of cells entering the connection. Cell loss can occur due to detected header errors or buffer overflow.
- Cell Insertion Ratio (CIR) — It is the ratio of the number of inserted cells to the total number cells entering the connection. Cell insertion can occur due to undetected header errors.
- Cell Information Field Bit Error Ratio (BER) — It is the ratio of the number of bit errors in the cell information field to the total number of bits in the information field.
- Cell Delay — It is the time which passes between the entrance of a cell in the network and its exit. Cell delays are caused by queueing delay, processing delay and transmission delay.
- Cell Delay Jitter — It is the difference between the maximum and the minimum of the end-to-end cell delay.

In this thesis, the cell loss ratio and the upper limit of cell delay are monitored for congestion controls. Cell loss is assumed to be caused primarily by buffer overflow; cell delay is assumed to be caused primarily by queueing delay. Network performance requirements depend on the type of teleservice being supported, since users respond differently to the adversaries of cell delay and cell loss ratio for different teleservices. For instance, for voice telephony teleservice which requires voice circuit-mode bearer service, users are more sensitive to cell delay than to cell loss ratio. On the other hand, for data handling teleservice which requires data packet-mode bearer service, users are more sensitive to cell loss ratio than to cell delay.

Performance requirements for ATM-based network elements (e.g. UNI access-node, switching node) can be derived from the values of the ATM specific QOS subattributes of the bearer services by means of Reference Connections. The reference connections for voice telephony and data handling teleservices illustrated in Fig. 6 represent the longest connections and have been derived from existing CCITT Recommendation G.104.

The estimated required values of the network-oriented QOS subattributes (CLR and cell delay) for each bearer service, and the estimated required values of the network performance parameters (CLR and cell delay) for each ATM-based network element are outlined in Table 1 and Table 2, respectively.

Bearer Service	Network-Oriented QOS Subattributes	
	Cell Loss Ratio	Cell Delay (ms)
Voice Circuit-Mode (64 Kbps)	10^{-3}	160
Data Packet-Mode (10 Mbps)	10^{-6}	200

Table 1 Network-oriented QOS Requirements for Bearer Services [32]

Bearer Service	Network Performance Parameters	
	Cell Loss Ratio	Cell Delay (ms)
Voice Circuit-Mode (64 Kbps)	10^{-4}	1
Data Packet-Mode (10 Mbps)	10^{-7}	4

Table 2 Performance Requirements for ATM-Based Network Elements [32]

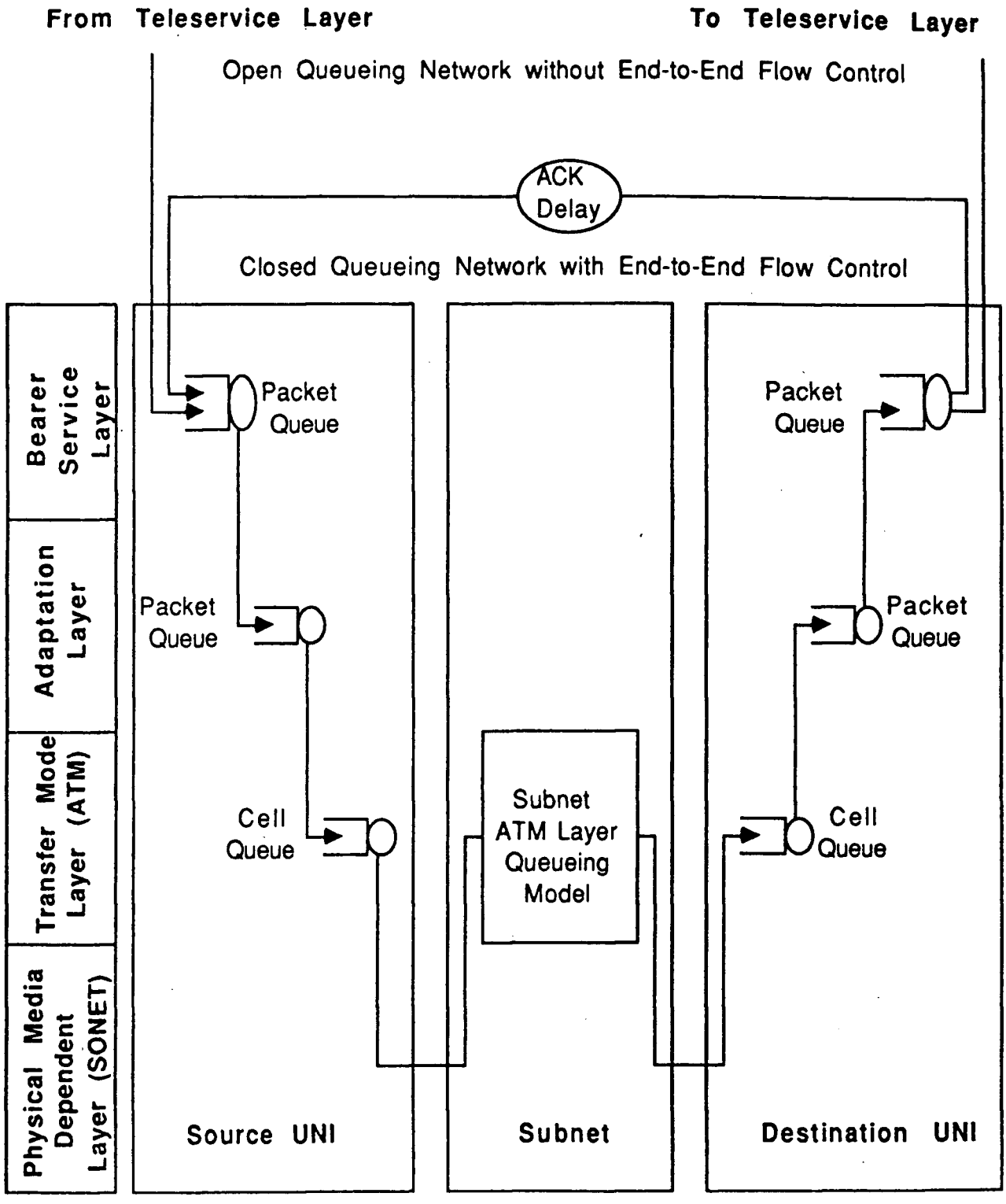


Fig. 3. B-ISDN Hierarchical Decomposition Model

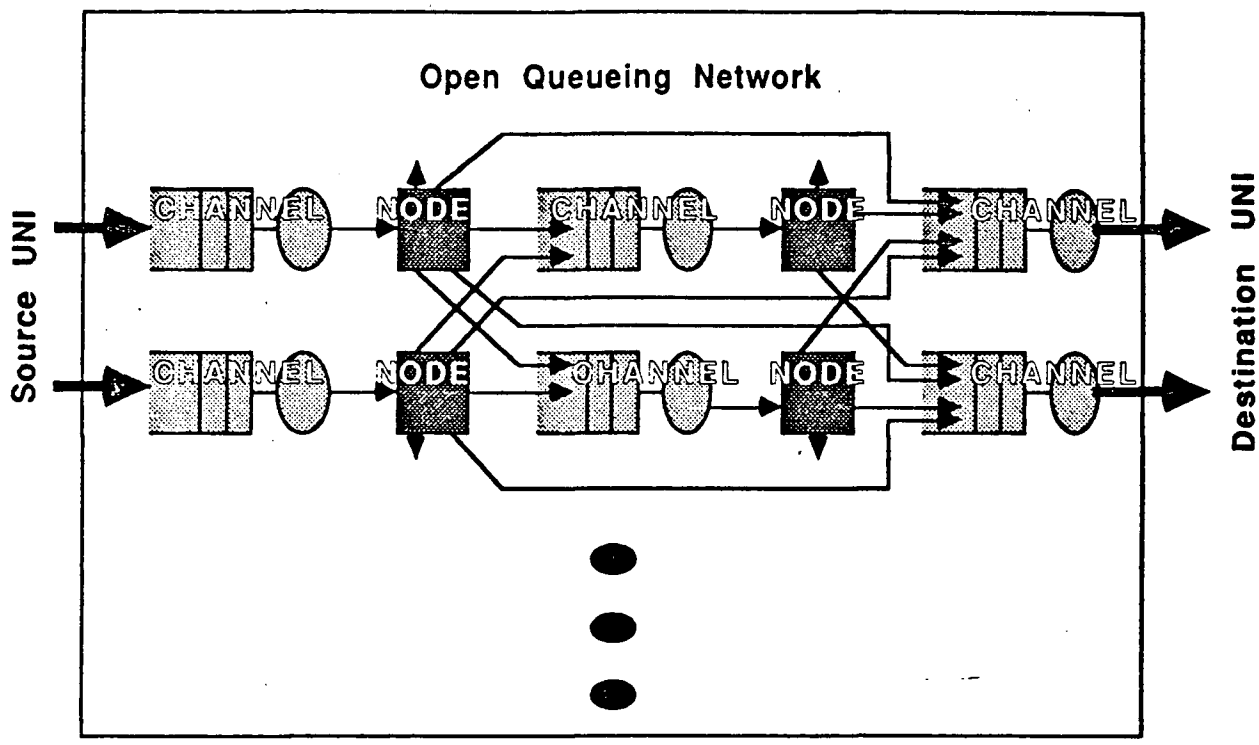


Fig.4. Subnet ATM Layer Queueing Model

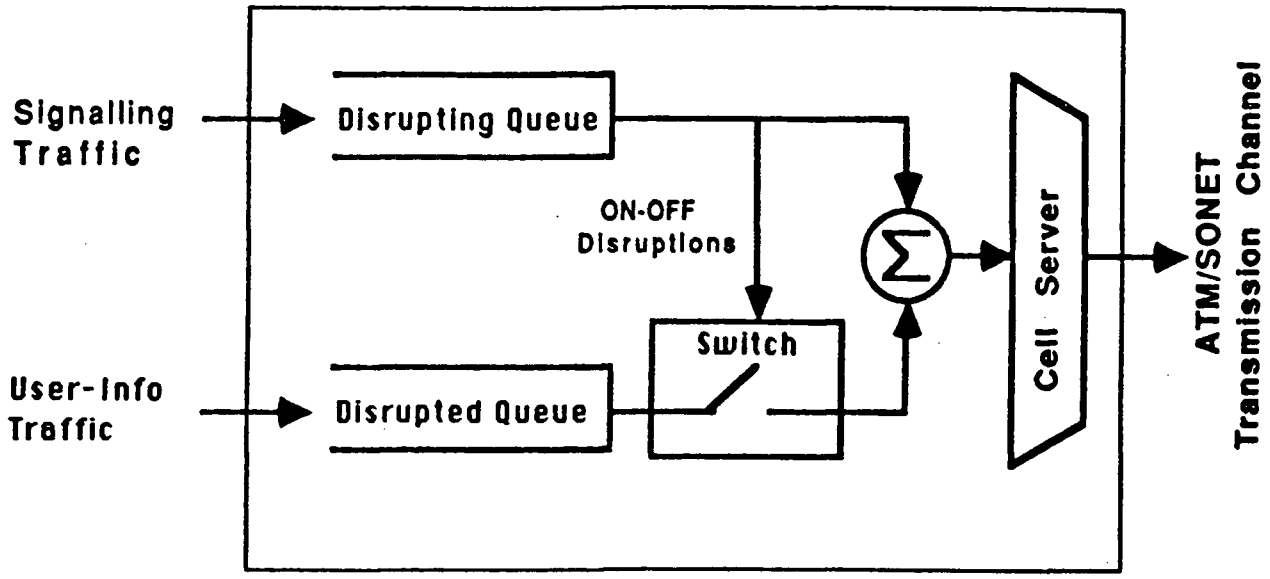
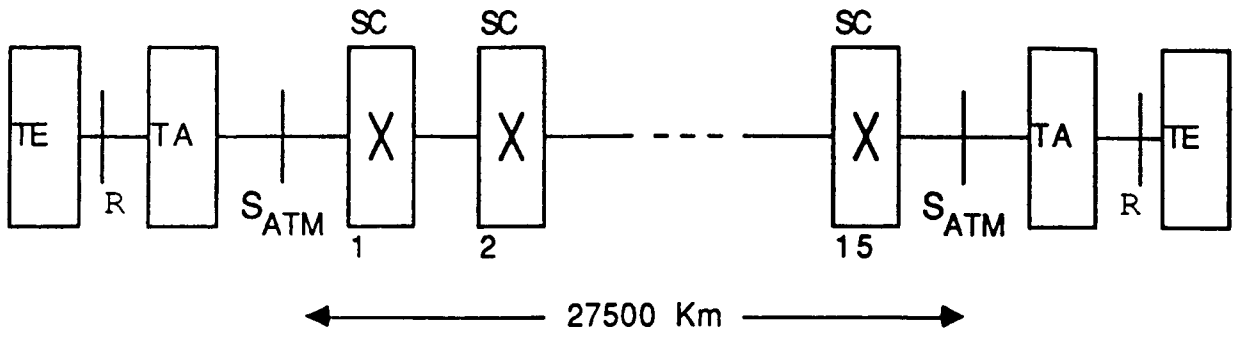


Fig. 5. UNI Traffic Multiplexing Model



TE: Terminal Equipment
TA: Terminal Adaptor
SC: Switching Center

Reference point R (rate): Provides a non-ISDN interface between user equipment (non-ISDN compatible) and adaptor equipment.

Reference point S (system): Corresponds to the interface of individual ISDN terminals, it separates user terminal equipment from network-related communications functions.

Fig. 6. Reference Connection for Voice Telephony and Data Handing Teleservices

Chapter 3 Modeling of Input Traffic to ATM-Based Network

This chapter examines the characterization of UNI input traffic process associated with the ATM transfer mode layer. The traffic process associated with each of the protocol layers can be characterized by appropriate traffic variables. Since teleservice layer combines the information processing function with the information transfer function provided by the bearer service layer, the traffic processes associated with these two layers can be described by the same set of traffic variables. The traffic variables associated with teleservice or bearer service layer can be classified by the following scopes:

- Call Access: Access and disengagement phases variables (e.g. average rate of call attempts).
- Cell Transfer: Information transfer phase variables (e.g. mean call connection time).
- Cell Generation: Source generation variables (e.g. data rate).

As for the ATM transfer mode layer, the associated traffic process is characterized by the information transfer phase variables; i.e. it is a subset of the associated traffic process of the teleservice or bearer service layer. For connection-oriented teleservice, it is an essential feature of an ATM-based network that the bit rates of user traffic can be variable during a call. Consequently, individual cell arrival process generated by a traffic source is generally classified by its bit rate variability as follows: (illustrated in Fig. 7)

- Constant Bit Rate (CBR)

  (e.g. call control and network management signalling traffic sources)

- Variable Bit Rate (VBR)

  a. Discrete on/off bi-states (e.g. voice telephony source).

  b. Discrete multiple states (e.g. multi-media and VBR video traffic sources).

  c. Continuous varying states (e.g. multi-media and VBR video traffic sources).

While the CBR traffic process can be described appropriately as deterministic process, the VBR traffic process has to be described as a stochastic (random) process. The modeling of the UNI input traffic process associated with the ATM protocol layer resulting from user teleservice traffic and signalling traffic will be discussed in the following sections.

3.1 User Traffic Modeling

Most user teleservice sources generate VBR traffic which has to be described by a stochastic process. In this thesis, the stochastic input traffic process at the UNI is described in terms of the cell-arrival random process $\{A(t);\ t \geq 0\}$; where $A(t=\tau)$ is a random variable representing the number of arrivals in time interval $[0, \tau]$.

The superposition of individual cell-arrival processes (homogeneous or heterogeneous) results in an aggregate cell-arrival process. The aggregate cell-arrival process is stochastically more complicated than the individual cell-arrival process because of the potential dependencies or correlations among individual processes.

The cell-arrival random process $\{A(t);\ t \geq 0\}$ has two properties: (1) random variable $A(t=\tau)$ is integral-valued; and (2) sample path $A(t)$ is a monotone nondecreasing function of t. Stochastic processes with these properties belong to the "counting" or "point" stochastic process class [33].

A point process class can be further divided into subclasses according to the following set of stochastic parameters:

- Arrival increments (number of arrivals in disjoint intervals): independent or dependent.
- Arrival increment distribution: stationary (increment distribution is independent of the epoch where the increment is located) or non-stationary.
- Arrival order (number of arrivals in an interval of length h as $h \rightarrow 0$): orderliness (one-at-a-time) or batch.

The subclasses of the point process class in ascending order of complexity are as follows: (1) Poisson renewal Process; (2) general or non-Poisson renewal process; (3) Doubly Stochastic Renewal process (DSRP) [33, 34], e.g. Markov Modulated Poisson Process (MMPP) and Switched Batch Bernoulli Process (SBBP); (4) non-stationary point process.

The mapping of the properties of the stochastic parameters for each point process subclass is illustrated in Table 3.

Point Process Subclass	**Stochastic Parameters**		
	Arrival Increment	**Arrival Increment Distribution**	**Arrival Order**
Poisson Renewal	Independent	Stationary (Poisson)	Orderliness/Batch
General Renewal	Independent	Stationary (Non-Poisson)	Orderliness/Batch
Doubly Stochastic Renewal	Dependent	Stationary	Orderliness/Batch
Non-stationary Point	Independent/ Dependent	Non-stationary	Orderliness/Batch

Table 3 Point Process Subclasses Classification

The modelling of individual and aggregate cell-arrival processes at the UNI by means of approximate point processes will be discussed in the following subsections.

3.1.1 Modelling of Individual Cell-Arrival Process

For connectionless-oriented teleservice, e.g. computer-to-computer data, the input traffic at the UNI is usually assumed to follow a Poisson distribution [21]. For connection-oriented teleservice, the bit rate of the input traffic at the UNI would be either (1) stream (CBR), e.g. call control or network management signalling traffic; (2) bursty (VBR with on/off states), e.g. voice telephony; (3) piecewise-constant-rate (VBR with multiple states), e.g. video telephony.

Previous researches [22, 23, 26] have shown that MMPP is appropriate for modelling cell-arrival processes from VBR sources: bi-state MMPP for modelling cell-arrival processes from bursty traffic sources and multiple-state MMPP for modelling cell-arrival processes from piecewise-constant-rate traffic sources. These researches also showed that multiple-state MMPP could be approximated by bi-state MMPP for analysis simplification.

This thesis focuses on bursty traffic sources and the bi-state MMPP is employed for modelling the corresponding cell-arrival process. The imbedded bi-state bit rate process of such cell-arrival process is determined by a bi-state Markov chain as illustrated in Fig. 8. The cell-arrival process modelled by a bi-state MMPP has the following common features: (1) it is characterized by two call states, "active" or "idle", and the state holding times are geometrically distributed; (2) it generates cells only when it is in the "active" state, and the cell interarrival time during the "active" state is geometrically distributed.

The statistical parameters for characterizing the cell-arrival process modelled by bi-state MMPP are defined as follows:

- Transitional probabilities between call states —

  Transitional probability from active to idle state α; and transitional probability from idle to active state β.

- Steady state probabilities of call states —

  At each cell slot, the *i-th* call is either active or idle. Call active status is represented by a random variable X_i where $i = 1, 2, \cdots, N$ and N is the number of calls accessing the network through the UNI; with

  $P_{X_i}(0)$ = Probability that the ith call is idle at a slot

  $P_{X_i}(1)$ = Probability that the ith call is active at a slot.

- Steady state probabilities of cell generation during call active state —

  When the *i-th* call is active, cell arrival at UNI resulting from cell generation at each slot is defined by a random variable Y_i; with

  $P_{Y_i}(0)$ = Probability that the ith call not generating a cell at a slot

  $P_{Y_i}(1)$ = Probability that the ith call generates a cell at a slot

The long-term-time-averaged observable parameters, illustrated in Fig. 9, are defined as follows:

- Average duration of call active state $\overline{T_A}$.
- Average duration of call idle state $\overline{T_I}$.
- Average cell arrival rate during call active state $\overline{R_C}$.
- Average interarrival time between cells $\overline{T_C} = \frac{1}{\overline{R_C}}$
- Maximum cell service rate μ.

The statistical parameters of the model are derived from the observable parameters as follows:

- Steady state probabilities of call states

$$P_X(1) = \frac{\overline{T_A}}{\overline{T_A}+\overline{T_I}}; \qquad P_X(0) = 1 - P_X(1) = \frac{\overline{T_I}}{\overline{T_A}+\overline{T_I}}$$

- Steady state probabilities of cell generation during call active state

$$P_Y(1) = \frac{\overline{R_C}}{\mu}; \qquad P_Y(0) = 1 - P_Y(1) = 1 - \frac{\overline{R_C}}{\mu} \text{ for } \overline{R_C} \leq \mu$$

3.1.2 Modelling of Aggregate Cell-Arrival Process

Table 4 illustrates the general point process characteristics of the aggregate cell-arrival process resulting from the superposition of individual cell-arrival processes, and the corresponding specialized point process appropriate for modelling.

Individual Cell-Arrival Process Modelling	**Stochastic Parameters of Aggregate Cell-Arrival Process**			**Aggregate Cell-Arrival Process Modelling**
	Arrival Increment	**Arrival Increment Distribution**	**Arrival Order**	
(PR)	Independent	Stationary (Poisson)	Batch	Poisson Renewal (PR)
(GR)	Independent	Stationary (Non-Poisson)	Batch	General Renewal (GR)
(PR), (GR), (DSR)	Dependent	Stationary	Batch	Doubly Stochastic Renewal (DSR)
(PR), (GR), (DSR), (NSP)	Dependent/ Independent	Non-stationary	Batch	Non-Stationar Point (NSP)

Table 4 Modelling of Aggregate Cell Arrival Process

Doubly Stochastic Renewal Process modelling [22, 23, 24, 26, 29] accounts for the property of dependent arrival increment. On the other hand, the simpler Poisson Renewal Process modelling [30, 14, 18] ignores this property. However, Sriram et al. [35] showed that Poisson Renewal Process modelling is a good approximation when the multiplex load is not too high. On the other hand, fluid modelling [24] is appropriate for heavy traffic condition.

Heffes et al. [22] modelled individual arrival processes as general (non-Poisson) renewal processes; and the aggregate arrival process as a bi-state MMPP. The process parameters of the two-state MMPP are estimated from a simpler Poisson renewal process. They derived the four parameters (state transition rate and arrival rate for each state) of the MMPP process from the four parameters (mean arrival rate, variance-to-mean ratio in short term, variance-to-mean ratio in long term and the third moment in short term) of the Poisson renewal process.

Saito et al. [23] modelled both individual and aggregate arrival processes from packetized voice sources as two-state MMPPs, and from packetized video sources as multi-state MMPPs. Nagarajan et al. [24] modelled both individual and aggregate arrival processes in three different ways: i.e. Poisson renewal processes, two-state MMPPs and fluid processes.

Norros et al. [25] modelled both individual and aggregate arrival processes as a two-component process: fluid process to describe the long-term time component and DSRP to describe the short-term time component. Baiocchi et al. [26] modelled both individual and aggregate arrival processes as two-state MMPPs.

Hahsida et al. [27] modelled both individual and aggregate arrival process as two-

state SBBPs. Dron et al. [28] modelled both individual and aggregate arrival processes in two different ways: Poisson renewal processes and DSRPs.

Le Boudec [29] modelled individual arrival processes from voice sources as Poisson (Bernoulli) renewal processes and from video sources as two-state MMPPs; and the aggregate arrival process as a multi-state MMPP. Kroner et al. [30] modelled both individual and aggregate arrival processes as Poisson Renewal Processes.

Hirano et al. [14] and Kamitake et al. [18] modelled individual arrival processes as two-state MMPP; and the aggregate arrival process as a general (non-Poisson) renewal process.

In this thesis, individual cell-arrival process from bursty sources are modelled as bi-state Markov modulated Poisson processes (MMPPs). The aggregate cell-arrival process resulting from bursty traffic sources is modelled as a fluid process for heavy traffic condition.

When the individual cell-arrival processes from bursty sources are modelled by bi-state MMPPs, the resulting aggregate cell arrival process in general is characterized by the following stochastic properties: (1) the arrival increment is dependent because the instantaneous cell arrival rate varies relatively slowly with burst arrivals and departures, tending to produce a positive correlation between the numbers of cell arrivals in successive slot intervals; (2) arrival may come in batch, i.e. more than one arrival in one slot interval.

Accordingly to Table 4, non-stationary point process should be employed to model the aggregate cell arrival process with the above stochastic properties. However, the analysis of non-stationary point process is intractable in general. Therefore, approximation models such as simpler point process, diffusion approximation process and fluid process have to

be employed in order to obtain tractable analysis.

When the aggregate cell-arrival process $A = \{A(t),\ t \geq 0\}$ is modelled as a MMPP, it can be shown that the MMPP is the resultant sum of a series of MMPPs. Let $G^{(N)} = \left\{ G_{i,}^{(N)}\ i \geq 0 \right\}$ be the cell generation (per slot) process modelled as a MMPP; where $G_{i=k}^{(N)}$ is the random variable denoting the number of cells generated by N calls at slot k. Then

$$A_{t=\tau} = G_1 + G_2 + \cdots G_{\frac{\tau}{h}} \tag{3}$$

where h is the duration of each cell slot.

The cell generation process is determined by the imbedded active-call (per slot) process $V^{(N)} = \left\{ V_i^{(N)},\ i \geq 0 \right\}$; where $V_{i=k}^{(N)}$ is the random variable denoting the number of calls being in the active state out of N calls at slot k. Recall that a call is either idle or active, and the transitions between idle and active states form a bi-state Markov chain as illustrated in Fig. 8.

The statistics required to describe the aggregate cell-arrival process as a fluid process are: (1) the mean number of active calls per slot, $\overline{V}$; (2) the mean cell-arrival rate per slot λ. The derivations of these statistics will be shown in the following subsection.

3.1.2.1 Mean Number of Active Calls per Slot

The statistical parameters characterizing an individual cell-arrival process from a

bursty source as a MMPP are as follows:

α $\equiv$ Transitional Probability from active to idle call state.

β $\equiv$ Transitional Probability from idle to active call state.

$P_{X_i}(0) \equiv$ Prob. the *ith* call is idle at a slot, $\mathrm{i} = 1, 2, \cdots, \mathrm{N}$.

$P_{X_i}(1) \equiv$ Prob. the *ith* call is active at a slot, $\mathrm{i} = 1, 2, \cdots, \mathrm{N}$.

$P_{Y_i}(0) \equiv$ Prob. the *ith* call does not generate a cell at a slot, $\mathrm{i} = 1, 2, \cdots, \mathrm{N}$.

$P_{Y_i}(1) \equiv$ Prob. the *ith* call generates a cell at a slot, $\mathrm{i} = 1, 2, \cdots, \mathrm{N}$.

$N \equiv$ Number of calls.

(a) Homogeneous Traffic Sources

For homogeneous traffic sources, $P_{X_i}(x) = P_X(x)$ and $P_{Y_i}(x) = P_Y(x)$ for $1 \leq i \leq N$.

Let $V^{(N)}$ be the random variable representing the number of active calls at a slot out of N calls. Then $P_{V^{(N)}}(v)$ is the probability that v calls are active at a slot out of N calls and it is given as follows:

$$P_{V^{(N)}}(v) = \binom{N}{v} P_X(1)^v P_X(0)^{N-v} \tag{5}$$
$$= \left[\frac{N!}{v!(N-v)!}\right]\left[\frac{\overline{T_A}}{\overline{T_A}+\overline{T_I}}\right]^v\left[\frac{\overline{T_I}}{\overline{T_A}+\overline{T_I}}\right]^{N-v}$$

where $\overline{T}_A$ and $\overline{T}_I$ are the average durations of the active and idle states, respectively, of a typical call.

Then, the mean number of active calls per slot is given as follows:

$$\overline{\mathrm{V}} = \mathrm{E}\left[V^{(N)}\right] = \sum_{v=0}^{N} v P_{V_N}(v) \tag{6}$$
$$= \sum_{v=0}^{N} v\left[\frac{N!}{v!(N-v)!}\right]\left[\frac{\overline{T_A}}{\overline{T_A}+\overline{T_I}}\right]^v\left[\frac{\overline{T_I}}{\overline{T_A}+\overline{T_I}}\right]^{N-v}$$

(b) Heterogeneous Traffic Sources

For heterogeneous traffic sources, in general $P_{X_i}(x) \neq P_{X_j}(x)$ for $i \neq j$. Therefore the probability $P_{V^{(N)}}(v)$ has to be obtained from the following recursive relationship:

$$P_{V^{(N)}}(v) = P_{V^{(N-1)}}(v-1)P_{X_N}(1) + P_{V^{(N-1)}}(v)P_{X_N}(0); \quad (7)$$

$$\text{for } 0 \leq v \leq \text{N}, \quad \text{where } P_{V^{(1)}}(v) = \begin{cases} P_{X_1}(0), & v = 0 \\ P_{X_1}(1), & v = 1 \\ 0 \quad , & v > 1 \end{cases}$$

Then, the mean number of active calls per slot is given as follows:

$$\overline{\text{V}} = \text{E}\left[V^{(N)}\right] = \sum_{v=0}^{N} v P_{V_N}(v) \quad (8)$$

3.1.2.2 Mean Number of Cell Arrivals per Slot

There are two approaches to estimate the statistics of the mean number of cell arrivals per slot: (1) mean of a limiting or stationary distribution; (2) time average.

(a) Mean of the Limiting or Stationary Distribution

Let G be the random variable representing the number of cell arrivals at a slot from N calls. Then $P_{G|V^{(N)}}(g \mid v)$ is the conditional probability that g cells are generated at a slot from v active calls out of N calls. For homogeneous traffic sources, it is given as follows:

$$P_{G|V^{(N)}}(g \mid v) = \binom{v}{g} P_Y(1)^g P_Y(0)^{v-g} \quad (9)$$

$$= \left[\frac{v!}{g!(v-g)!}\right]\left[\frac{\overline{R}_C}{\mu}\right]^g \left[1 - \frac{\overline{R}_C}{\mu}\right]^{v-g}$$

where $\overline{R_C}$ is the average cell arrival rate during the call active state, and μ is the maximum cell service rate.

For heterogeneous traffic sources, it is given as follows in a recursive relationship:

$$P_{G|V^{(N)}}(g \mid v) = P_{G|V^{(N-1)}}(g-1 \mid v)P_{Y_N}(1) + P_{G|V^{(N-1)}}(g \mid v)P_{Y_N}(0) \qquad (10)$$

$$\text{for } 0 \le g \le N \text{ and where } P_{G|V^{(1)}}(v) = \begin{cases} P_{Y_1}(0), & v = 0 \\ P_{Y_1}(1), & v = 1 \\ 0 \quad\quad , & v > 1 \end{cases}$$

Let $P_G(g)$ be the unconditional probability that g cells are generated at a slot from N calls. For homogeneous traffic sources, it is given as follows:

$$P_G(g) = \sum_{v=0}^{N} P_{G|V^{(N)}}(g \mid v)P_{V^{(N)}}(v) \qquad (11)$$

$$= \sum_{v=0}^{N} \left\{ \left[\frac{v!}{g!(v-g)!}\right] \left[\frac{\overline{R_A}}{\mu}\right]^g \left[1 - \frac{\overline{R_A}}{\mu}\right]^{v-g} \left[\frac{N!}{v!(N-v)!}\right] \left[\frac{\overline{T_A}}{\overline{T_A} + \overline{T_I}}\right]^v \left[\frac{\overline{T_I}}{\overline{T_A} + \overline{T_I}}\right]^{N-v} \right\}$$

For heterogeneous traffic sources, it is given as follows:

$$P_G(g) = \sum_{v=0}^{N} P_{G|V^{(N)}}(g \mid v)P_{V^{(N)}}(v) \qquad (12)$$

$$= \sum_{v=0}^{N} \left[P_{G|V^{(N-1)}}(g-1 \mid v)P_{Y_N}(1) + P_{G|V^{(N-1)}}(g \mid v)P_{Y_N}(0)\right] \times$$

$$[P_{V^{(N-1)}}(v-1)P_{X_N}(1) + P_{V^{(N-1)}}(v)P_{X_N}(0)]$$

Let λ be the mean number of cell arrivals per slot from N calls. It is given by the mean of the stationary distribution as follows:

$$\lambda = \mathrm{E}[G] = \sum_{g=0}^{N} gP_G(g) \qquad (13)$$

(b) Time Average

Since the derivation of the mean cell-arrival rate λ from the mean of the stationary distribution does not result in a convenient closed form expression, the derivation of λ from the time average in terms of continuous-time domain and elementary renewal theorem will now be attempted.

λ^k of the k_{th} individual cell-arrival process will be derived via elementary renewal theorem. Let $\left\{X_j^k\right\}$ be a sequence of independent, non-negative random variables of interarrival times between cells. Let $\overline{X}^k$ be the mean interarrival time; i.e. $\mathrm{E}\left[X_j^k\right] = \overline{X}^k$ Let $M^k(t)$ be the number of renewal or cell arrivals that have occurred in epoch [0, t], $t \geq 0$.

Applying the elementary renewal theorem :

$$\lim_{t \to \infty} \frac{M^k(t)}{t} = \frac{1}{\overline{X}^k} \quad \text{(with probability one) and} \qquad (14)$$
$$\lim_{t \to \infty} \frac{\mathrm{E}\left[M^k(t)\right]}{t} = \frac{1}{\overline{X}^k}$$

Then the mean cell arrival rate is given as follows:

$$\lambda^k = \frac{\mathrm{E}\left[M^k(t)\right]}{t} = \frac{1}{\overline{X}^k} \qquad (15)$$

Referring to Section 3.1.1 for the definitions of $\overline{T}_A$, $\overline{T}_I$, $\overline{R}_C$ and $\overline{T}_C$,

$$\lambda^k = \frac{1}{\overline{X}^k} = \overline{R_C}\left[\frac{\overline{T_A}}{\overline{T_A} + \overline{T_I}}\right] = \overline{R_C}\left[\frac{1}{(1 + \overline{T_I}/\overline{T_A})}\right] \qquad (16)$$

The average durations of call idle and active states are given as follows:

$$\overline{T_I} = \overline{T_C} \sum_{i=1}^{\infty} i\beta(1-\beta)^{i-1} = \overline{T_C}/\beta \qquad (17a)$$

$$\overline{T_A} = \overline{T_C} \sum_{i=1}^{\infty} i\alpha(1-\alpha)^{i-1} = \overline{T_C}/\alpha \qquad (17b)$$

The superposition of cell-arrival processes from N traffic sources will now be considered. For homogeneous traffic sources, the mean cell-arrival rate λ of the aggregate cell-arrival process is given as

$$\lambda = N\lambda^k = N\overline{R_C}\left[\frac{1}{(1+\alpha/\beta)}\right] = N\overline{R_C}\left[\frac{1}{(1+\overline{T_I}/\overline{T_A})}\right] \tag{18}$$

For heterogeneous traffic source,

$$\lambda = \sum_{k=1}^{N} \lambda^k \tag{19}$$

3.2 Signalling Traffic Modeling

The signalling traffic at the UNI access-node carries the following messages originated from the user nodes:

- Call control messages.
- Network management messages.

Call control messages go through switching systems via permanent virtual circuits to establish and release call connections (non-permanent virtual circuits) for transporting user traffic. Network management messages go to distributed or centralized network management centers to update information on congestion, fault, billing, configuration and etc.

Call control messages must be transported quickly since they directly affect call establishment time and network efficiency. Some types of network management message can be urgent and asynchronous in nature if they are concerned with fault, reconfiguration

and etc. Other types of network management messages can be less time critical and synchronous in nature, e.g. regular status report. Consequently, signalling traffic must have a higher transport priority than the user traffic at the UNI.

In this thesis, only call control signalling traffic is considered (the contribution of network management messages to the signalling traffic is ignored). For independent users, the aggregate call arrival of call control signalling traffic is found to be Poisson in nature [21]. In terms of the discrete time scale of cell slot, the following assumptions are made with regard to the aggregate arrival of call control signalling messages: (1) geometric distribution for the interarrival time; (2) deterministic or fixed message length.

During the transfer of a signalling message, the cell input rate at the UNI is of CBR (Constant Bit Rate) in nature and the output rate is determined by the access link capacity.

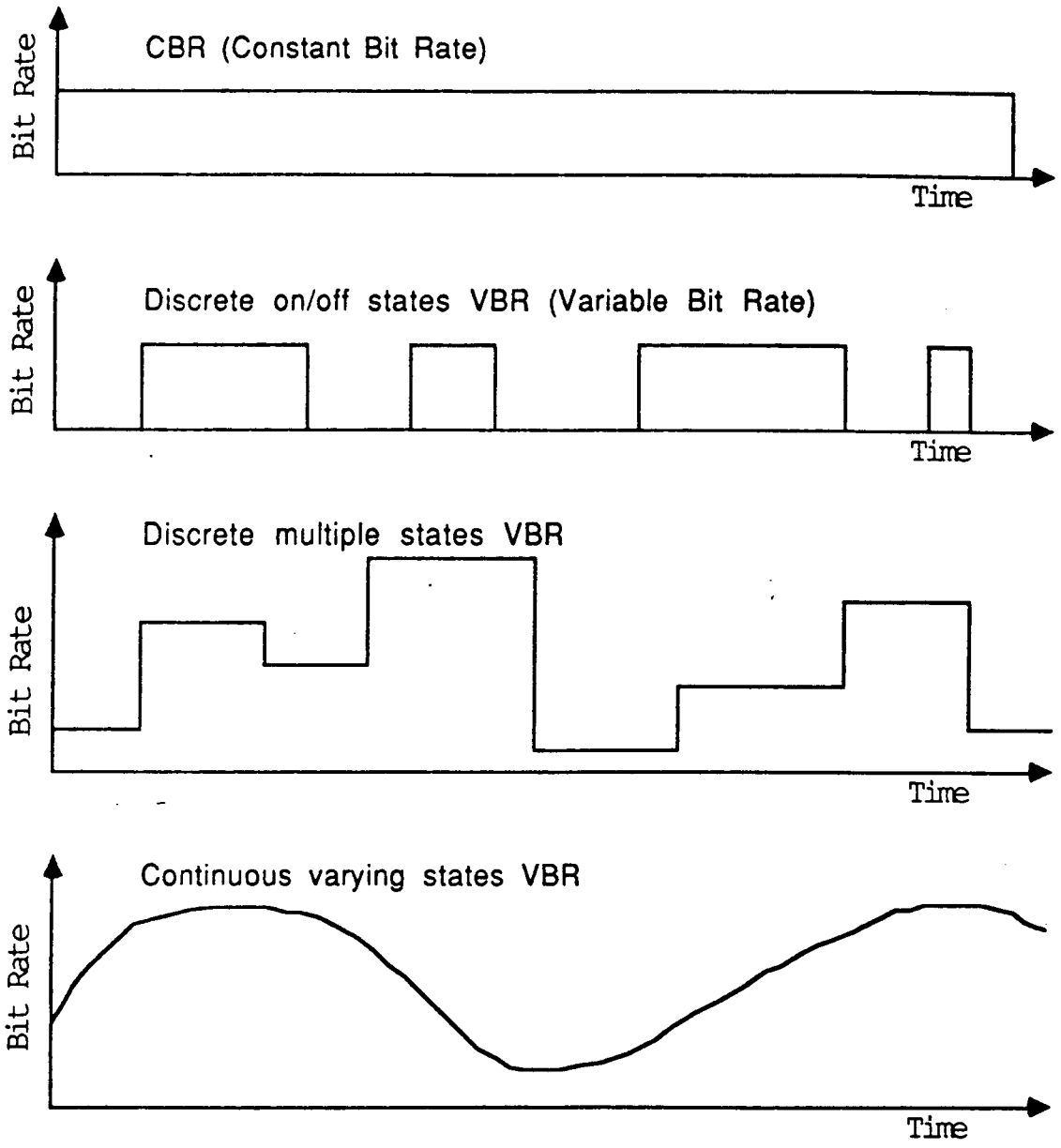
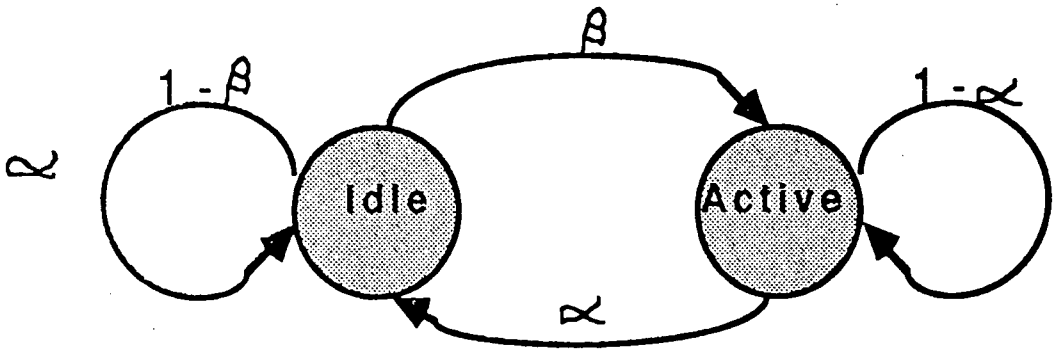


Fig. 7. Bit Rate Characterization of ATM Traffic Process

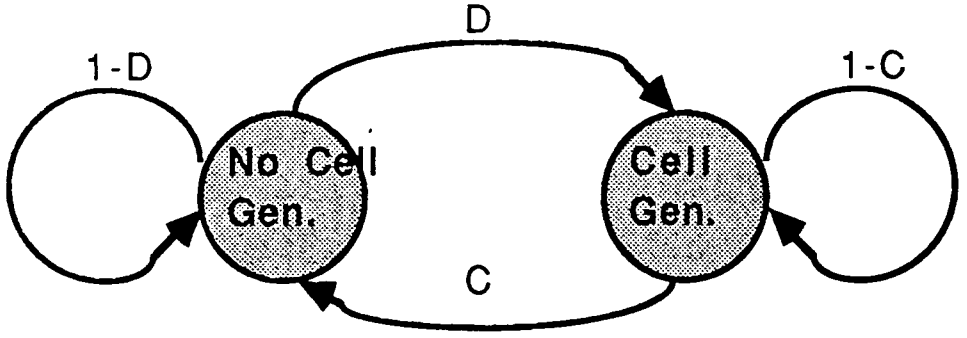
Bi-State Idle/Active Model for a Bursty Call



α : Transitional Probability from active to idle state

β : Transitional Probability from idle to active state

Bi-State Cell Generation Model for an Active Call



C : Transitional Probability from cell generation to no cell generation

D : Transitional Probability from no cell generation to cell generation

Fig. 8. Bi-state Models for a Bursty Call

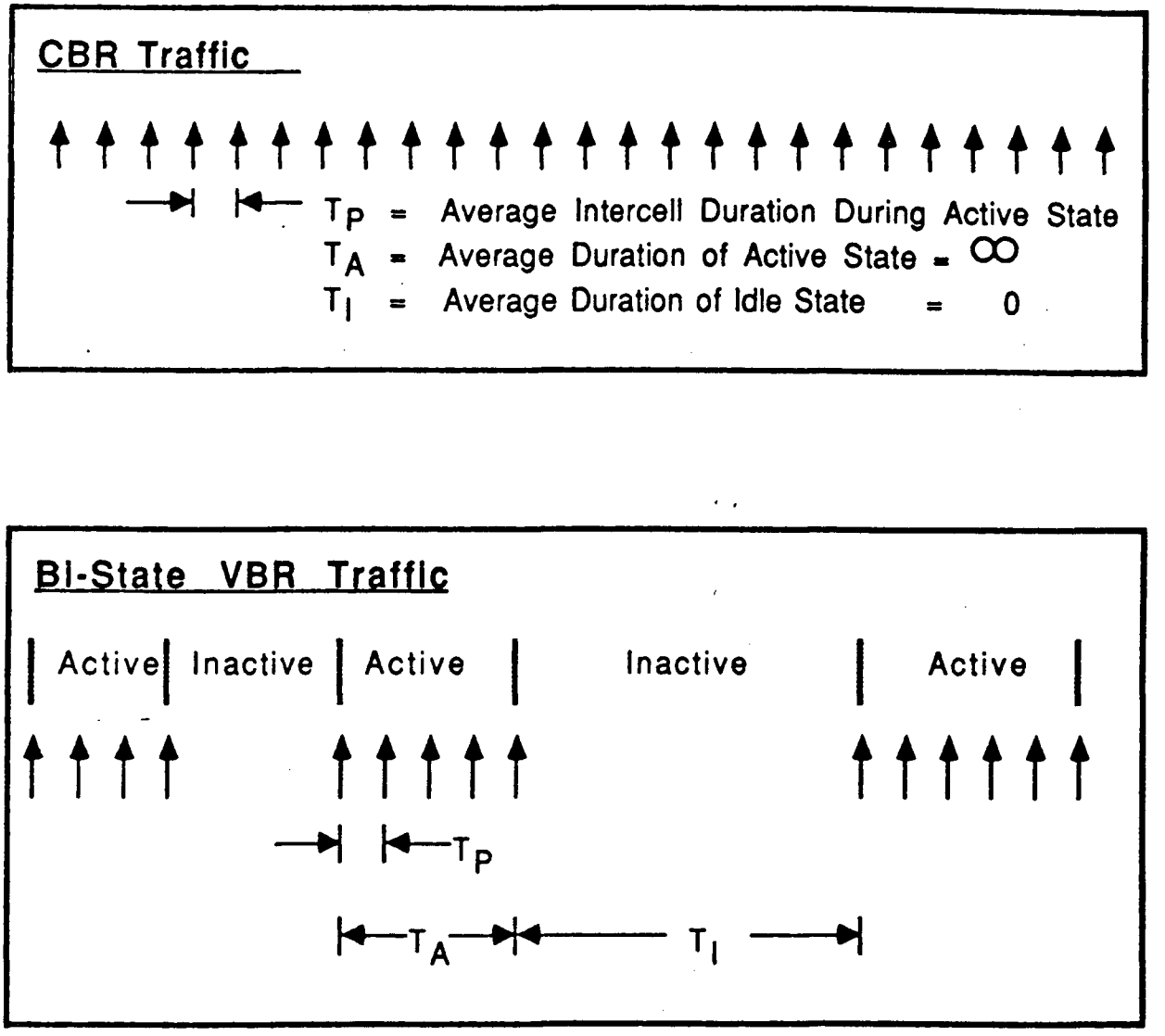


Fig.9. Long-Term-Time-Averaged Observable Traffic Parameters

Chapter 4 UNI Queue Length Process of User Traffic

Under heavy traffic condition, the multiplexing of user traffic at the UNI access-node is modelled as a fluid model with random service disruptions due to the higher priority signalling traffic. In Chapter 2, the disrupted user traffic cell-service process $S = \{S(t),\ t \geq 0\}$ has been described. In Chapter 3, the aggregate cell-arrival process of user traffic $A = \{A(t),\ t \geq 0\}$ which is modelled as a fluid process under traffic condition has been described. In this chapter, the stochastic queueing behaviour of user traffic is modelled by the random process $Q = \{Q(t),\ t \geq 0\}$; where sample path $Q(t)$ is the number of cells queueing in the user traffic buffer at time t.

In this thesis, the performance requirements for the UNI access-node are defined in terms of the stochastic cell loss requirement and the deterministic cell delay (upper-bound cell delay) requirement. The phenomenon of cell losses due to buffer overflows is modelled by the cell loss ratio process $L = \{L_k, k \geq 0\}$; where L_k is the ratio of the number of cell losses (due to buffer overflow) to the number of cell arrivals at slot k. Obviously, the cell loss ratio process (to be analyzed in Chapter 5) depends on the cell-arrival process A, the cell-service process S and the queue length process Q analyzed in this chapter.

The user traffic queue length process $Q = \{Q(t),\ t \geq 0\}$, is a function of the aggregate cell-arrival process $A = \{A(t),\ t \geq 0\}$, cell service process $S = \{S(t),\ t \geq 0\}$ and the cumulative busy time process $B = \{B(t),\ t \geq 0\}$ and the buffer capacity M. The dynamics of the process Q can be expressed as follows:

$$Q(t) = [Q(0) + A(t) - S(B(t))]^{\pm};\ \ \mathrm{x}^{\pm} = \min\{\max\{\mathrm{x}, 0\},\ \mathrm{M}\}; \qquad (20)$$

$$\text{where, } B(t) = \int_0^t 1[Q(\tau) > 0,\ I(\tau) = 1]d\tau;$$

$$\text{and } 1[Q(\tau) > 0,\ I(\tau) = 1] = \begin{cases} 1 & \text{if } Q(\tau) > 0 \text{ and } I(\tau) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Recalling from 2.2 that $I(\tau)$ is the indication function of the state of the user traffic servicing; it is in down state when $I(\tau) = 0$ and in up state when $I(\tau) = 1$.

The discrete-state/discrete-time queue length process can be described in different levels of details with regard to the *time* and *state spaces*:

A. Microscopic

a. It discloses the highest level of details. It captures the short-term process behaviour.

b. Complete queue state dependencies are taken into account.

B. Macroscopic

a. It discloses lower level of details. It captures the long-term heavy traffic (overloaded) process behaviour. Under this condition, the time and state spaces tend to be continuous.

b. Discrete queue state dependencies vanish; and the process is described by the following approximation models:

☐ Fluid Approximation

- A deterministic process is obtained at the limit.
- The first order moment is used for approximation.

☐ Diffusion Approximation

- The Markov continuous-path process is obtained as the limit of the non-Markov jump process.
- The first and second order moments are used for approximating the deviation of the system from its fluid approximation.

A fluid model [36, 37, 38] is characterized by a deterministic continuous flow of fluid. It is often used to capture the asymptotic behaviour of queueing systems in the form of a FSSLN (functional strong law of large numbers); i.e. to smooth out discrete processes. For example, an arrival process X = {X(t), t ≥ 0) can be smoothed out by averaging many independent copies of its sample path, as described in [21]. However, a cruder alternative is to rescale units of measurements and approximate X by the limit $\overline{X} = \lim_{k\to\infty} X^k = \left\{ \lim_{k\to\infty} \frac{X(kt)}{k}, t \geq 0 \right\}$; which leads to the fluid model. A fluid model with random disruptions [39] is characterized by a stochastic sequence of up and down intervals; and the flow of fluid is disrupted during the down intervals. While the fluid model is always deterministic, the diffusion model approximates the deviation of the system from its fluid model approximation.

To investigate the instantaneous cell loss behaviour at each cell-slot in the next chapter, the queue behaviour must first be examined over the short-term time range of a cell-slot interval. Under this condition, the analysis of the queue length process must take into account of the discrete-state dynamics in a time span equivalent to a cell-slot interval. However, the derivation of stationary discrete-state *p.m.f.* (probability mass function) in the discrete-state and discrete-time domains usually results in a recursive

form of solution. In general, it is easier to obtain a close form solution when the analysis is done in the continuous-state and continuous-time domains; which is applicable to queue behaviour under heavy traffic or long-term time range condition.

Therefore, the following technique is devised to obtain a close form solution of the stationary discrete-state *p.m.f.* (probability mass function) with finite buffer capacity:

- Investigate the queue behaviour under long-term heavy traffic condition with infinite buffer capacity via a fluid model with random disruptions.
    a. Under this condition, the analysis is done in the continuous-time and continuous-state domains.
    b. Derive the stationary continuous-state *c.d.f.* (cumulative distribution function) with infinite buffer capacity in close form.
- Investigate queue behaviour under short-term condition (discrete-time and discrete-state spaces) with finite buffer capacity via a M/D/1 with random disruptions:
    a. Under this condition, the analysis is done in discrete-time (in unit of cell-slot interval) and discrete-state domains.
    b. Determine the effect of buffer capacity on the discrete-state dynamics.
- Determine the limiting condition under which the fluid model with random disruptions is a reasonable approximation to the M/D/1 model subjected to the same random disruptions in describing the queue behaviour under short-term heavy traffic condition. Under this limiting condition, derive stationary discrete-state *p.m.f.* with finite buffer capacity from the continuous *c.d.f.* with infinite buffer capacity while taking into account of the effect of buffer capacity on discrete-state dynamics.

4.1 Long-Term Heavy Traffic Queue Behavior via Fluid Model with Random Disruptions

In this section, the fluid model with random disruptions is applied to capture the long-term heavy traffic behaviour of the user traffic queue length process. The model has the following characteristics:

- The aggregate cell-arrival process of user traffic is assimilated to a continuous flow of fluid.
- Time scale becomes continuous.
- Service of user traffic is randomly disrupted by the service of the higher-priority signalling traffic.
- The interarrival times and service times of signalling traffic are exponentially distributed.

In [39], the fluid model with random service disruption was applied for manufacturing systems environment; and the stationary behaviour is studied. In this thesis, this model is applied to study the long-term stationary behaviour of the user traffic queue length process under long-term heavy traffic condition.

The corresponding stationary *c.d.f.* of the user traffic queue length process under infinite buffer capacity is derived by applying the analytical approach presented in [39] as follows:

- Formulate a discrete embedded random walk process associated with the continuous queue length process. (Note: a random walk process is a regenerative renewal process with regenerative points being the renewal indices.)

- Show that the continuous queue length process is a regenerative process with regenerative points being determined by the weak descending ladder epochs or points [40, 33] of the embedded random walk process.
- Obtain the Laplace transform of the stationary *c.d.f.* of the queue length process by integrating a defined sample path functional over a regenerative interval.
- Obtain the stationary *c.d.f.* of the queue length process by taking the inverse of the above Laplace transform.

4.1.1 Embedded Random Walk Process

A discrete embedded random walk process associated with the continuous queue length process is now formulated.

Theorem 4.1.1: If time is indexed by the renewal epochs of the down/up cycle of the user traffic servicing, then the resulting embedded process with infinite buffer capacity $\hat{Q}^{\infty}$ is a regenerative random walk process with regeneration points being defined by the weak descending ladder points or epochs of the associated ladder point processes. (The ladder point processes associated with a random walk process are illustrated in Fig. 10. The sections of the random walk process between weak descending ladder points are just independent identically distributed replicates.)

Proof:

Given :

$\{\delta_i\} \equiv$ sequence of i.i.d. durations of the down / up cycles of

of the user traffic service process $= \{D_i + U_i\}$; where

D_i and U_i are defined in Section 2.2.2.

Define :

Renewal epoch process : $T = \left\{ T_j = \sum_{i=1}^{j} \delta_i,\ j \geq 0 \right\}$

$T_j \equiv$ Renewal epoch of the jth renewal (down / up cycle)

Renewal count process : $C = \{C(t) = sup\{n : T_n \leq t\},\ t \geq 0\}$

$C(t) \equiv$ Number of renewals that have occurred by time t

$R_j \equiv$ Net contribution to queue length due to cell arrival and

cell service during the *jth* down / up cycle.

$= max\{[\lambda D_j - (\mu - \lambda)U_j],\ 0\},\ k \geq 1$

$\hat{Q}^{(\infty)} \equiv$ Embedded random walk process of queue length process

with infinite buffer threshold.

$$= \left\{ \hat{Q}^{\infty}_{C(t)} = \sum_{j=1}^{C(t)} R_j, C(t) \geq 0 \right\}$$

$\because \{\delta_j = D_j + U_j\}$ are i.i.d. random variables and R_j depends on δ_j

but is independent of $\{\delta_i : i \neq j\}$;

$\therefore R_j$ are i.i.d random variables.

So, $\hat{Q}^{(\infty)}$ is a random walk process by definition.

The ladder epochs of the embedded random walk process are defined as follows:

$\{\tau_-(n), n \geq 1\} \equiv$ Sequence of weak descending ladder epochs

$$\tau_-(n+1) = inf\left\{k > \tau_-(n) :\ \hat{Q}^{\infty}_k \leq \hat{Q}^{\infty}_{\tau_-(n)}\right\}$$

By definition, $\hat{Q}^{\infty}$ has regenerative points being $\{\tau_-(n), n \geq 1\}$.

4.1.2 Regenerative Queue Length Process

Theorem 4.1.2: The queue length process with infinite buffer capacity Q^{∞} is a regenerative process with regenerative points being $\{T_{\tau_-(n)},\ n \geq 1\}$. It has a stationary distribution if and only if $E[T_{\tau_-}] < \infty$. (T_j is defined in the proof of Theorem 4.1.1 as the renewal epoch of the *j*th renewal or down/up cycle of the user traffic service process.)

Proof:

The sample path of Q^{∞} is derived from the sample path of $\hat{Q}^{\infty}$ as follows (*C(t)* is defined

in the proof of Theorem 4.1.1 as the number of renewals or down/up cycles of the user traffic process that have occurred by time t):

$$Q^{(\infty)}(t) = \begin{cases} \hat{Q}^{(\infty)}_{C(t)} + \lambda\big(t - T_{C(t)}\big), \\ \qquad \text{for } T_{C(t)} \le t < T_{C(t)} + D_{C(t)+1} \\ \Big[\hat{Q}^{(\infty)}_{C(t)} - (\mu - \lambda)\big(t - T_{C(t)} - D_{C(t)+1}\big)\Big]^{+}, \\ \qquad \text{for } T_{C(t)} + D_{C(t)+1} \le t < T_{C(t)+1} \end{cases}$$

where $x^{+} = max\{x, 0\}$,

$Q^{(\infty)} \equiv$ Queue Length Process $= \Big\{Q^{(\infty)}(t),\ t \ge 0\Big\}$,

$\hat{Q}^{(\infty)} \equiv$ Embedded Queue Length Process $= \Big\{\hat{Q}^{(\infty)}_{C(t)},\ C(t) \ge 0\Big\}$.

From Theorem 4.1.1, $\hat{Q}^{\infty}$ is regenerative with regenerative points being $\{\tau_{-}(n),\ n \ge 1\}$; therefore, Q^{∞} is regenerative with regenerative points being $\big\{T_{\tau_{-}(n)},\ n \ge 1\big\}$ and it has a stationary distribution if and only if $\mathrm{E}[T_{\tau_{-}}] < \infty$.

4.1.3 Stationary Continuous-State Distribution with Infinite Buffer Capacity

Theorem 4.1.3: The Laplace Transform of the stationary distribution of the queue length process with infinite buffer capacity can be determined by integrating the sample path functional of the queue length process; and it is given by $\frac{\mathrm{E}\left[\int_0^{T_{\tau-}} e^{-\alpha Q^{(\infty)}(t)}\,dt\right]}{\mathrm{E}[T_{\tau-}]}$. (A typical sample path of the queue length process is illustrated in Fig. 11.)

Proof:

Given : $Q^{(\infty)} = \left\{ Q^{(\infty)}(t),\ t \geq 0 \right\}$ is a regenerative queue length process with

regeneration points being $\left\{ T_{\tau-(n)},\ n \geq 1 \right\}$, and

$f()$ is a bounded measurable function.

Then : $Q^{(\infty)}(t)$ *and* $f\left(Q^{(\infty)}(t)\right)$ are sample paths, *and*

$Q^{(\infty)}(\infty)$ and $f\left(Q^{(\infty)}(\infty)\right)$ are random variables.

Time Average : $\overline{f\left(Q^{(\infty)}(\mathbf{t})\right)} = \frac{\mathrm{E}\left[\int_0^{T_{\tau-}} f\left(Q^{(\infty)}(t)\right)dt\right]}{\mathrm{E}[T_{\tau-}]}$ (renewal reward theorem [27, 28])

Mean of Limiting Distribution : $\lim_{t\to\infty}\mathrm{E}\left[f\left(Q^{(\infty)}(t)\right)\right] = \mathrm{E}\left[f\left(Q^{(\infty)}(\infty)\right)\right]$

$\because Q^{(\infty)}$ is a regenerative process with $\mathrm{E}[\mathrm{T}_{\tau\_}] < \infty$ and $\mathrm{E}\left\{ \int_0^{\mathrm{T}_{\tau\_}} \left| Q^{(\infty)}(t) \right| dt \right\} < \infty$

$\therefore$ Time Average = Mean of Limiting Distribution

i.e. $$\frac{\mathrm{E}\left[\int_0^{T_{\tau\_}} f\left(Q^{(\infty)}(t) \right) dt \right]}{\mathrm{E}[T_{\tau\_}]} = E\left[f\left(Q^{(\infty)}(\infty) \right) \right]$$

$\widetilde{F_Q^{(\infty)}}(\alpha) \equiv$ Laplace Transform of the stationary distribution of $Q^{(\infty)}(\infty)$

$$= \int_0^{\infty} e^{-\alpha q} dF_{Q^{(\infty)}(\infty)}(q)$$

$$= E\left[f\left(Q^{(\infty)}(\infty) \right) \right] \quad \text{when } f(x) = e^{-\alpha x}$$

$$= \frac{\mathrm{E}\left[\int_0^{T_{\tau\_}} e^{-\alpha Q^{(\infty)}(t)} dt \right]}{\mathrm{E}[T_{\tau\_}]}$$

Applying Theorem 4.1.3, the Laplace transform of the stationary distribution of the Q^{∞} is given by:

$$\widetilde{F_Q^{(\infty)}}(\alpha) = \frac{-\frac{1}{\alpha(\mu-\lambda)} + \frac{\mu}{\alpha\lambda(\mu-\lambda)} \mathrm{E}\left[\sum_{i=0}^{\tau_- - 1} \left[e^{\alpha \hat{Q}_i{}^{(\infty)}} - e^{-\alpha\left(\hat{Q}_i{}^{(\infty)} + \lambda D_{i+1} \right)} \right] \right]}{\mathrm{E}[T_{\tau\_}]} \tag{27}$$

where,

$\lambda =$ Mean cell arrival rate (Section 3.1.2.2)

$\mu =$ Mean cell service rate (Section 2.2.2)

$D_i = ith$ duration of down state of user traffic servicing (Section 2.2.2)

The stationary cumulative distribution function is obtained by taking the inverse Laplace transform of Eqn. 27 and it is given as follows:

$$F_Q^{\infty}(q) \equiv Prob\{Q^{\infty} \leq q\} = Inverse\ Laplace\ Transform\ of\ \widetilde{F_Q^{\infty}}(\alpha)$$
$$= \frac{1-\sigma}{1+\gamma} + \frac{\sigma}{\rho(1-\lambda)}\left[1 - \exp\left(-\frac{(1-\sigma)q}{\lambda \mathrm{E}[D]}\right) + (1-\sigma)\exp\left(-\frac{q}{\lambda \mathrm{E}[D]}\right)\right]$$
$$\text{where} \quad \rho = \frac{\lambda}{\mu}, \qquad \gamma = \frac{\mathrm{E}[D]}{\mathrm{E}[U]}, \qquad \sigma = \frac{\lambda \mathrm{E}[D]}{(\mu-\lambda)\mathrm{E}[U]},$$
$$F_Q^{\infty}(-\infty) = 0 \text{ and } F_Q^{\infty}(\infty) = 1. \tag{28}$$

4.2 Short-term Queue Behaviour

In the last section, the queue length process under long-term heavy traffic was analyzed in continuous-time and continuous-state domains via the fluid model with random disruptions, and the stationary continuous-state *c.d.f.* with infinite buffer capacity was derived in close form. In this section, the queue length process under short-term condition is examined in discrete-time and discrete-state domains.

Then the stationary discrete-state *p.m.f.* with finite buffer capacity in close form is derived as follows:

- Determine the effect of buffer capacity on the discrete-state dynamics via the M/D/1 model subjected to the same random disruptions as the fluid model.
- Determine the limiting condition under which the fluid model with random disruptions is a reasonable approximation to the M/D/1 model subjected to the same random disruptions in describing the queue behaviour under short-term heavy traffic condition.

- Under the limiting condition, the stationary discrete-state *p.m.f.* with finite buffer capacity is derived from the continuous *c.d.f.* with infinite buffer capacity while taking into account of the effect of buffer capacity on discrete-state dynamics.

4.2.1 Effect of Buffer Capacity on Discrete-State Dynamics via M/D/1 Model with Random Disruptions

In the discrete-state domain, define $P_Q^{(\infty)}(q)$ and $P_Q^{(M)}(q)$ as the *p.m.f.* with infinite buffer capacity and finite buffer capacity M respectively. In [41], it was shown that for the M/D/1 queueing model with random service disruption and with a given buffer capacity M, the ratio r_q of each non-empty state probability $\left\{P_Q^{(M)}(q),\ 1 \leq q \leq M\right\}$ to empty state probability $P_Q^{(M)}(0)$ is independent of the buffer capacity; however, empty state probability varies with the buffer capacity to cause the redistribution of the state probabilities. This relationship is shown as follows:

for $M = 1, \cdots, \infty$ *and* $q = 0, 1, \cdots, M$;

$$r_0 = 1$$

$$r_q = \frac{P_Q^{(1)}(q)}{P_Q^{(1)}(0)} = \cdots = \frac{P_Q^{(M)}(q)}{P_Q^{(M)}(0)} = \cdots = \frac{P_Q^{(\infty)}(q)}{P_Q^{(\infty)}(0)} \tag{29}$$

i.e. r_q is independent of M.

4.2.2 Comparisons of Fluid Model and M/D/1 Model Subjected to the Same Random Disruptions

Consider the fluid (D/D/1) model and the M/D/1 model with the common model parameters as follows:

- Mean arrival rate λ.

- Mean service rate μ
- Subjected to random disruptions characterized by exponential up and down states with means $\overline{U}$ and $\overline{\mathrm{D}}$ respectively. ($\rho = \frac{\lambda}{\mu}$, $\gamma = \frac{\overline{D}}{\overline{U}}$, $\sigma = \frac{\lambda\overline{D}}{(\mu-\lambda)\overline{U}}$)

For the fluid model with random disruptions, the stationary average queue length is:

$$\mathrm{E}\left[Q^{(\infty)}(\infty)\right] = \frac{(2-\sigma)\sigma^2\mu\overline{D}}{(1-\sigma)(1+\gamma)} \tag{30}$$

The M/D/1 model with random disruptions is equivalent to the M/G/1 model without random disruptions. For the M/G/1 model, the mean and variance of the service times are

$$\frac{1}{\mu}\left(1+\frac{\overline{D}}{\overline{U}}\right) \quad \text{and} \quad \frac{2\overline{D}^2}{\mu\overline{U}} + \frac{1}{\mu^2}\left(1+\frac{\overline{D}}{\overline{U}}\right)^2 \tag{31}$$

Applying the Pollaczek-Khinchin formula, the stationary average queue length is :

$$\mathrm{E}\left[\breve{Q}^{(\infty)}(\infty)\right] = \frac{\sigma}{1-\sigma}\left(1+\frac{\overline{U}}{\overline{D}}+\lambda\overline{D}\right) \tag{32}$$

Comparing Eqn. 30 and Eqn. 32:

$$\frac{\mathrm{E}\left[Q^{(\infty)}(\infty)\right]}{\mathrm{E}\left[\breve{Q}^{(\infty)}(\infty)\right]} = \frac{\sigma(2-\sigma)}{\frac{1}{\mu\overline{U}}\left(1+\frac{1}{\gamma}\right)^2 + \rho(1+\gamma)} \tag{33}$$

Therefore, the limiting condition for $\mathrm{E}\left[Q^{(\infty)}(\infty)\right] \rightarrow \mathrm{E}\left[\breve{Q}^{(\infty)}(\infty)\right]$ is: $\overline{U} >> \frac{1}{\mu}$ and $\sigma \rightarrow 1$. It is reasonable to assume that the average duration of the up state of user traffic servicing is much larger then the average service time per cell, i.e. $\overline{U} >> \frac{1}{\mu}$. When $\sigma \rightarrow 1$, heavy traffic condition prevails.

4.2.3 Stationary Discrete-State Distribution with Finite Buffer

Providing the limiting condition ($\overline{U} >> \frac{1}{\mu}$ and $\sigma \rightarrow 1$) is met, the stationary discrete-state *p.m.f.* with finite buffer capacity under short-term condition can be approximated as follows:

- Stationary discrete-state *p.m.f.* with infinite buffer capacity is mapped from the stationary continuous-state *c.d.f.* with infinite buffer capacity.
- Stationary discrete-state *p.m.f.* with finite buffer capacity is determined by the relation between buffer capacity and *p.m.f.*

The discrete-state *p.m.f.*s can be mapped from the corresponding continuous-state *c.d.f.*s — $F_Q^{(\infty)}(q)$ and $F_Q^{(M)}(q)$ — as follows:

$$P_Q^{(\infty)}(q) \equiv Prob\left\{q-1 \leq Q^{(\infty)} \leq q\right\} \equiv F_Q^{(\infty)}(q) - F_Q^{(\infty)}(q-1) \qquad (34)$$
$$P_Q^{(M)}(q) \equiv Prob\left\{q-1 \leq Q^{(M)} \leq q\right\} \equiv F_Q^{(M)}(q) - F_Q^{(M)}(q-1)$$

In Section 4.2.1, the relation between buffer capacity and *p.m.f.* is as follows:

for $M = 1, \cdots, \infty$ *and* $q = 0, 1, \cdots, M$;

$$r_0 = 1$$
$$r_q = \frac{P_Q^{(1)}(q)}{P_Q^{(1)}(0)} = \cdots = \frac{P_Q^{(M)}(q)}{P_Q^{(M)}(0)} = \cdots = \frac{P_Q^{(\infty)}(q)}{P_Q^{(\infty)}(0)} \qquad (35)$$

ie, r_q is independent of M.

Therefore,

$$P_Q^{(M)}(q) = r_q P_Q^{(M)}(0) \qquad (36)$$

Substituting Eqn. 34 into Eqn. 35, we have:

$$r_q = \frac{P_Q^{(M)}(q)}{P_Q^{(M)}(0)} = \frac{P_Q^{(\infty)}(q)}{P_Q^{(\infty)}(0)} = \frac{F_Q^{(\infty)}(q) - F_Q^{(\infty)}(q-1)}{F_Q^{(\infty)}(0)} \tag{37}$$

Since $\sum_{q=0}^{M} P_Q^{(M)}(q) = 1$; and $P_Q^{(M)}(q) = r_q P_Q^{(M)}(0)$

$$\begin{aligned} \therefore P_Q^{(M)}(0) &= \frac{1}{\sum_{q=0}^{M} r_q} \\ &= \left[1 + \frac{F_Q^{(\infty)}(1) - F_Q^{(\infty)}(0)}{F_Q^{(\infty)}(0)} + \cdots\cdots \frac{F_Q^{(\infty)}(M) - F_Q^{(\infty)}(M-1)}{F_Q^{(\infty)}(0)}\right]^{-1} \\ &= \frac{F_Q^{(\infty)}(0)}{F_Q^{(\infty)}(M)} \end{aligned} \tag{38}$$

Substituting Eqn. 37 and Eqn. 38 into Eqn. 36, the stationary *p.m.f.* with finite buffer capacity M is:

$$P_Q^{(M)}(q) = r_q P_Q^{(M)}(0) = \frac{F_Q^{(\infty)}(q) - F_Q^{(\infty)}(q-1)}{F_Q^{(\infty)}(M)}, \quad q = 0, 1, \cdots, M. \tag{39}$$

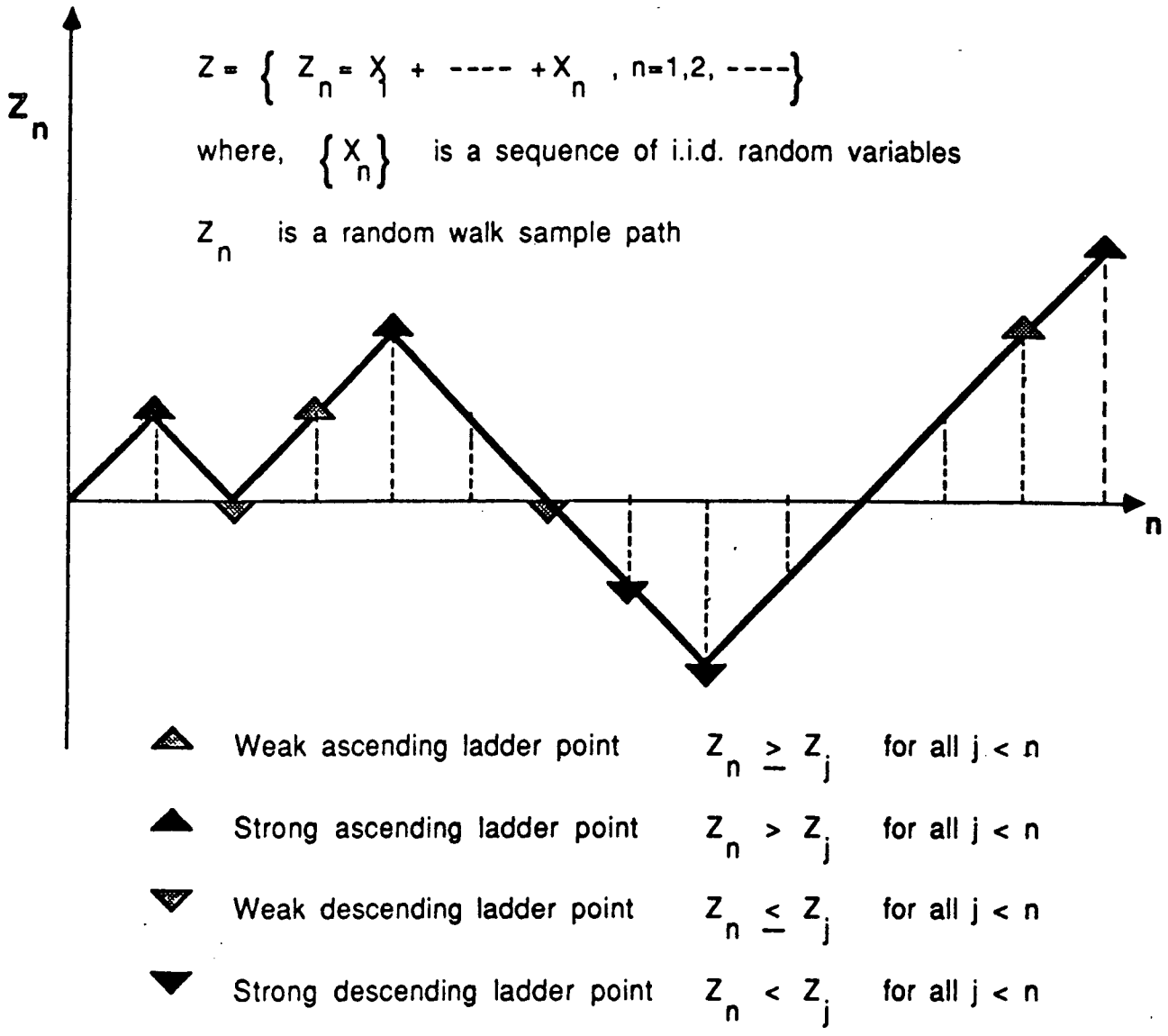
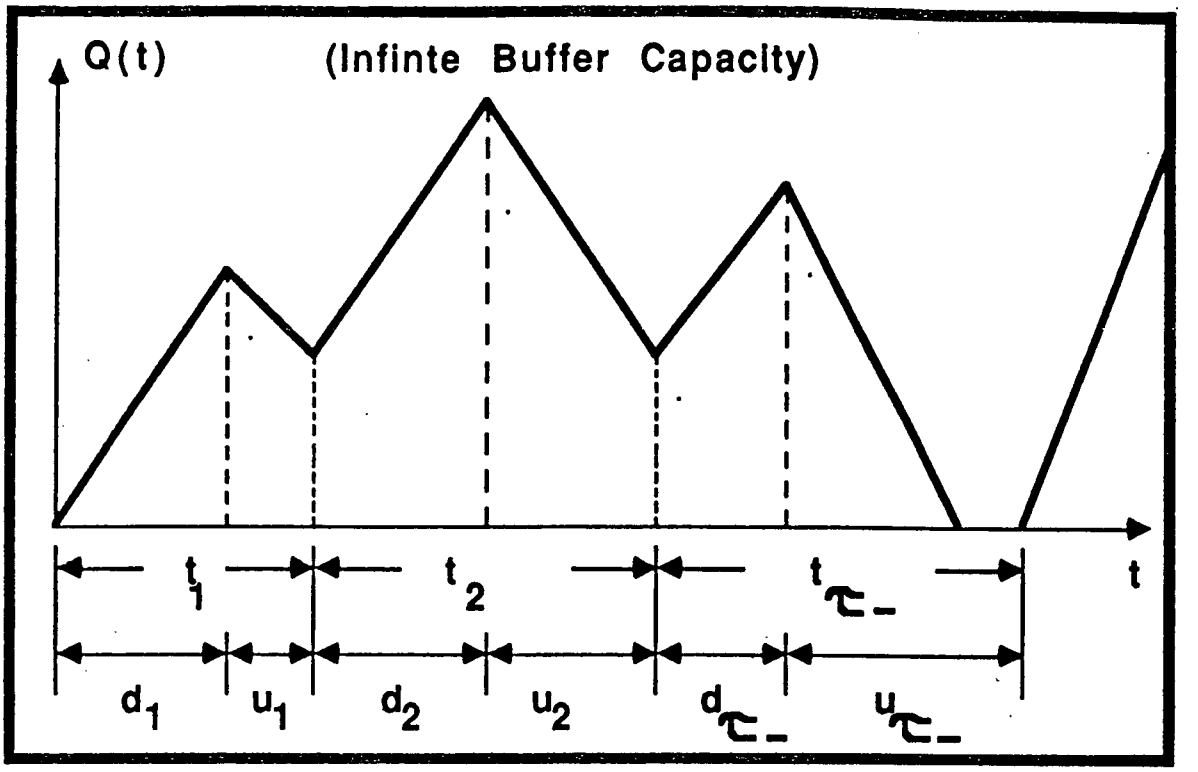


Fig. 10. Ladder Points of a Random Walk Sample Path



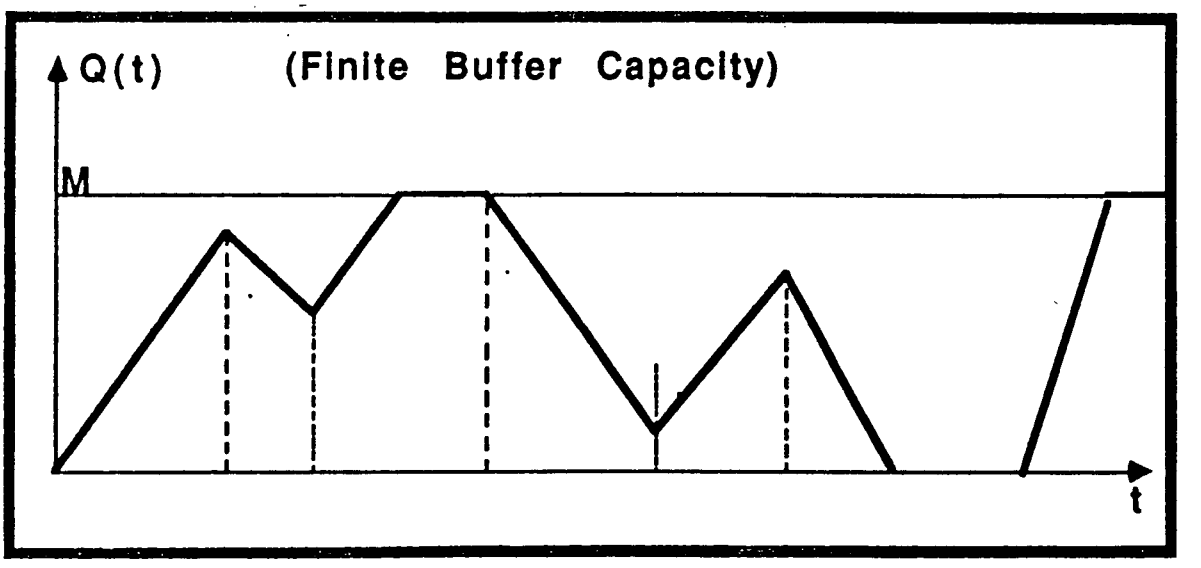


Fig. 11. A Typical Sample Path of the Queue Length Process

Chapter 5 UNI Cell Loss Ratio Process of user Traffic

This chapter investigates the cell loss phenomenon due to user traffic buffer overflow during user traffic multiplexing at the UNI, which is modelled as a fluid model with random service disruptions due to the higher-priority signalling traffic. In this thesis, the cell loss phenomenon is modelled as a discrete-state/discrete-time stochastic cell loss ratio process, $L = \{L_i,\ i \geq 0\}$, where $L_{i=k}$ is a random variable denoting the ratio of the number of cell losses (due to buffer overflow) to the number of cell arrivals at slot k and $L_i, i \geq 0$ is a sample path of the random process.

The cell loss ratio process is a function of the aggregate cell arrival process $A = \{A_i,\ i \geq 0\}$, cell service process $S = \{S_i,\ i \geq 0\}$, the queue length process $Q = \{Q_i,\ i \geq 0\}$ and the buffer capacity M. The dynamics of the process can be expressed as follows:

$$L_i = \frac{max\{[Q_{i-1} + A_i - S_i - M], 0\}}{A_i} \tag{40}$$

The cell loss ratio process is modelled as a discrete doubly stochastic process where the state of the cell loss ratio process is determined by the imbedded active-call process $V^{(N)} = \left\{V_i^{(N)},\ i \geq 0\right\}$ (Section 3.1.1); where $V_{i=k}^{N}$ is the random variable denoting the number of calls being in the active state out of N calls at slot k. Recall that a call is either idle or active, and the transitions between idle and active states form a bi-state Markov chain as illustrated in Fig. 8.

We will derive the stationary tail distribution of the cell loss ratio process by the following procedures:

1. To derive the instantaneous cell loss ratio at a slot by mapping between states of the active calls process and the cell loss ratio process.
2. To establish the necessary condition for the existence of a stationary cell loss ratio process.
3. To derive the stationary tail distribution of the cell loss ratio process in terms of the stationary distribution of the active call process.

5.1 Derivation of Instantaneous Cell Loss Ratio

The instantaneous cell loss ratio at slot k is derived by employing the technique in reference [18]. Given the following system parameters:

- M = Upper-bound queue size allowed for buffer.
- N = Number of homogeneous user traffic calls.

Then, in slot k:

- the stationary queue length is q with probability $P_{Q(M)}(q)$,
- the number of active calls out of N calls is v with probability $P^{k}_{V(N)}(v)$,
- the number of cells generated is g with conditional probability $P^{k}_{G|V(N)}(g)$, and
- the stationary number of cells serviced is s with probability $P_S(s)$.

The probabilities $P_{Q(M)}(q)$, $P^{k}_{VN}(v)$, $P^{k}_{G|V(N)}(g)$ and $P_S(s)$ have been obtained in Sections 4.2, 3.1.2.1, 3.1.2.2 and 2.2.2 respectively.

The number of lost cells due to buffer overflow is then $max[(q+g-s-M),0]$. The normalized instantaneous cell loss ratio at slot k, i.e. the ratio of the number of cell losses to the number of cell arrivals at slot k, (L_k) as a function of the number of active calls v becomes:

$$L_k = \frac{\sum_{q=0}^{M} P_{Q^{(M)}}(q) \sum_{g=0}^{v} max\left[\sum_{s=0}^{1} (q+g-s-M) P_S(s), 0\right] P^k_{G|V^{(N)}}(g)}{\sum_{g=0}^{v} g P^k_{G|V^{(N)}}(g)} \tag{41}$$

5.2 Stationary or Steady-State Cell Loss Ratio

Because of the dynamic nature, an exact analysis of the cell loss ratio process would require the complete knowledge about the history of the changes in the number of active calls, and it is not tractable.

However, if changes in the number of active calls happen at a moderate rate, and the time required for a network to reach the steady state after a change is short compared to the time to the next change in the number of active calls, then it would be justified to assume the cell loss ratio process as being stationary.

An intuitive argument for the above assumption can be illustrated by the example of voice conversions, where the average talkspurt length of 0.185 second and the cell slot length of 3.7 μ seconds are assumed, values typical for ATM networks. In this example, the average talkspurt length is 50,000 slots; i.e. once a call becomes active, it remains active for an average of 50,000 slots. Statistical fluctuations in the number of active calls occur very slowly in the time scale of slots. It is, therefore, reasonable to assume that the instantaneous cell loss ratio converges to the steady-state value long before the next change in the number of active calls occurs.

Then the instantaneous cell loss ratio converges to its stationary or steady-state value as k approaches infinity. The stationary cell loss ratio as a function of v is given as follows:

$$L = \lim_{k \to \infty} L_k \tag{42}$$
$$= \frac{\sum_{q=0}^{M} P_{Q(M)}(q) \sum_{g=0}^{v} max\left[\sum_{s=0}^{1} (q+g-s-M) P_S(s), 0\right] P_{G|V(N)}(g)}{\sum_{g=0}^{v} g P_{G|V(N)}(g)}$$

where,

$$P_{G|V(N)}(g) \equiv \lim_{k \to \infty} P^k_{G|V(N)}(g)$$

5.3 Distribution of the Stationary Cell Loss Ratio

The tail *c.d.f.* (cumulative distribution function), $\overline{F_L}(l)$, of the stationary cell loss ratio is defined as follows: $\overline{F_L}(l) = Prob.\{L > l\} = 1 - F_L(l) = 1 - Prob.\{L \leq l\}$ Recalling from Section 5.2 that $L = fcn(V)$ or $V = fcn^{-1}(L)$, and from Section 3.1.2 that $P_{V(N)}(v)$ is the conditional probability that v calls are active at a slot out of N calls.

Let the stationary cell loss ratio L assumes a value of l; then the probability that L remains greater than l is the probability that the number of active calls remains greater than or equal to $v = fcn^{-1}(l)$. Therefore,

$$\overline{F_L}(l) = Prob.\{L > l\} = \sum_{v=fcn^{-1}(l)}^{N} P_{V(N)}(v) \tag{43}$$

Chapter 6 UNI Congestion Management Model

The traffic process associated with each of the protocol layers can be characterized by appropriate traffic variables. Since teleservice layer combines the information processing function with the information transfer function provided by the bearer service layer, the traffic processes associated with these two layers can be described by the same set of traffic variables. The traffic variables associated with teleservice or bearer service layer can be classified according to the following phases of the user traffic:

- Call access and disengagement phases (e.g. average rate of call attempts).
- Information transfer phase
    - Cell Transfer (e.g. mean call connection time).
    - Cell Generation (e.g. mean offered traffic intensity in Erlang).

As for the ATM transfer mode layer, the associated traffic process is characterized by the information transfer phase variables; i.e. it is a subset of the associated traffic process of the teleservice or bearer service layer.

In this thesis, an integrated congestion management platform is employed at the UNI. Integrated congestion management dictates that congestion control schemes are applied during the call access phase and the information transfer phase of user traffic source.

- Call access phase control
    a. Call-level access control scheme via call admission control.
- Information transfer phase control
    a. Cell-level transfer control scheme via intra-node buffer control.

b. Cell-level generation control scheme via source policing control.

The congestion management architecture of the UNI access-node as illustrated in Fig. 12 consists of the following components: (1) ATM asynchronous statistical multiplexer; (2) call-level congestion controller to implement call-level access control scheme; (3) cell-level congestion controller to implement cell-level transfer control scheme; (4) user traffic source enforcer to implement cell-level generation control scheme.

The servicing of user traffic through the ATM asynchronous statistical multiplexer is modelled as a fluid model with random disruptions due to the servicing of signalling traffic and it is associated with the following system parameter:

- Aggregate Cell Arrival Process A (analyzed in Section 3.1.2)
  - $\overline{T_A}$ = Average duration of call active state (traffic dependent).
  - $\overline{T_I}$ = Average duration of call idle state (traffic dependent).
  - $\overline{R_C}$ = Average cell arrival rate during call active state (traffic dependent).
  - μ = Maximum cell service rate (system configuration dependent).
  - N = Number of homogeneous user traffic calls (controllable).
- Cell Service Process S (analyzed in Section 2.2)
  - $\overline{D}$ = Average duration of the *down* state of user traffic servicing. (dependent on signalling traffic statistics)
  - $\overline{U}$ = Average duration of the *up* state of user traffic servicing (dependent on signalling traffic statistics).

- Queue Length Process Q (analyzed in Section 4 as a function of Aggregate Cell Arrival Process and Cell Service Process)
  - B = Buffer capacity (system configuration dependent).
  - M = Upper-bound queue size allowed for buffering; $M \leq B$ (controllable).
- Performance Related Processes
  - Cell Loss Ratio Process L (analyzed in Chapter 5 as a function of Aggregate Cell Arrival Process, Cell Service Process and Queue Length Process)

    $L(N,M)$ = Stationary cell loss ratio random variable (Section 5.2) as a function of N and M. The derivation of $L(N,M)$ from the system parameters including N and M is illustrated in Fig. 13.
    - Tail *c.d.f.* (cumulative distribution function) of L: $\overline{F}_L(l) = P\{L \geq l\}$ (Eqn. 43)
    - $L(N,M)$ increases stochastically[1] as N increases for a given M, as illustrated in Fig. 14 (e.g. $L(N_1, M) \overset{st}{\leq} L(N_2, M)$); and decreases stochastically as M increases for a given N as illustrated in Fig. 15.
  - Cell Delay Process

    $D(M)$ = Upper-bound cell delay as a function of M.

    It is directly proportional to M; i.e. $D = kM$ where k is the cell transmission

---

[1] To understand the meaning of "stochastically smaller", it is necessary to distinguish stochastic ordering [33, 34] from deterministic ordering. Let X and Y be two random variables. X is deterministically larger than Y, written as $X \geq Y$ (deterministic ordering) if

$$X(\omega) \geq Y(\omega) \quad \text{for every point } \omega \text{ in the sample space } \Omega.$$

X is stochastically larger than Y, written as $X \overset{st}{\geq} Y$ (stochastic ordering) if

$$P(X > t) \geq P(Y > t) \quad \text{for all } t.$$

time. For a link transmission rate of 150 Mbps and cell size of 53 bytes, k = 2.867 μs/cell.

- Performance Requirements
  - Let L_{thd} = Stationary cell loss ratio threshold random variable. Due to system dynamics considerations, it is required that $L(M,N) \overset{st}{\leq} L_{thd}$, i.e. L must be stochastically smaller than L_{thd} for all feasible values of *(M, N)*.
  - Let D_{thd} = Upper-bound cell delay threshold. $D(M) \leq D_{thd}$ must be satisfied for all feasible values of M. Consequently, $D(M) \leq D_{thd}$ for $M \leq M_{thd}$, where M_{thd} = Upper-bound queue size threshold above which the upper-bound cell delay threshold will be exceeded.

Cell-level generation control schemes implemented by the user traffic source enforcer [19, 20, 42, 43] ensure that the user sources generate traffic according to the values declared in the call setup phase. The enforcement control acts on each source before traffic from all sources is multiplexed. Some algorithms for triggering the enforcement control include the leaky bucket mechanism [44] and the virtual leaky bucket mechanism [20]. Enforcement controls can be of cell discard or cell tagging schemes.

In this thesis, the cell-level generation control scheme will not be discussed, i.e. traffic sources are assumed to be well-behaved with respect to declared traffic values. Instead, call access and cell-level transfer control schemes are proposed and analyzed.

6.1 Call-Level Access Control Scheme

The objective of the call-level access control scheme is to maintain the required network performance assigned to the UNI access-node by exerting call admission control

in the call access phase of each traffic source.

The network performance parameter chosen for call level access congestion control is the cell loss ratio. The cell loss ratio process as a stochastic process has been defined and analyzed in Chapter 5. The formulation of this call access control scheme is as follows:

Given the cell service process, buffer capacity, the current number of call connections and the corresponding aggregate traffic process, the call-level access control scheme ensures that admitting an additional incoming call will not cause the stationary cell loss ratio to stochastically exceed the cell loss ratio threshold (derived from the network-oriented QOS), subject to the constraint that the upper-bound cell delay is less than or equal to the upper-bound cell delay threshold.

6.1.1 Control Model

An input-limit static control model with the following parameters is employed to model the call-level access control system:

- System Parameters (fixed): $\overline{T_A}$, $\overline{T_I}$, $\overline{R_C}$, μ, B, M, $\overline{D}$ and $\overline{U}$.
- Input Parameter (controllable): N
- Output Parameters (to be monitored): $L(N, M = M_{thd})$

  Note that M is set to M_{thd} to satisfy the constraint that the upper-bound cell delay does not exceed the upper-bound cell delay threshold, i.e. $D(M_{thd}) = D_{thd}$.

6.1.2 Control Scheme

The call-level access control scheme is described as follows:

- Control model: Input-limiting static control model.

- Control epoch: Control applied during the access phase of a user source.
- Input: N; output: $L(M,N)$.
- Control action: Admit incoming call (increase N by 1) providing $L(N+1, M=M_{thd}) \overset{st}{\leq} L_{thd}$, where $D(M_{thd}) = D_{thd}$. Otherwise, reject incoming call.

6.2 Cell-Level Transfer Control Scheme

The objective of the cell-level transfer control scheme is to optimize the network performance of the UNI access-node by exerting intra-node buffer control in the aggregate information transfer phase of the user traffic sources.

The network performance parameters chosen for cell-level congestion control is the cell loss ratio for loss sensitive teleservice, or the upper-bound cell delay for delay-sensitive teleservice. The cell loss ratio process as a stochastic process has been defined and analyzed in Chapter 5. The upper-bound cell delay is the maximum delay that can be experienced by a cell transiting the UNI access-node. The formulation of the cell-level control scheme is as follows:

Given the traffic type, the cell service process and the number of call connections, the information transfer phase control scheme rejects incoming cells from the aggregate traffic stream if the current queue size q has reached the upper-bound queue size M, as determined below: (1) for loss sensitive teleservice — M is maximized so as to minimize the stationary cell loss ratio; (2) for delay sensitive teleservice — M is minimized so as to minimize the upper-bound cell delay; subjecting to the constraints (during the data transfer phase of call handling) that the stationary cell loss ratio being stochastically

smaller than the cell loss ratio threshold and M is less than the upper-bound queue size threshold M_{thd}.

6.2.1 Control Model

This optimization problem can be considered as a sequential decision problem faced by the cell level controller as a decision maker. A sequential decision process is a model for a dynamic system under the control of a decision maker. At each point in time at which a decision is to be made, the decision maker observes the state of the system. Based on the information from this observation, an action is chosen from a set of available alternatives. The consequences of this action are twofold: the decision maker receives an immediate reward or incurs an immediate cost, and the state that the system will occupy at subsequent decision epochs is influenced either deterministically or probabilistically. The problem faced by the decision maker is to choose a sequence of actions that will optimize the performance (minimize the cost or maximize the reward) of the system over the decision-making horizon (the number of decision-making stages, each stage being a point in time at which a decision is made).

The characteristics of the sequential decision process of the cell-level controller are identified as follows:

1. Decision Epochs & Decision Making Horizon

   The control decision epochs are discrete intervals equal to an integral number of cell slots; and the decision making horizon is infinite.

2. State Space

   The observable state Q_k is the queue length at the k_{th} cell arrival. The mathematical

property of the discrete-time queue length process, $Q = \{Q_k : k \geq 0\}$, has been defined previously in Section 4.2.

3. Control Rule Space

A control rule selects an action based on the system information. A control rule is characterized as follows:

a. Markovian vs. History Dependent

A control rule is Markovian if it depends only on the current state and stage of the system and not on its past. A control rule is history dependent if it depends on the entire past history of the system as summarized in the sequence of previous states and control actions.

b. Deterministic vs. Randomized

A control rule is deterministic if it selects an action with certainty. A control rule is randomized if it specifies a probability distribution on the set of allowable actions.

In Section 4.2.3, the resulting queue length process is found to be Markovian in nature under short-term heavy traffic condition. For Markovian queueing system, it is shown in [45, 46] that the Markovian and deterministic control rule is optimal, which is therefore employed by the cell-level transfer controller. The control rule is a function $\psi_k : x_k \rightarrow \phi_k$ that specifies the control action ϕ_k when the system is in state x_k. It is defined as follows:

$$\phi_k = \begin{cases} \psi_k(x_k) = 0, & \text{cell rejection} \\ \psi_k(x_k) = 1, & \text{cell acceptance} \end{cases} \tag{46}$$

$$\phi_k \in \Phi_k, \text{ where } \Phi_k \equiv \text{control variable space}$$

$$\psi_k \in \Psi_k, \text{ where } \Psi_k \equiv \text{control rule space}$$

4. Optimality Criterion

Each cell rejection is asserted with a cost c since rejected cells are lost. The decision objective is to minimize the total expected discounted cost: $E\left\{\int_0^\infty e^{-\alpha t}Q_t dt + \sum_{n=1}^{\infty} ce^{-\alpha J_n}\right\}$ where the J_n are the rejection times, Q_t is the queue length at time t and α is the discount factor. Thus, the optimal policy has to compromise between congestion (as measured by the queue length) and the cost of lost cells due to rejection.

5. Control Policy Space

A control policy π specifies the sequence of control rules to be used by the decision maker over the course of the decision-making horizon. A control policy space Π specifies all the possible control policies; that is,

$$\pi = \{\psi_k,\ k \geq 0\} \tag{47}$$

$$where,\ \psi_k \in \Psi_k,$$

$$\Pi = \{\Psi_k,\ k \geq 0\}$$

$$where,\ \pi \in \Pi$$

In this thesis, a stationary control policy is employed by the cell-level transfer controller which uses the identical control rule in each control stage; i.e. $\pi = \{\psi, \psi, \psi, \cdots\}$.

When the decision making horizon is infinite, stationary condition frequently results

[47]. This means that the set of states, the set of allowable actions in each sate, the rewards, the probability transition functions and the decision sets are the same at every stage. Under most optimality criteria, stationary policies are optimal under stationary condition.

The problem is now to find a stationary, Markovian and deterministic control policy for the cell level controller that minimizes the total discounted cost functional. Under this problem formulation, it can be proven that the optimal control policy is a threshold policy; i.e. to reject cells which arrive when the queue length exceeds some threshold value. The proof for a similar problem formulation can be found in [45].

6.2.2 Control Scheme

The cell-level transfer control scheme is described as follows:

- Control model: Sequential decision control model analyzed by dynamic programming.
- Control epoch: Control applied during the information transfer phases of all user sources.
- State Space: Queue length at control instant.
- Control Rule Space: Cell acceptance or cell rejection at control instant.
- Control Policy: Threshold policy, i.e. reject incoming cells if the current queue size reaches the maximum queue size M which is determined as follows:

  a. Loss sensitive teleservice

  The cell loss ratio L is minimized by maximizing M while ensuring that $L(M, N) \overset{st}{\leq} L_{thd}$ and $D(M) \leq D_{thd}$. These conditions are always satisfied by setting $M = M_{thd}$.

b. Delay sensitive teleservice

The upper-bound cell delay D is minimized by minimizing $M = M_{thd} - \Delta M$ while ensuring that $L(M = M_{thd} - \Delta M, N) \overset{st}{\leq} L_{thd}$.

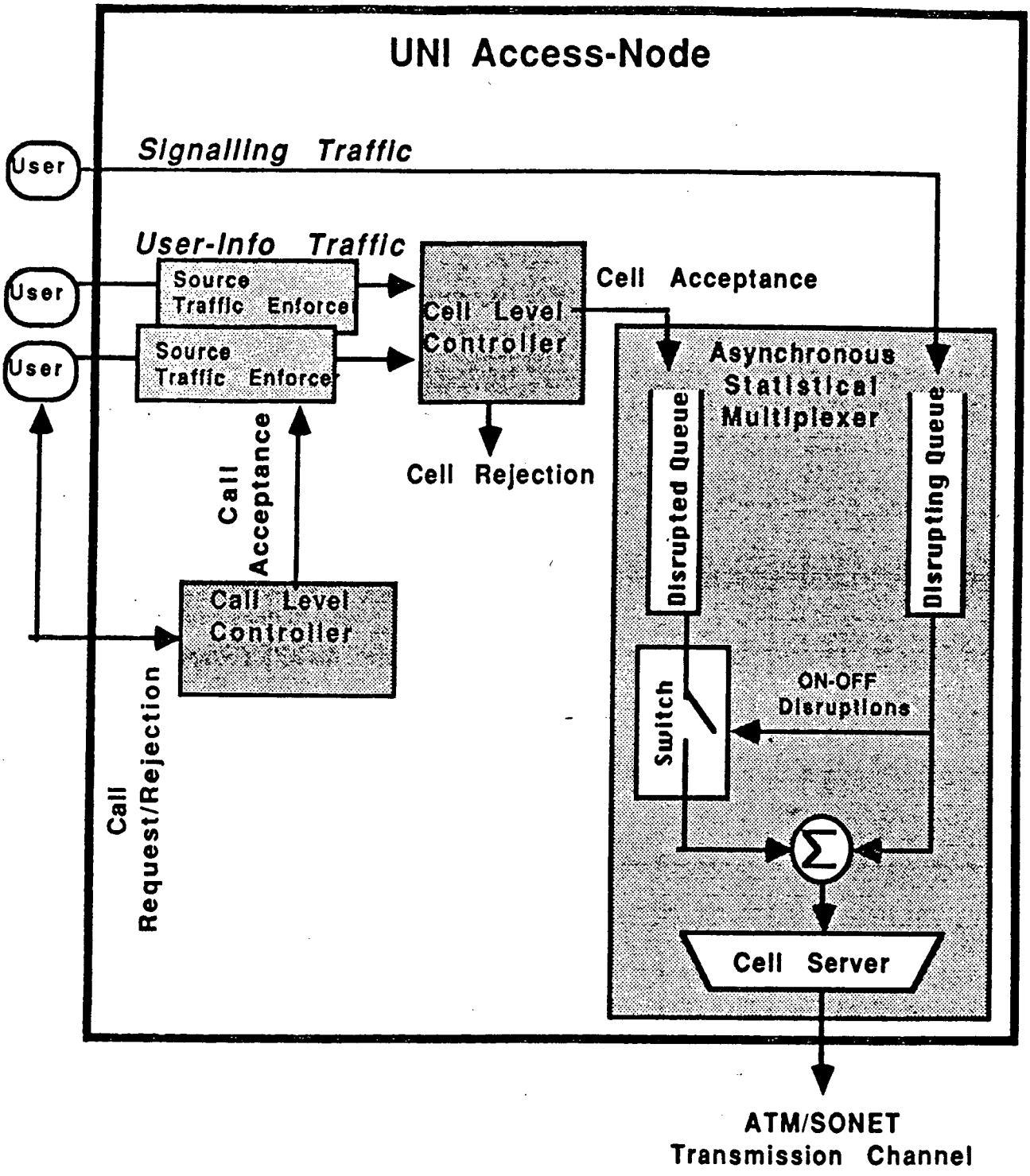
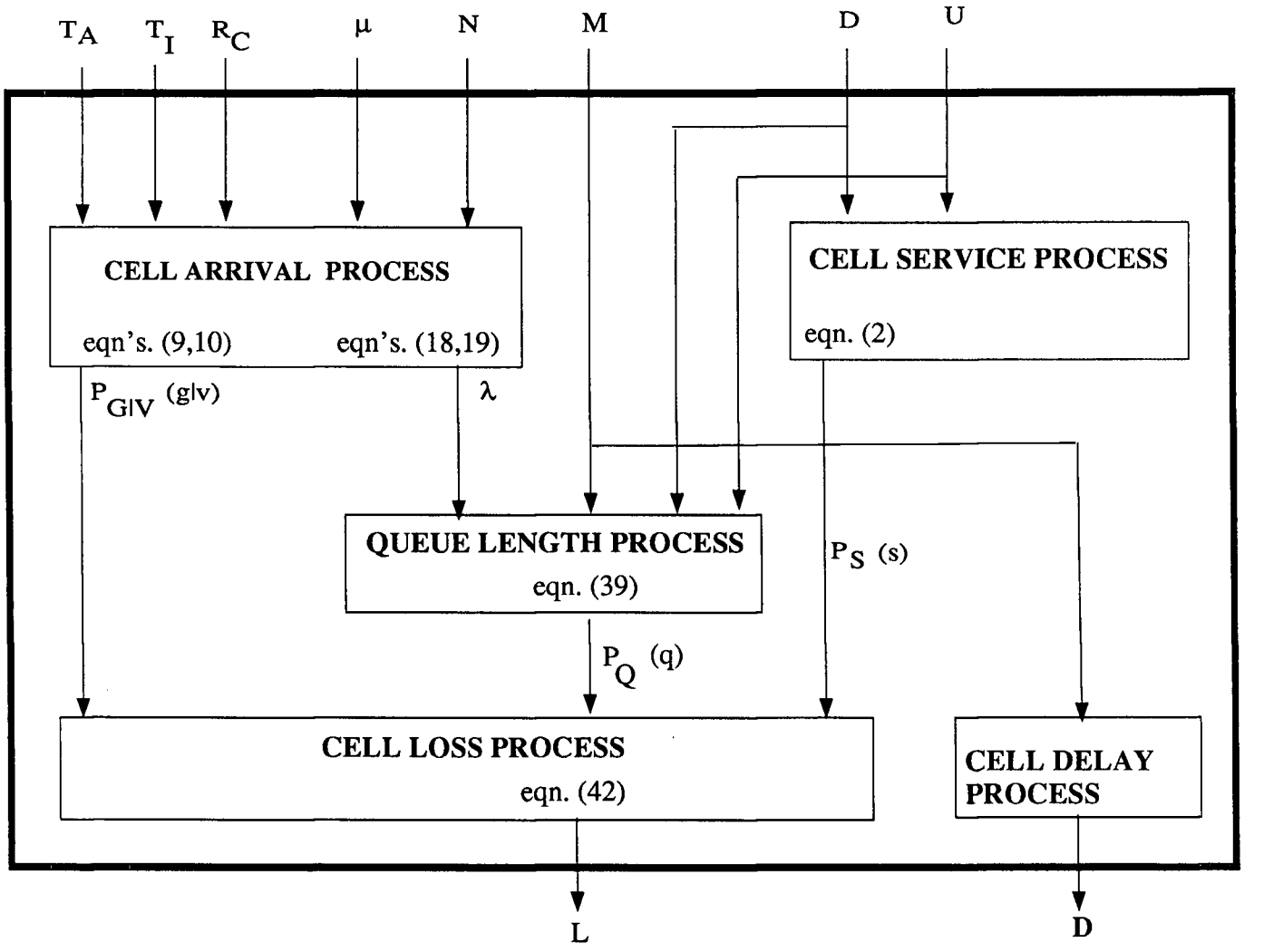


Fig. 12. Congestion Management Architecture of the UNI Access-Node



$P_{G|V}$ (g|v) : Conditional probability that g cells are generated at a slot from v active calls out of N calls.

P_Q (q) : Probability that q cells are queued in the buffer.

P_S (s) : Probability that s cells are serviced at a slot.

λ : Mean cell arrival rate.

Figure 13 System Parameters Associated with ATM Asynchronous Statistical Multiplexer

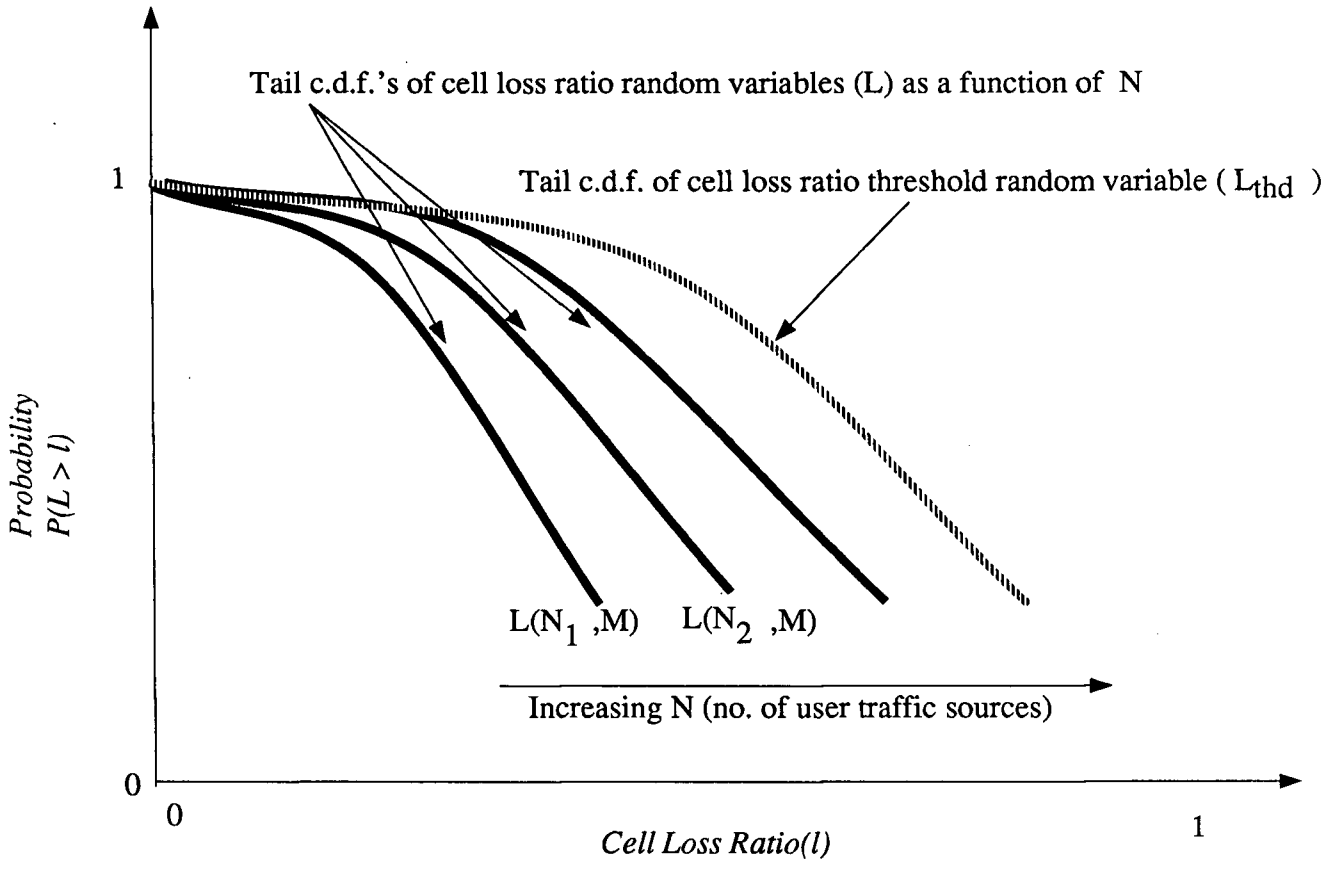


Figure 14 Tail c.d.f. of Cell Loss Ratio Threshold Random Variable as a Function of the Number of User Traffic Sources

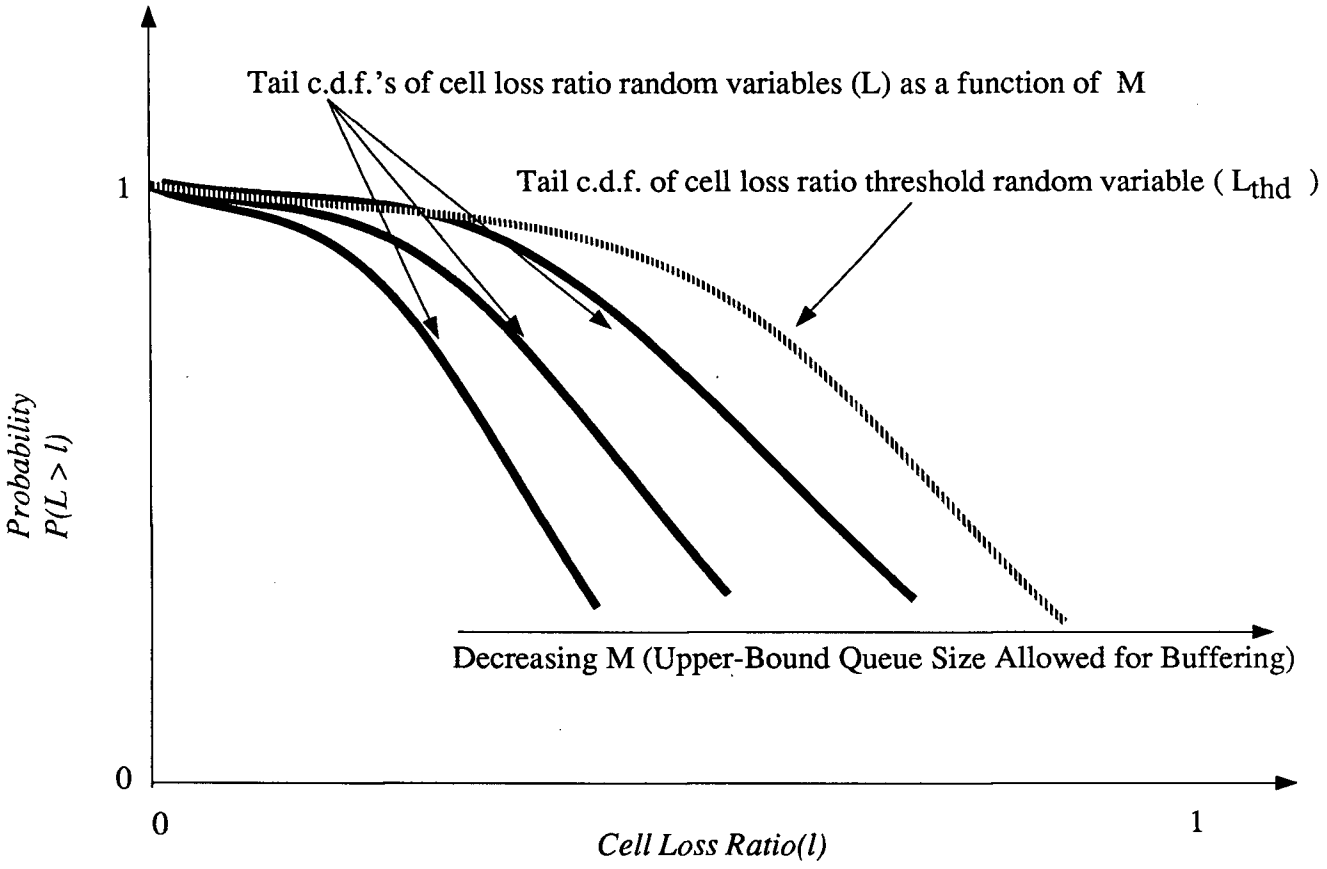


Figure 15 Tail c.d.f. of Cell Loss Ratio Threshold Random Variable as a Function of the Upper-Bound Queue Size Allowed for Buffering

Chapter 7 Application and Results

This chapter illustrates the applications of the integrated congestion management (call-level access and cell-level transfer congestion control schemes) to homogeneous user traffic sources. The first case deals with homogeneous telephony teleservice sources that require the support of circuit-mode bearer service. The second case deals with homogeneous data-handling teleservice sources that require packet-mode bearer service.

The call-level access and cell-level transfer control schemes are applicable to both homogeneous and heterogeneous traffic. The performance requirements for heterogeneous traffic can be determined by the performance requirements of the most demanding component traffic type, or a compromise among all the component traffic types. Since the computations of joint probability distributions for heterogeneous traffic require recursive calculations or a large number of convolutions (Section 3.1.2), homogeneous traffic are considered in these examples for computational ease.

7.1 Homogeneous Voice-Telephony Sources

In a B-ISDN ATM network, voice-telephony teleservice requires circuit-mode bearer service (e.g. 64 Kbps voice traffic). User's perception of the voice quality defines the user-oriented QOS attribute of the telephony teleservice. User perception is more sensitive to the QOS delay subattribute than to the QOS loss subattribute.

The required values of the network-oriented QOS delay and loss subattributes of the bearer service for voice is derived from the values of the corresponding user-oriented QOS subattributes of the telephony teleservice contributed by the information transport layers only (i.e. minus the contributions by the information processing layers).

The cell delay and cell loss ratio performance requirements of the voice bearer service for each network exchange (e.g. UNI access-node, network switch-node) is derived from the network-oriented QOS delay and loss subattributes of voice bearer service over the reference connection [32] which represents the longest connection as derived from CCITT Recommendation G.104. The values of these subattributes are as follows:

- Cell Delay of 160 ms.
- Cell Loss Ratio of 10^{-3}.

The corresponding estimates of the cell delay and cell loss ratio performance requirements of the voice bearer service for each network exchange are as follows:

- Cell Loss Delay of 1 ms.
- Cell Loss Ratio of 10^{-4}.

These requirements and the following parameters are employed to determine the call-level and cell-level control parameters:

- System Parameters (fixed):
  - $\overline{T_A}$ = Average duration of call active state = 0.185 seconds.
  - $\overline{T_I}$ = Average duration of call idle state = 1.31 seconds.
  - $\overline{R_C}$ = Average cell arrival rate during call active state = 64 Kbps
  - R_O = Maximum cell service rate = 150 Mbps.
  - $\overline{D}$ = Average duration of the *down* state of user traffic servicing = 1 cell slot.
  - $\overline{U}$ = Average duration of the *up* state of user traffic servicing = 10^6 cell slots.
- Controllable Parameters
  - M = Upper-bound queue size allowed for buffering.

- N = Number of homogeneous user traffic calls.

- Performance Requirements
  - M_{thd} (the upper-bound queue size threshold to meet the upper-bound cell delay requirement of 1 ms) = $1\text{ms} \times \frac{150\text{Mb/s}}{(53\times8)\text{b/cell}} = 353\text{cells}$
  - L_{thd} (the stationary cell loss ratio threshold random variable) is defined via the tail *c.d.f.*, i.e. $P(L_{thd} > l)$ where $0 < l \leq 1$. The cell loss ratio performance requirement of 10^{-4} means that $P(L_{thd} > 10^{-4})$ must be very small, e.g. 10^{-9}, and this becomes a point of the tail *c.d.f.* as follows:

Cell Loss Ratio (l)	1.0x 10^{-14}	1.0x 10^{-10}	1.0x 10^{-7}	1.0x 10^{-6}	1.0x 10^{-5}	1.0x 10^{-4}	1.0x 10^{-3}	1.0x 10^{-2}
Probability $P(L_{thd}>l)$	1.0x 10^{-2}	5.0x 10^{-3}	1.0x 10^{-3}	1.0x 10^{-5}	1.0x 10^{-7}	1.0x 10^{-9}	1.0x 10^{-11}	1.0x 10^{-13}

7.1.1 Call-Level Congestion Control

For call-level congestion control, N is subjected to limit-control while M is set to M_{thd} to meet upper-bound cell delay requirement. Fig. 16 shows the tail *c.d.f.*, $\overline{F_L}(l)$, of the stationary cell loss ratio random variable L, as a function of the number of user traffic calls (N). The following observations can be made:

- The tail *c.d.f.* curves shift rightward as N increases; i.e. L becomes stochastically larger as N increases. Recall that L_2 is stochastically greater than L_1 if and only if $\overline{F}_{L2}(l) \geq \overline{F}_{L1}(l)$ for all l.
- Up to 9000 calls can be accommodated while ensuring that the cell loss ratio stochastic requirement and the upper-bound cell delay requirement are met, i.e. the

call level controller would admit incoming call as long as the number of calls does not exceed 9000.

In this case of N = 9000 calls of voice-telephony traffic at 64Kbps,

$$\text{Output Rate} = 9000 \times 64\text{Kbps} \times \frac{0.185\text{sec}}{0.185\text{sec} + 1.31\text{sec}} = 71.3\text{Mbps}$$

$$\text{Link Utilization for User Traffic} = \frac{\text{Output Rate}}{\text{Link Transmission Rate}} = \frac{71.3\text{Mbps}}{150\text{Mbps}}$$

$$= 47.5\%$$

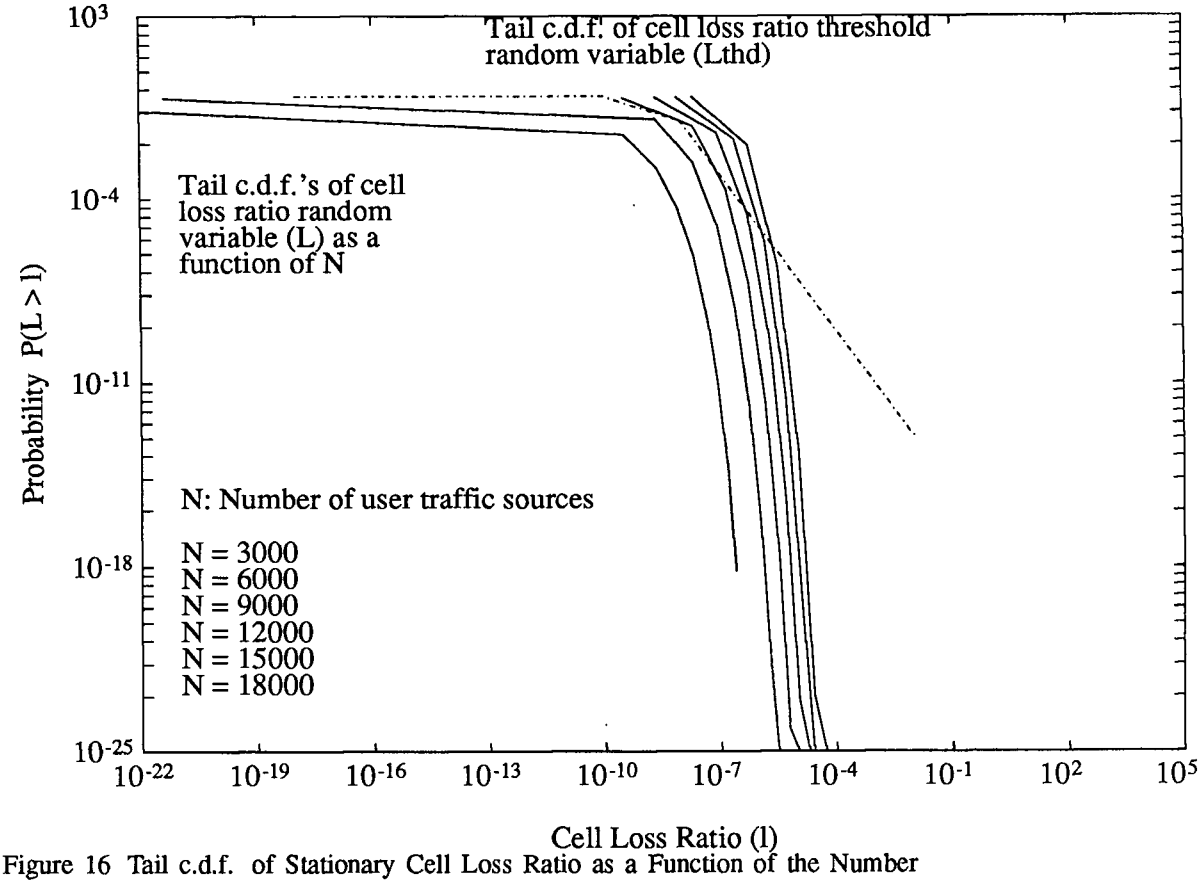


Figure 16 Tail c.d.f. of Stationary Cell Loss Ratio as a Function of the Number of Calls (Voice-Telephony Teleservice)

7.1.2 Cell-Level Congestion Control

For cell-level congestion control, the control policy is a threshold policy which rejects incoming cell if the current queue size reaches some upper-bound queue size M for the current number of homogeneous user traffic calls N. N is set at 5000 calls for this example.

For voice telephony teleservice, user is more sensitive to cell delay than to cell loss. Therefore, the upper-bound cell delay D is minimized by minimizing M while ensuring that $L(M,\ N = 5000) \overset{st}{\leq} L_{thd}$ and $D(M) \leq D_{thd}$. M is determined to be 200 cells from Fig. 17, which shows the tail *c.d.f.* of the stationary cell loss ratio random variable L ($\overline{F_L}(l)$) as a function of the upper-bound queue size M. The following observations can be made:

- The tail *c.d.f.* curves shift rightward as M decreases; i.e. L becomes stochastically larger as M decreases.
- M can be reduced from M_{thd} = 353 cells down to 200 cells while ensuring that the cell loss ratio stochastic requirement and the upper-bound cell delay requirement are met.

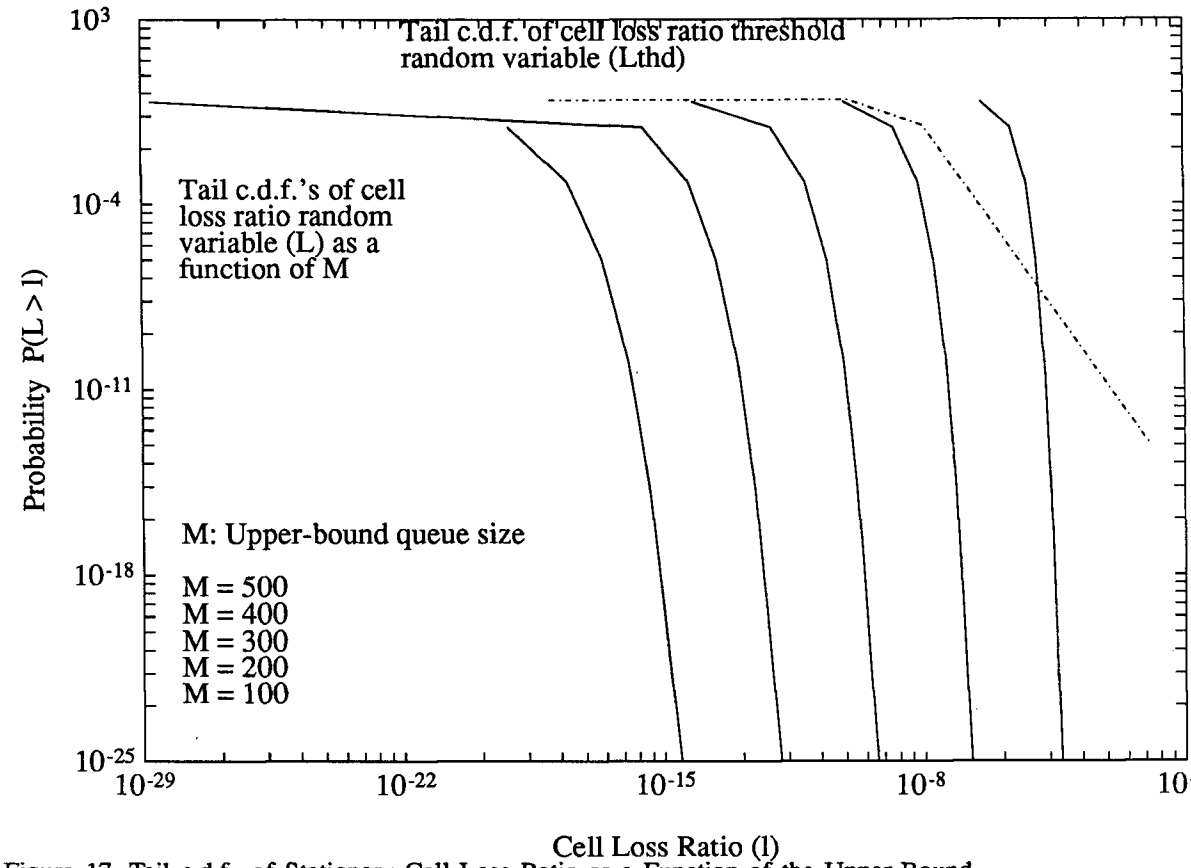


Figure 17 Tail c.d.f. of Stationary Cell Loss Ratio as a Function of the Upper-Bound Queue Size (Voice-Telephony Teleservice)

7.2 Homogeneous Data-Handling Sources

In a B-ISDN ATM network, data-handling teleservice requires packet-mode bearer service (e.g. 10 Mbps data traffic). User's perception of the quality of the data-handling teleservice defines the user-oriented QOS attribute of the data-handling teleservice. User perception is more sensitive to the QOS loss subattribute than to the QOS delay subattribute.

The required values of the network-oriented QOS delay and loss subattributes of bearer service for data is derived from the values of the user-oriented QOS delay and

loss subattributes of the message handling teleservice contributed by the information transport layers only (i.e. minus the contribution by the information processing layers).

The cell delay and cell loss ratio performance requirements of the data-handling bearer service for each network exchange (e.g. UNI access-node, network switch-node) is derived from the network-oriented QOS delay and loss subattributes of data-handling bearer service over the reference connection [32] which represents the longest connection as derived from CCITT Recommendation G.104. The values of these subattributes are as follows:

- Cell Delay of 200 ms.
- Cell Loss Ratio of 10^{-6}.

The corresponding estimates of the cell delay and cell loss ratio performance requirements of the data-handling bearer service for each network exchange are as follows:

- Cell Loss Delay of 4 ms.
- Cell Loss Ratio of 10^{-7}.

These requirements and the following parameters are employed to determine the call-level and cell-level control parameters:

- System Parameters (fixed):
  - $\overline{T_A}$ = Average duration of call active state = 1.728 seconds.
  - $\overline{T_I}$ = Average duration of call idle state = 156 seconds.
  - $\overline{R_C}$ = Average cell arrival rate during call active state = 10 Mbps
  - μ = Maximum cell service rate = 150 Mbps.
  - $\overline{D}$ = Average duration of the *down* state of user traffic servicing = 1 cell slot.

- $\overline{U}$ = Average duration of the *up* state of user traffic servicing = 10^6 cell slots.

- Controllable Parameters
  - M = Upper-bound queue size allowed for buffering.
  - N = Number of homogeneous user traffic calls.
- Performance Requirements
  - M_{thd} (the upper-bound queue size threshold to meet the upper-bound cell delay requirement of 4 ms) = 4ms $\times \frac{150\text{Mb/s}}{(53\times8)\text{b/cell}} = 1415\text{cells}$
  - L_{thd} (the stationary cell loss ratio threshold random variable) is defined via the tail *c.d.f.*, i.e. $P(L_{thd} > l)$ where $0 < l \leq 1$. The cell loss ratio performance requirement of 10^{-7} means that $P(L_{thd} > 10^{-7})$ must be very small, e.g. 10^{-9}, and this becomes a point of the tail *c.d.f.* as follows:

Cell Loss Ratio (l)	1.0x 10^{-14}	1.0x 10^{-12}	1.0x 10^{-10}	1.0x 10^{-7}	1.0x 10^{-6}	1.0x 10^{-5}	1.0x 10^{-4}	1.0x 10^{-3}
Probability $P(L_{thd} > l)$	1.0x 10^{-2}	0.5x 10^{-2}	1.0x 10^{-3}	1.0x 10^{-9}	1.0x 10^{-11}	1.0x 10^{-13}	1.0x 10^{-15}	1.0x 10^{-17}

7.2.1 Call-Level Congestion Control

For call-level congestion control, N is subjected to limit-control while M is set to M_{thd} to meet upper-bound cell delay requirement. Fig. 16 shows the tail *c.d.f.*, $\overline{F_L}(l)$, of the stationary cell loss ratio random variable L, as a function of the number of user traffic calls (N). The following observations can be made:

- As in the previous example, the tail *c.d.f.* curves shift rightward as N increases; i.e. L becomes stochastically larger as N increases.

- Up to 60 calls can be accommodated while ensuring that the cell loss ratio stochastic requirement and the upper-bound cell delay requirement are met. Therefore, the call level controller would admit incoming call as long as the number of calls does not exceed 60.

  In this case, with $N = 60$ calls of data-handling traffic at 10Mbps,

$$\text{Output Rate} = 60 \times 10\text{Mbps} \times \frac{1.728\text{sec}}{1.728\text{sec} + 156\text{sec}} = 6.573\text{Mbps}$$

$$\text{Link Utilization for User Traffic} = \frac{\text{Output Rate}}{\text{Link Transmission Rate}} = \frac{6.573\text{Mbps}}{150\text{Mbps}}$$

$$= 4.4\%$$

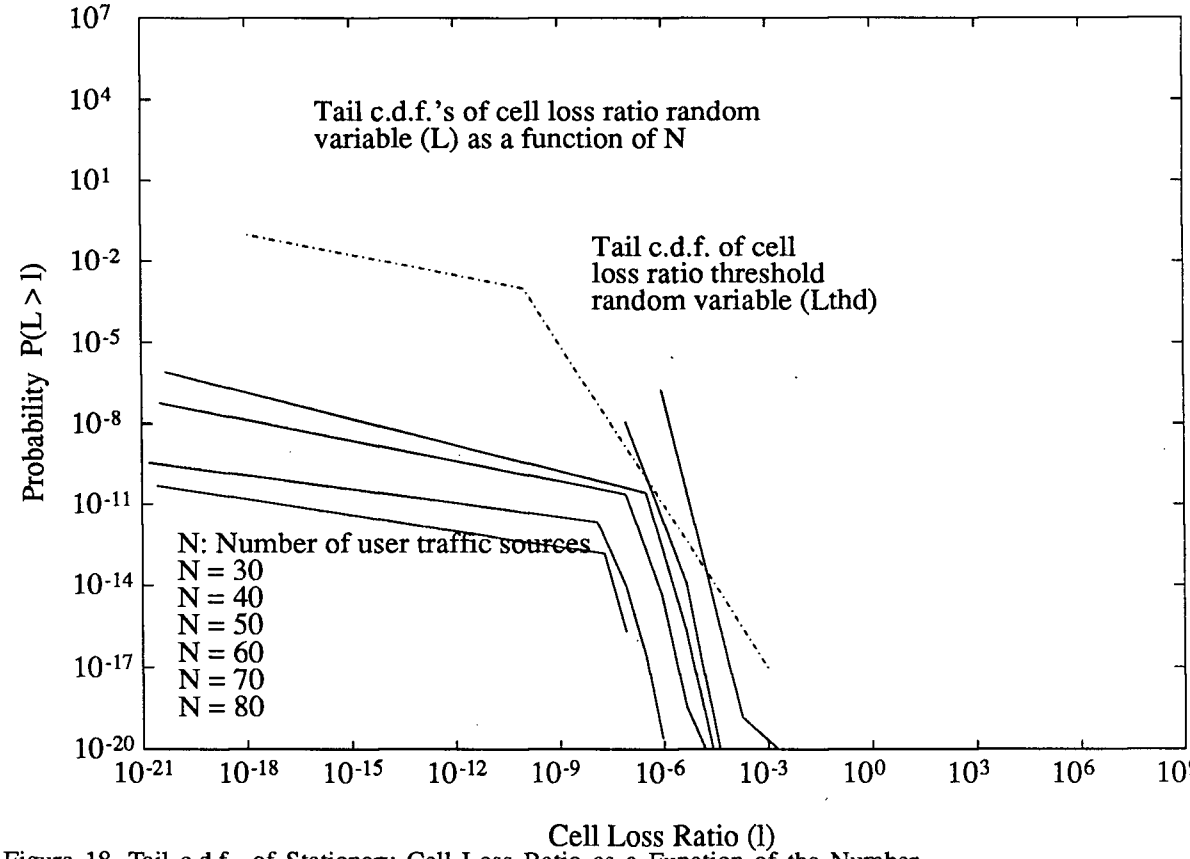


Figure 18 Tail c.d.f. of Stationary Cell Loss Ratio as a Function of the Number of Calls (Data-Handling Teleservice)

7.2.2 Cell-Level Congestion Control

For cell-level congestion control, the control policy is a threshold policy which rejects incoming cell if the current queue size reaches some upper-bound queue size M for the current number of homogeneous user traffic calls N. N is set at 30 calls for this example.

For data-handling teleservice, user is more sensitive to cell loss than to cell delay. Therefore the cell loss ratio L is minimized by maximizing M while ensuring that $L(M, N = 30) \overset{st}{\leq} L_{thd}$ and $D(M) \leq D_{thd}$. These conditions are always satisfied by setting $M = M_{thd} = 1415$ cells.

7.3 Comparisons of Results

The differences between results for the voice-telephony sources and the results for the data-handling sources are outlined as follows:

- *C.d.f.* curves for the voice-telephony sources are steeper than those for the data-handling sources. This indicates that the fluctuation of the cell loss for the the voice-telephony sources is smaller than for the data-handling sources.
- Link utilization for the voice-telephony sources is much greater than that for the data-handling sources.

Obviously, these differences in results are caused by the difference in the source rates. The source rate is 64 Kbps for voice-telephony sources, which is much slower than the 10 Mbps source rate for the data-handling sources. Without statistical multiplexing, 2340 voice-telephony calls or 15 data-handling calls can be time-division-multiplexed onto one link. With statistical multiplexing, the link can accommodate 9000 voice-telephony calls

or 60 data-handling calls. Therefore, a multiplexing gain of 3.8 and 4 is achieved for the voice-telephony sources and data-handling sources, respectively.

Since a large number of voice-telephony calls (9000) can be statistically multiplexed onto one link, the burstiness of the multiplexed voice traffic is very much reduced compared with that of each individual voice-telephony call. In contrast, only a small number of data-handling calls (60) can be statistically multiplexed onto one link, so that the burstiness of the multiplexed data traffic is not significantly diminished compared to each individual data-handling call.

When the bursty nature of the multiplexed traffic is diminished, fluctuation of cell loss becomes smaller and this results in more efficient link utilization.

Chapter 8 Conclusions

The followings have been accomplished in this research:

- The effect of the higher-priority signalling traffic on the multiplexing of user traffic through the ATM statistical multiplexer at the UNI has been studied. Consequently, a novel modeling technique for user-application traffic multiplexing through the ATM statistical multiplexer at the UNI has been proposed: it is characterized by a queueing model with random service disruptions due to the transport of the higher-priority signalling traffic.
- The congestion performance requirements of the user traffic for the UNI are studied in terms of the stochastic cell loss requirement and the deterministic upper-bound cell delay requirement. However, in order to investigate the stochastic cell loss phenomenon due to buffer overflow, the stochastic queue behaviour must first be examined. Consequently, a novel algorithm to solve the stationary distribution of the queue length process under short-term and finite buffer capacity conditions has been presented:
  - Analyze the continuous-state queue length process under long-term heavy traffic and infinite buffer capacity conditions via the fluid model with random disruptions. Determine the effect of buffer capacity on the discrete-state dynamics via the M/D/1 model subjected to the same random disruptions as the fluid model.
  - Determine the limiting condition under which the fluid model with random disruptions is a reasonable approximation to the M/D/1 model subjected to the same random disruptions in describing the queue behaviour under short-term

heavy traffic condition. Under the limiting condition, the stationary discrete-state distribution function with finite buffer capacity is derived from the continuous-state distribution function with infinite buffer capacity while taking into account of the effect of buffer capacity on discrete-state dynamics.

- An integrated congestion management platform at the UNI has been proposed. Integrated congestion management dictates that congestion control schemes are applied during the call access phase (call admission control scheme) and the information transfer phase (intra-node buffer control scheme) of user traffic source. The congestion control schemes are devised to meet the congestion performance requirements and to optimize the performance if possible.

  - A novel UNI call admission control scheme (implemented via a call-level access controller) has been proposed, and its objective is to maintain the required network performance assigned to the UNI access-node by exerting call admission control in the call access phase of each user traffic source. The control scheme has been analyzed using an input-limit static control model employing stochastic ordering between the cell loss ratio random variable and the desired threshold random variable as a criterion to decide if a new call should be admitted. The cell loss ratio random variable has been chosen as the performance objective rather than the long-term-time-averaged cell loss ratio, so as to take into account of the dynamic nature of bursty traffic sources.
  - A novel UNI intra-node buffer control scheme (implemented via cell-level transfer controller) has been proposed, and its objective is to optimize the network performance of the UNI access-node by exerting buffer control in the aggregate

information transfer phase of the user traffic sources. The network performance parameters chosen for cell-level congestion control is the cell loss ratio for loss sensitive teleservice, or the upper-bound cell delay for delay sensitive teleservice. The control scheme has been analyzed by means of a sequential decision process model characterized by a stationary, Markovian and deterministic threshold control policy.

Areas for further research would include the followings:

- Extend the cell-level and call-level congestion control schemes to the ATM switching node.
- Examine the performance requirements for heterogeneous traffic so that the cell-level and call-level congestion control schemes can be applied transparently in both the single-media case or in the multi-media case where user traffic sources are heterogeneous in nature.
- Extend the proposed congestion management model to heterogeneous networks architecture; where the B-ISDN ATM-based network acts as an backbone network interconnecting LANs and radio networks via MAN.

Appendix A List of Acronyms and Abbreviations

ATM Asynchronous Transfer Mode

ISDN Integrated Services Digital Network

B-ISDN Broadband ISDN

CCITT International Telegraph and Telephone Consultative Committee

CBR Constant Bit Rate

c.d.f. Cumulative Distribution Function

CIR Cell Insertion Ratio

CLR Cell Loss Ratio

CCS Common Channel Signalling

CL-VBR Connectionless-oriented Variable Bit Rate

CO-CBR Connection-oriented Constant Bit Rate

CO-VBR Connection-oriented Variable Bit Rate

DSRP Doubly Stochastic Renewal Process

FIFO First-In-First-Out

MAN Metropolitan Area Network

MMPP Markov Modulated Poisson Process

N-ISDN Narrowband ISDN

OSI Open Systems Interconnection

p.m.f. Probability Mass Function

QOS Quality of Service

SBBP Switched Batch Bernoulli Process

SNACP SubNetwork Access Convergence Protocol

SNDCP SubNetwork Dependent Convergence Protocol

SNICP SubNetwork Independent Convergence Protocol

SONET Synchronous Optical Network

STM Synchronous Transfer Mode

SC Switching Center

TA Terminal Adaptor

TE Terminal Equipment

UNI User Network Interface

VBR Variable Bit Rate

Bibliography

[1] W. Stallings, *ISDN An Introduction.* MacMillan, New York, 1989.

[2] J. Luetchford, "CCITT Recommendations on the ISDN: A Review," *IEEE Journal on Selected Areas in Communications*, May 1986.

[3] R. Potter, "ISDN Protocol and Architecture Models," *Computer Networks and ISDN Systems*, no. 10, 1985.

[4] E. I. T. K. K. Murano, Koso Murakami and H. Ogasawara, "Technologies Towards Broadband ISDN," *IEEE Communications*, vol. 28, pp. 66–70, April 1990.

[5] S. Yoneda and H. Salloum, "B-ISDN User Network Interface: Performance Monitoring Functions Using SONET Overhead," *ICC'90*, vol. 3, April 1990.

[6] H. Armbruster, "Broadband Communication and Its Realization with Broadband ISDN," *IEEE Communications Magazine*, November 1987.

[7] H. M. X. Liu, "Design of a Hybrid Transfer Mode Broadband ISDN," *ICC'90*, vol. 2, pp. 315.1.1–315.1.5, April 1990.

[8] K. Ross and D. Tsang, "Optimal Circuit Access Policies in an ISDN Environment: A Markov Decision Approach," *IEEE Trans. Commun.*, vol. 37, pp. 934–939, September 1989.

[9] B. Maglaris and M. Schwartz, "Performance Evaluation of a Variable Frame Multiplexer for Integrated Switched Networks," *IEEE Trans. Commun.*, vol. 29, pp. 800–807, June 1981.

[10]B. Maglaris and M. Schwartz, "Optimal Fixed Frame Multiplexing in Integrated Line- and Packet-Switched Communication Networks," *IEEE Trans. Information Theory*, vol. 28, pp. 263–273, March 1982.

[11]B. Kraimeche and M. Schwartz, "Analysis of Traffic Access Control strategies in Integrated Service Networks," *IEEE Trans. Commun.*, vol. 33, pp. 1085–1093, October 1985.

[12]F. D. M. Aicardi, R. Bolla and R. Minciardi, "Optimization of Capacity Allocation among Users and Services in Integrated Networks," *ICC'90*, vol. 2, pp. 302.3.1–302.3.7, April 1990.

[13]I. Viniotis and A. Ephremides, "Optimal Switching of Voice and Data at a Network Node," *Proc. 26th Conf. Decision and Control*, pp. 1504–1507, December 1987.

[14]M. Hirano and N. Watanabe, "Characteristics of a Cell Multiplexer for Bursty ATM Traffic," *ICC'89*, p. 13.2, June 1989.

[15]M. Decina and T. Toniatti, "On Bandwidth Allocation to Bursty Virtual Connections in ATM Networks," *ICC'90*, vol. 3, p. 318.6, April 1990.

[16]G. R. G. Gallassi and L. Fratta, "ATM: Bandwidth Assignment and Bandwidth Enforcement Policies," *Proc. GLOBECOM'89*, pp. 49.6.1–49.6.6, November 1989.

[17]P. Joos and W. Verbiest, "A Statistical Bandwidth Allocation and Usage Monitoring Algorithm for ATM Networks," *ICC'89*, pp. 13.5.1– 13.5.8, June 1989.

[18]T. Kamitake and T. Suda, "Evaluation of an Admission Control Scheme for an ATM Network Considering Fluctuations in Cell Loss Rate," *GLOBECOM'89*, p. 49.4, November 1989.

[19]E. Rathgeb, "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 325–334, April 1991.

[20]E. C. M. Butto and A. Tonietti, "Effectiveness of the 'Leaky Bucket' Policing Mechanisms in ATM Networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 335–342, April 1991.

[21]L. Kleinrock, *Queueing Systems*. John Wiley and Sons, 1976.

[22]H. Heffes and D. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 856–868, September 1986.

[23]H. Saito and M. Kawarasaki, "An Analysis of Statistical Multiplexing in an ATM Transport Network," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 359–367, April 1991.

[24]J. K. R. Nagarajan and D. Towsley, "Approximation Techniques for Computing Packet Loss in Finite-Buffered Voice Multiplexers," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 368–377, April 1991.

[25]A. S. I. Norros, J.W. Roberts and J. Virtamo, "The Superposition of Variable Bit Rate Sources in an ATM Multiplexer," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 378–387, April 1991.

[26]M. L. A. R. A. Baiocchi, N.B. Melazzi and B. Winkler, "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 388–393, April 1991.

[27]Y. T. O. Hashida and S. Shimogawa, "Swithced Batch Bernoulli Process (SBBP) and the Discrete-Time SBBP/G/1 Queue with Application to Statistical Multiplexer Performance," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 394–401, April 1991.

[28]G. R. L.G. Dron and B. Sengupta, "Delay Analysis of Continuous Bit Rate Traffic Over an ATM Network," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 402–407, April 1991.

[29]J.-Y. L. Boudec, "An Efficient Solution Method for Markov Models of ATM Links with Loss Priorities," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 408–417, April 1991.

[30]P. B. H. Kroner, G. Hebuterne and A. Gravey, "Priority Management in ATM Switching Nodes," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 418–427, April 1991.

[31]L. Dittmann and S. Jacobsen, "Statistical Multiplexing of Identical Bursty Sources in an ATM Network," *GLOBECOM'88*, November 1988.

[32]G. S. M.E. Theológou and E. Protonotarios, "Service Performance Requirements in the Asynchronous Transfer Mode (ATM) Environment," *IEEE Proc. of the workshop on Network Management and Control*, pp. 117–128, September 1990.

[33]R. Wolff, *Stochastic Modelling and the Theory of Queues*, ch. 8. Prentice Hall, 1989.

[34]S. Ross, *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.

[35]K. Siriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 833–

846, September 1986.

[36]D. Mitra, "Stochastic Theory of a Fluid Model of Multiple Failure-Susceptible Producers and Consumers Coupled by a Buffer," *Adv. Appl. Prob.*, pp. 646–676, 1988.

[37]A. M. H. Chen, "Discrete Flow Networks: Bottleneck Analysis and Fluid Approximations," *Math of OR.*, 1990.

[38]J. Vandergraft, "A Fluid Flow Model of Network of Queues," *Mgmt. Sci.*, vol. 29, pp. 1198–1208, 1983.

[39]D. Y. H. Chen, "A Fluid Model for Systems with Random Disruptions," *Paper DDM-89-09972*, February 1990.

[40]S. Asmussen, *Applied Probability and Queues*, ch. 7. John Wiley and Sons, Chichester, 1987.

[41]H. Kekre and M. Khalid, "Buffer Design in a Closed Form with Hybrid Input and Random Server Interruptions," *IEEE Proc.*, vol. 127, pp. 448–455, December 1980.

[42]S. J. Dittmann and K. Moth, "Flow Enforcementt Algorithms for ATM Networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 343–350, April 1991.

[43]K. K. C. Rasmussen, J.H. Sorensen and S. Jacobsen, "Source-Independent Call Acceptance Proceduress in ATM Networks," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 351–358, April 1991.

[44]G. Neistegge, "The Leaky Bucket Policing Method," *Int'l J. of Digital and Analog Cabled System*, vol. 2, June 1990.

[45]J. Walrand, *Introduction to Queueing Networks*, ch. 8, pp. 278–281. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[46]D. Bertsekas, *Dynamic Programming*. Prentice Hall, Englewood Cliffs, 1987.

[47]M. Puterman, *Markov Decision Processes*. Wiley, New York, 1991.