ABSTRACT

BUFFER MANAGEMENT AND CELL SWITCHING MANAGEMENT IN WIRELESS PACKET COMMUNICATIONS

by

Jongho Bang

The buffer management and the cell switching (e.g., packet handoff) management using buffer management scheme are studied in Wireless Packet Communications.

First, a throughput improvement method for multi-class services is proposed in Wireless Packet System. Efficient traffic management schemes should be developed to provide seamless access to the wireless network. Specially, it is proposed to regulate the buffer by the "Selective-Delay Push-In (SDPI)" scheme, which is applicable to scheduling delay-tolerant non-real time traffic and delay-sensitive real time traffic. Simulation results show that the performance observed by real time traffics are improved as compared to existing buffer priority scheme in term of packet loss probability.

Second, the performance of the proposed SDPI scheme is analyzed in a single CBR server. The arrival process is derived from the superposition of two types of traffics, each in turn results from the superposition of homogeneous ON-OFF sources that can be approximated by means of a two-state Markov Modulated Poisson Process (MMPP). The buffer mechanism enables the ATM layer to adapt the quality of the cell transfer to the QoS requirements and to improve the utilization of network resources. This is achieved by selective-delaying and pushing-in cells according to the class they belong to. Analytical expressions for various performance parameters and numerical results are obtained. Simulation results in term of cell loss probability conform with our numerical analysis.

Finally, a novel cell switching scheme based on TDMA protocol is proposed to support QoS guarantee for the downlink. The *new packets* and *handoff packets* for

each type of traffic are defined and a new cutoff prioritization scheme is devised at the buffer of the base station. A procedure to find the optimal thresholds satisfying the QoS requirements is presented. Using the ON-OFF approximation for aggregate traffic, the packet loss probability and the average packet delay are computed. The performance of the proposed scheme is evaluated by simulation and numerical analysis in terms of packet loss probability and average packet delay.

BUFFER MANAGEMENT AND CELL SWITCHING MANAGEMENT IN WIRELESS PACKET COMMUNICATIONS

by
Jongho Bang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering, NJIT

May 2001

APPROVAL PAGE

Buffer Management and Cell Switching Management in Wireless Packet Communications

Jongho Bang

4/18/01

Dr. Sirin Tekinay, Dissertation Advisor — Date
Assistant Professor, Department of Electrical and Computer Engineering, NJIT

4/18/2001

Dr. Nirwan Ansari, Dissertation Co-Advisor — Date
Professor, Department of Electrical and Computer Engineering, NJIT

4/18/01

Dr. Yun-Qing Shi, Committee Member — Date
Associate Professor, Department of Electrical and Computer Engineering, NJIT

04/18/01

Dr. Symeon Papavassiliou, Committee Member — Date
Assistant Professor, Department of Electrical and Computer Engineering, NJIT

04/18/01

Dr. Malathi Veeraraghavan, Committee Member — Date
Associate Professor, Department of Electrical Engineering, Polytechnic University

BIOGRAPHICAL SKETCH

Author : Jongho Bang

Degree : Doctor of Philosophy

Date : May 2001

Undergraduate and Graduate Education :

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology (NJIT), Newark, NJ, USA, 2001

- Master of Science in Electronics Engineering,
  Chung-ang University, Seoul, Korea, 1992

- Bachelor of Science in Electronics Engineering,
  Chung-ang University, Seoul, Korea, 1990

Major : Electrical Engineering

Presentations and Publications:

Jongho Bang, Sirin Tekinay and Nirwan Ansari,

"A Novel Capacity Maximization Scheme for Multimedia Wireless ATM System Providing QoS Guarantees for Handoffs,"

IEEE Vehicular Technology Conference, Tokyo, Japan, May 2000.

Jongho Bang, Nirwan Ansari and Sirin Tekinay,

"Selctive-Delay Push-In Buffering Mechanism for QoS Provisioning In ATM Switching Nodes Loaded with ON-OFF Arrival Processes,"

The 15th International Conference on Information Networking (ICOIN-15), Beppu, Japan, Jan. 2001.

To my mother

To my mother-in-law

To my wife, Leeyoung Whang

To my son, David Sungjun Bang

To my daughter, Ellen Eunseo Bang

ACKNOWLEDGMENT

I wish to express my sincere gratitude to my advisor, Professor Sirin Tekinay, who provided constant supervision, many suggestions, continuous support and encourage throughout the course of study.

I would like to express my deepest gratitude to my co-advisor, Dr. Nirwan Ansari. His advice, guidance and insight helped me enormously throughout this research.

Special thanks to Dr. Malathi Veeraraghavan for serving as members of committee, having kindly through the original manuscript, and providing valuable suggestions.

My gratitude is extended to Dr. Yun-Qing Shi, and Dr. Symeon Papavassiliou for serving as members on the dissertation committee and for their comments.

I would like to sincerely thank my mother, brother, and sister. Also, I would like to thank my mother-in-law, brothers-in-law, and my son, David Sungjun Bang, and my daughter, Ellen Eunseo Bang for their love and support.

I dedicate this dissertation with my love to my wife, Leeyoung Whang.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 General

Broad-band wireless network technologies such as IMT-2000 and Wireless ATM (WATM) are motivated by the increasing importance of portable computing and telecommunication applications. The rapid penetration of cellular phones and laptop PC's during the previous decade is proof that users place a significant value on portability as a key feature which enables tighter integration of such technologies with daily lives. This type of computing technologies will make it possible for services such as videotelephony, electronic banking, yellow pages, map services and local advertising to be provided over a wireless medium to a mobile user while on the move.

With the advent of the World Wide Web, the Internet has grown beyond reasonable imagination from a network that was intended for collaboration among a selective group of researchers to a network that is rapidly influencing our lives by changing the existing paradigms of communication and opening new avenues. Specifically, the Internet has paved the way for data networking, in which networks based on the IP packet-switched model will support voice, data, and video within a unified network infrastructure. Meanwhile, the cellular market continues to grow at an impressive pace. Not surprisingly, the volume of cellular data devices is expected to grow at a phenomenal rate. Thus, it appears that the cellular data sector, which is expected to benefit from growth in both the Web and cellular areas, promises to be an exciting area for technology innovation. So, high spectral efficiency and flexible data rate access are the main focus for future wireless network [1], as well as the development trend of existing networks toward the third generation (3G). To accomplish this goal, packet switching has been introduced to time-division multiple access (TDMA)-based systems. For instance, the proposed General Packet Radio Service (GPRS) for Global System for Mobile Communications (GSM) [2] and GPRS-136

for the North American Standard IS-136 [3] are forming the mainstream of evolution toward 3G.

Handoff is extremely important in any mobile network because of the default cellular architecture employed to maximize spectrum utilization. When the mobile terminal moves away from a base station, the signal level degrades and there is a need to switch communications to another base station. For a voice user, handoff results in an audible click interrupting the conversation for each handoff; and because of handoff, data users may lose packets and unnecessary congestion control measures may come to play. While significant work has been done handoff mechanisms in circuit-switched mobile networks [4], there is not much literature available on handoff in packet-switched mobile networks.

1.2 Cellular Packet Switched Network

From the user's perspective, wireless packet data networks (which employ packet-switching) offer an alternative that usually guarantees both cheaper and improved services in a vast range of applications.

1.2.1 CDPD

CDPD, Cellular Digital Packet Data, was initially designed as an overlay system on top of the AMPS networks. Subsequently, it was adapted to IS-95 and IS-136 networks. Services provided are access to networks based on IP and Connectionless Network Protocol (CLNP). Fig. 1.1 shows a network view of the CDPD network. The network nodes of CDPD are home and serving mobile data intermediate systems (MD-ISs) and the mobile data base station (MDBS). Basically, intermediate systems are IP-capable routers that form the backbone of the CDPD network. They are responsible for relaying user data, network administration, and mobility information. The home MD-IS stores the mobile station profile, authenticates the mobile station,

and provides the point of entry for IP datagrams destined for a mobile station, which are encapsulated and routed to the proper serving MD-IS. There are two layers of mobility management in the CDPD network. The home MD-IS performs macro mobility management of tracking which serving MD-IS is currently serving the mobile station, while the serving MD-IS is in charge of the micro mobility management of tracking the mobile station down to the cell level [5].

CDPD has its own set of databases (independent of the AMPS, IS-95, or IS-136 networks) for mobility management and subscriber profile information.

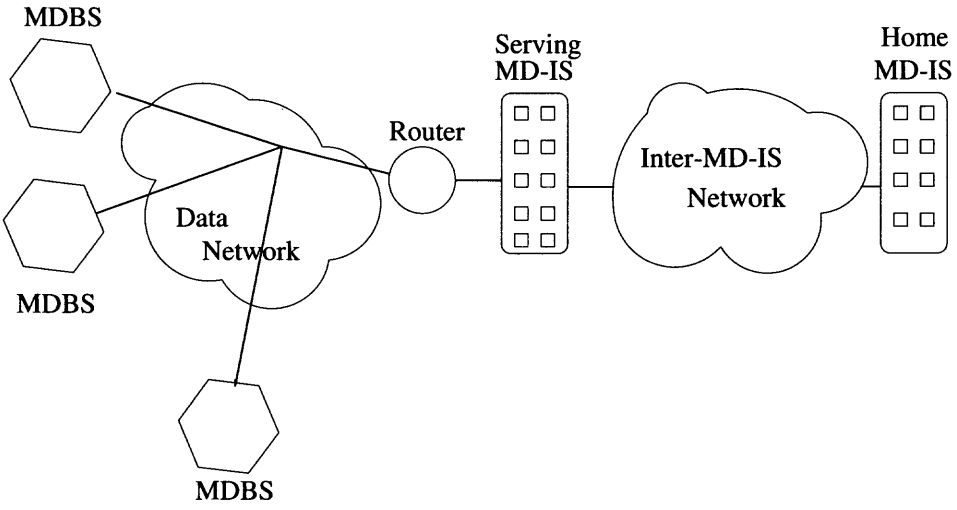

Figure 1.1 A network view of the CDPD network

The corresponding signaling protocol to manage the data structures is also specific to CDPD and unrelated to IS-41, the counterpart signaling protocol in the underlying AMPS, IS-95, or IS-136 cellular network.

1.2.2 GPRS

GPRS, General Packet Radio Service, is a new GSM service introduced in order to provide more efficient access to packet data networks from cellular networks.

GPRS is based on packet transmission over the air interface and in the network, and therefore allows more efficient resource utilization. GPRS is particularly well suited to carrying Internet traffic, which is often bursty with fluctuating data rate requirements. GPRS defines a general framework for cellular connection to a variety of packet data network. GPRS introduces a totally new backbone network based on IP, composed of new packet network nodes and traditional packet Internet nodes. Fig. 1.2 provides a network view of regular GPRS, as designed for GSM. GPRS adds two main network elements to the existing infrastructure: the serving GPRS support node (SGSN) and the gateway GPRS support node (GGSN). These elements interact with each other and with the existing cellular network elements over a set of new interfaces.

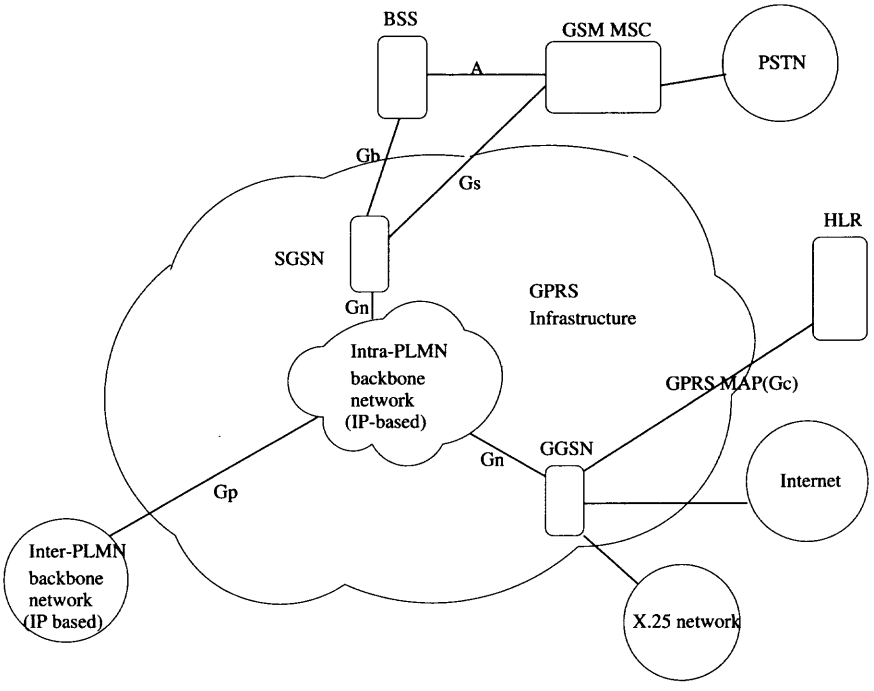

Figure 1.2 The regular GPRS network architecture

In particular, two new interfaces are standardized: a Gb interface between the BS subsystem (BSS) and the SGSN, and the Gs interface between the SGSN and the mobile switching center (MSC) [6].

SGSN takes care of terminal mobility and authentication functions, and is connected to the BSS over a frame relay network on one side and to the GGSN over an IP backbone network on the other. GGSN, in turn, provides connections and access to external networks. As regards the external IP network, GGSN can be seen as performing common IP router functions.

1.2.3 Wireless LAN

The IEEE 802.11 Wireless LAN (WLAN) is an extension to, or an alternative for, a wired LAN in a building or campus. WLANs provide the functionality of wired LANs, but without the physical constraints of the wire itself. Packets of data are converted into radio waves or infrared (IR) light pulses that are sent to other wireless devices or to a wireless access point - a device that bridges wireless traffic to a wired network.

1.3 An Overview of Handoff Management

Handoff is a basic mobile network capability for dynamic support of terminal migration. Handoff management is the process of initiating and ensuring a seamless and lossless handoff of a MT from the region covered by one base station to another base station.

1.3.1 Phases in a Handoff Procedure

There are three phases in a handoff procedure. These phases are shown in Fig. 1.3.

- *Measurements*: The mobile terminal as well as the base station do several measurements continuously. The signal strength is one parameter which might be measured by both the terminal and the base station.

- *Decision*: Based on the measurement taken, a decision is made as to whether is required. A decision to perform a handoff might be taken if the signal strength goes below a specified threshold.

- *Execution*: The actual handoff of the terminal from one cell to another is done in this phase. There are essentially two sub-phases in the execution of the handoff (e.g., new link establishment and release of old link).

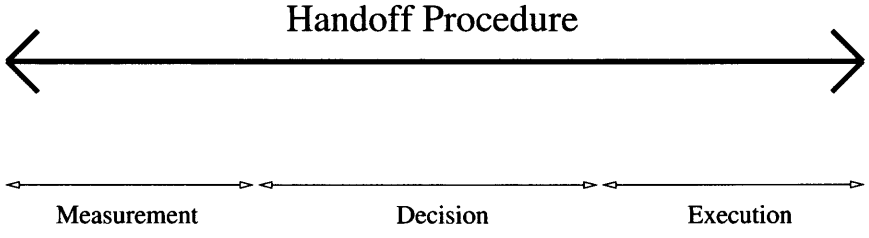

Figure 1.3 Phases in a Handoff Procedure

1.3.2 Handoff Types

The handoff procedures attempt to maintain the connections from a terminal as it migrates from one cell to another. There are various criteria base on which handoffs are classified.

1.3.2.1 Based on the Location of the Handoff Functions

- *Mobile Initiated Handoff*: The MT has to manage the handoff. That is, it takes the measurements on the downlink, processes them, takes the decision to do the handoff and decides the target base station.

- *Network Initiated Handoff*: The network manages the handoff, which includes taking measurements on the uplink, processing them, deciding to do the handoff, deciding the target base station.

- *Mobile Assisted Handoff*: This is similar to the network initiated handoff, except that the mobile assists the network by taking measurements along the downlink and relaying them back to the network.

1.3.2.2 Based on the Network Elements involved: The handoff procedures can be classified based on the network elements that are involved in the handoff.

- *Intra Cell*: This type of handoff is done within the current coverage area i.e., cell. The used channel is only changed for this type of handoff.

- *Inter Cell*: If the MT crosses cell boundaries, then it is referred to as inter cell handoff.

- *Inter Network*: If the handoff is done between two different networks, then it is referred to as inter network handoff.

1.3.2.3 Based on Number of Active Connections: The handoffs can also be classified based on the number of connections that a mobile terminal maintains during the handoff procedure.

- *Hard Handoff*: The MT switched the communication from the old link to the new link. Thus, there is only on active connection from the MT at any time. There is a short interrupt in the transmission. This interrupt should be minimized in order to make the handoff seamless.

- *Soft Handoff*: The MT is connected simultaneously to two access points. As it moves from one cell to another, it "softly" switches from on a base station to another. When connected to two base stations, the network combines information received from two different routes to obtain a better quality. This is commonly referred to as macro diversity.

1.3.2.4 Based on the Direction of the Handoff Signaling: Another way of classifying the handoffs is the direction of the handoff signaling.

- *Forward Handoff*: After the MT decides the cell to which it will make a handoff, it contacts the base station controlling the cell. The new base station initiates the handoff signaling to link the MT from the old base station. This is especially useful if the MT suddenly loses contact with the current base station.

- *Backward Handoff*: After the MT decides the cell to which it attempts to make a handoff, it contacts the current base station, which initiates the signaling to do the handoff to the new base station.

1.3.3 Requirements for a Handoff Scheme

- *Handoff Latency*: The time required to effect the handoff should be appropriate for the rate of mobility of the mobile terminal. That is, the decision to do the handoff should be valid for the current position of the mobile terminal after the handoff is completed.

- *Quality of Service*: In the context of Wireless ATM, the handoff procedure should attempt to maintain the requested QoS after the handoff is completed. However, since this is not always possible, a handoff mechanism should be capable of QoS re-negotiation.

- *Buffer Strategy*: The handoff strategy should avoid changes to the network switch buffer hardware implementations. The tradeoff between buffering (to ensure a lossless handoff) and packet loss (to ensure a seamless handoff) is made based on traffic class.

- *Group Handoff*: In the context of Wireless ATM, the handoff procedure should facilitate the handoff of a group of VCs. This property is especially useful in realizing a QoS controlled handoff.

1.3.4 Resource Allocation Schemes for Handoff in a Cellular Circuit Switched Network

In a cellular circuit-switched wireless network, a call can be terminated due to non availability of channels when handoff occurs and termination of an existing call has more impact on the system performance from the point of view of the user than the blocking of a new call. However, minimization of this can be achieved by sacrificing new call blocking performance as new calls compete for these channels. Handoff priority schemes have been proposed to give handoff preference in channel assignment over new arrivals.

- *Guard Channel scheme*: The classical handoff schemes considered the problem of sharing channels appropriately between new calls and handoff calls for one class of traffic, namely, voice conversation in a macrocellular environment. A relatively simple scheme called guard channel (e.g., cutoff priority) scheme, first proposed by Hong and Rappaport in [7], has been shown to be effective for such systems. In the guard channel scheme, new calls and handoff calls are treated equally on a FCFS basis for channel allocation until a predetermined channel utilization threshold is reached. At this point, new calls are simply blocked (e.g., cutoff), and only handoff call requests are honored.

- *Queueing Handoff Request scheme*: If the handoff attempt finds all channels in the target cell occupied, it can be queued. If any channel is released while the mobile is in the handoff area, the next queued handoff attempt is accomplished successfully. If the received power level from the source cell's base station falls below the receiver threshold level prior to the mobile being assigned a channel in the target cell, the call is forced into termination. When a channel is released in the cell, it is assigned to the next handoff call attempt waiting in the queue. If more than one handoff call attempt is in the queue, the FCFS [7] or dynamic priority [8] queueing discipline is used.

1.3.5 Rerouting Schemes for Handoffs (i.e., Network Handoff)

In wireless networks, a connection terminating at a mobile user may require dynamic reestablishment during the short time span necessary for terminal handoff due to its movement from one cell to another. The connection reestablishment procedure has to ensure in-sequence and loss-free delivery of the packets containing user data. There are several approaches proposed to handle network handoffs, which have completely different characteristics, performance, and impact on the wired network [10].

- *Connection Extension*: This approach prolongates the VC between the terminals by adding one hop that provides the connection from the source base station to the destination base station through the fixed network. This path extension can be performed by the source base station, as shown in Fig. 1.4. The advantage of this approach is twofold: simple and reasonably fast extension, and intrinsic preservation of packet sequence. But, the resource waste is remarkable [11].

- *Incremental Reestablishment*: This technique is appealing because it requires only the establishment of a new partial path (without the involvement of the remote terminal and network entities) which connects to a portion of the

original connection path, therefore allowing VCs to be partly reused. Because of spatial locality in movement, it is very likely that the reestablished path to the new location of the mobile user shares most of the original path. As a consequence, this technique is expected to be fast, efficient, and transparent, so it can be imagined that the end user does not perceive the network handoff as a service interruption [12]. Fig. 1.5 shows the path rerouting performed while the terminal moves through the network.

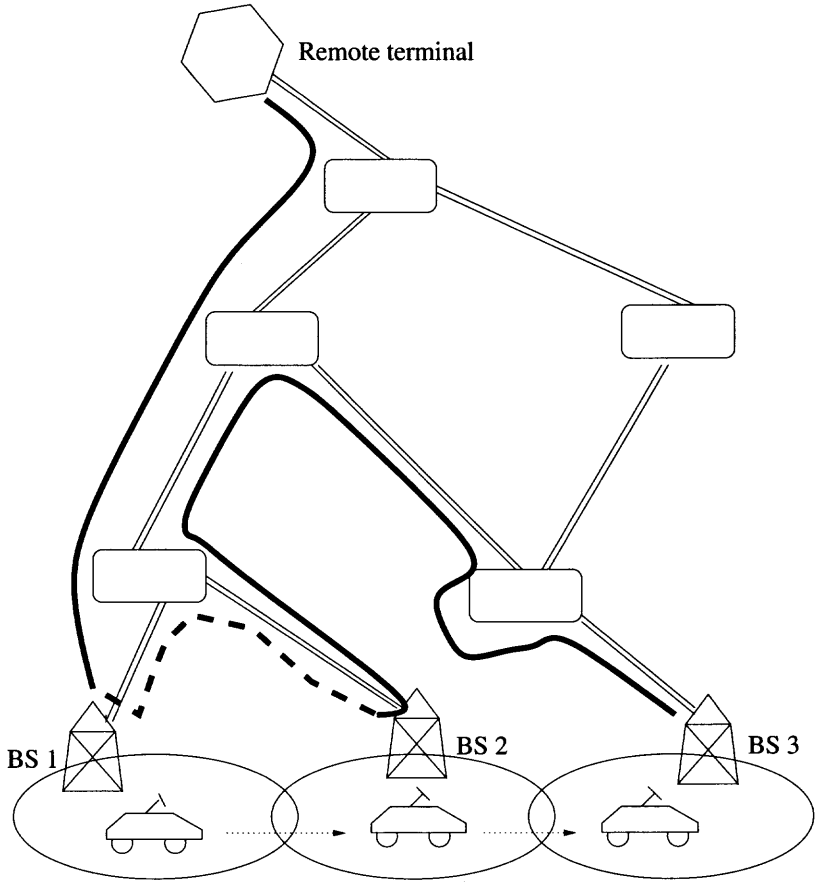

Figure 1.4 The connection extension case

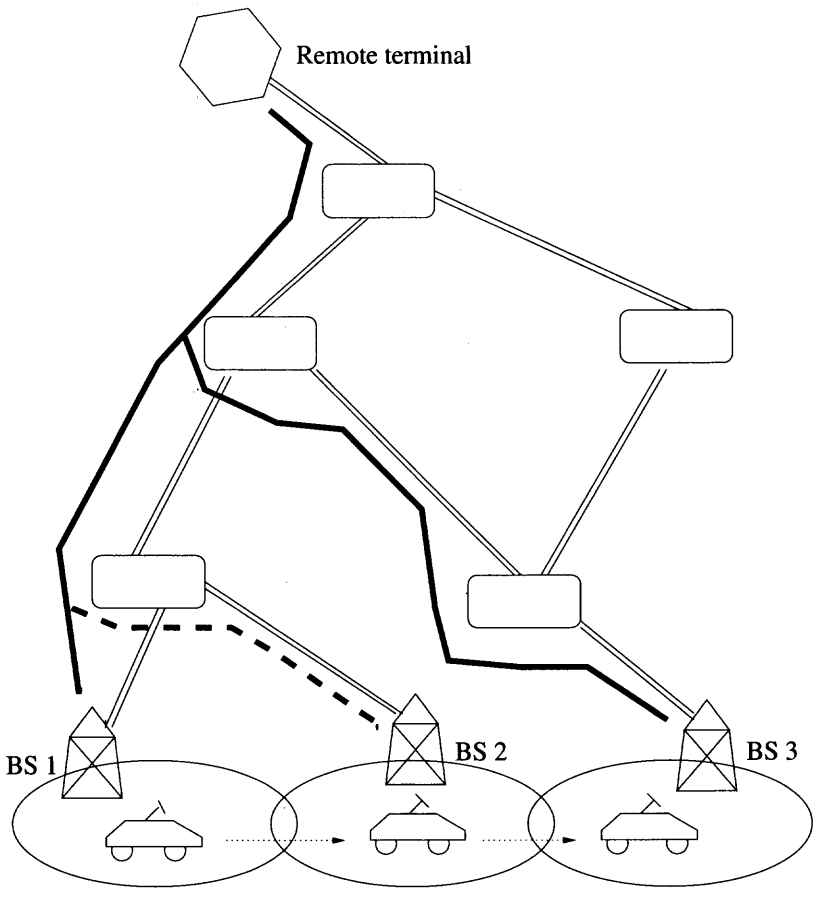

Figure 1.5 The incremental reestablishment case

- *Multicast Establishment*: This approach, which was proposed by Acampora and Naghshineh in [13], preallocates resource in the network portion surrounding the macrocell where the mobile user is located. When a new mobile connection is established, a set of virtual connections named a *virtual connection tree* , is created, reaching all base stations managing the macrocells toward which the mobile might move in the future. Thus, the mobile user can freely roam in the area covered by the tree without involving the network call acceptance capabilities during handoff. This approach is fast and statistically guarantees the QoS contract in case of network handoff. Since the QoS is negotiated only once at connection establishment, resources should be allocated within the

entire area where the mobile is expected to roam. However, this approach may not be efficient in terms of network bandwidth utilization, since it introduces the possibility of refusing a connection because of lack of resources that may never be needed, and high signaling overheads. Fig. 1.6 shows a multicast establishment, assuming that the MT moves within three macrocells.

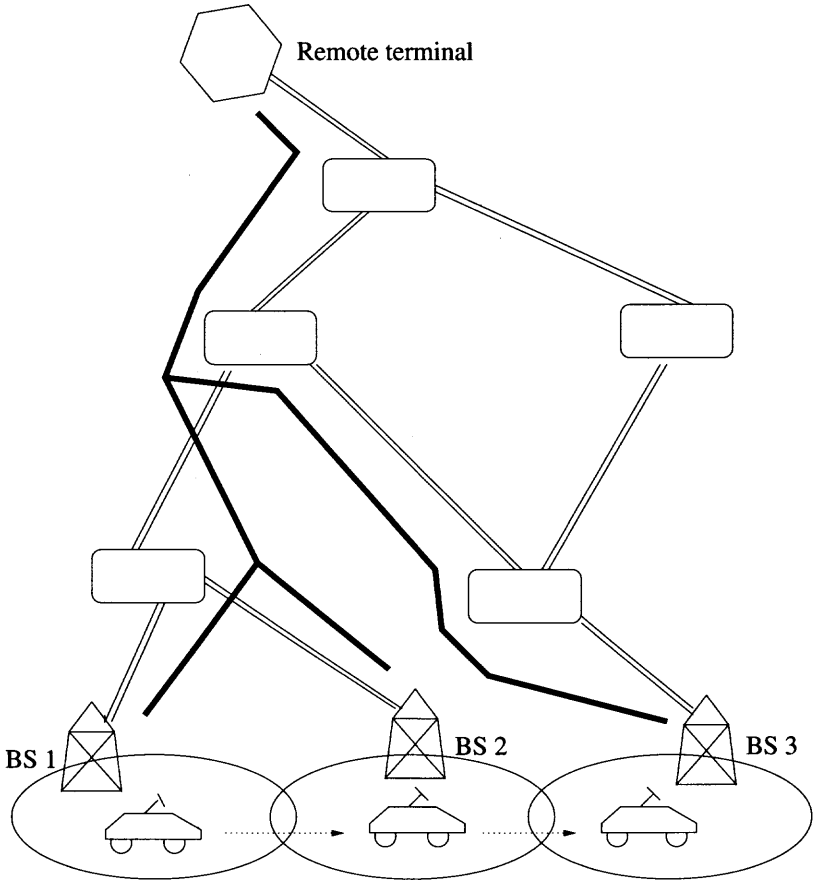

Figure 1.6 The multicast establishment case

1.4 Space Priority Schemes

Since future wireless network is high-speed network with a bandwidth larger than the existing wireless network, the packet loss ratio will be large due to congestion

in the wireless network when a large amount of traffic is transmitted to the mobile terminal. In order to solve this problem, several traffic management schemes are proposed. Especially, buffer control scheme has the advantage of providing QoS guarantee for various types of traffic.

To support multiple classes of traffic, priority mechanisms can be used to control packet loss rate. In this case, when network congestion occurs, different packet loss requirements can be satisfied by selectively discarding packets. Space priority schemes can be used as local congestion control schemes to satisfy different packet loss requirements of different classes of traffics. With a space priority scheme, when congestion is detected, the higher priority is given to loss-sensitive traffic over other traffic, and cells with lower priority are discarded first. Two space priority schemes have been proposed in the literature: partial buffer sharing and push-out scheme.

1.4.1 Partial Buffer Sharing

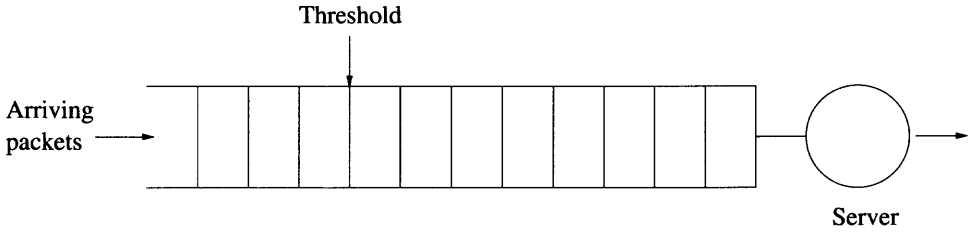

Figure 1.7 Arrangement of a Simple Threshold Based Scheme

Partial buffer sharing uses a threshold to determine whether an arriving packet should be allowed to enter the buffer [14][15][16][17][19][24]. The typical simplest threshold arrangement is to have two levels, high priority or loss sensitive and low priority or loss insensitive, and for the threshold mechanism only to operate on the low priority packets. Thus, when the queue occupancy is above the threshold

only high priority packets are admitted. The motivation behind this arrangement is principally to try and meet the diverse QoS requirements and this is achieved by improving the loss performance of the high priority traffic while degrading the performance of the low priority. This arrangement, which is depicted in Fig. 1.7, assumes that the buffer comprises a single FIFO queue.

1.4.2 Push-Out

The pure push-out policy is a classical space priority mechanism which has widely been discussed in the literature [15][16][19]. In general, the algorithm operates as follows. A shared memory type buffer is usually employed, either as a shared memory switch fabric, or as a shared memory output buffer. Arriving packets typically have two priority levels, namely high and low, and are all stored while there is space in the buffer, which is depicted in Fig. 1.8.

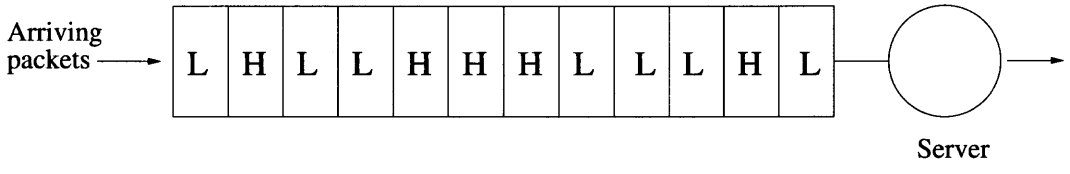

Figure 1.8 Schematic of the Push-out Arrangement

When the buffer becomes full, arriving low priority packets are dropped immediately. High priority packets arriving at a full buffer may be stored by pushing out a packet already stored in the buffer. The decision as to which packet to push out can be either selective or non-selective. With the non-selective scheme an arriving high priority packet simply pushes out the packet at the head of line position. If the buffer consists of several logical queues, the packet pushed out may

not belong to the queue to which the arriving high priority packet joins, typically the packet pushed out is in the longest queue.

In the selective push-out scheme, an arriving high priority packet to a full buffer may pushout a low priority packet stored within the buffer in order that it can be stored at the queue. Again, if the buffer contains several logical queues the packet pushed out may not necessarily be from the queue which the high priority packet is to join. The decision as to which low priority packet in the buffer to pushout can either be:

- the packet nearest the head of line position, or first-in first-dropped (FIFD) or
- the packet nearest the tail of the queue, or last-in-first-dropped (LIFD) or
- a packet chosen at random [20]

An extension to the pure selective push-out scheme is the probabilistic push-out scheme. Here, a high priority packet arriving at a full buffer may pushout a low priority packet with a given probability. While this provides a control parameter which may be adjusted, and may prove easier to do so on-line, compared to say the threshold in partial buffer sharing, decreasing the probability below 1 (equivalent to the selective scheme) only serves to degrade the performance of the high priority packets, and thus may be best suited to where there are several logical queues.

While the selective push-out scheme yields a better overall performance, the non-selective scheme offers the following two advantages compared to a simple FIFO buffer:

- high priority packets experience better loss performance than low priority packets
- the length of the logical queues tend to be equalized, leading to a degree of fairness

In addition, the non-selective scheme is simpler to implement than the selective scheme owing to the fact that a fewer number of pointers are required, leading to a smaller processing overhead.

Compared to partial buffer sharing, the selective push-out scheme offers a number of advantages which include:

- in [18], it was reported that selective push-out offered the best performance results in terms of the packet loss ratio compared to a hybrid non-selective push-out scheme with a global threshold and a simple global threshold scheme. In general, the selective push-out scheme gives very good performance results especially in terms of the low priority packet throughput for a given level of high priority packet performance.

- buffer memory is utilized more efficiently owing to the fact that packets are not discarded until the buffer is full.

The main disadvantage to push-out based schemes, compared to threshold based ones, is that they are complex to implement.

1.4.3 Hybrid scheme

Hybrid schemes which attempt to combine the performance advantage of the push-out scheme with the implementation simplicity of the threshold scheme have been proposed. Here, we give two examples of such schemes. The first [15] uses a FIFO buffer into which both high and low packets are places, an arrangement which is shown in Fig. 1.9, where the arriving high and low priority packets are denoted C_h and C_l, respectively.

The buffer space from the head of the line position to the threshold indicator operates as a normal FIFO buffer, that is, no packets may be pushed out. The buffer space from the point of the threshold indicator operates as a push-out scheme described in the previous section. There is a trade off with this schemes; assuming

that the threshold is said to increase the further away from the head of line position it is, the buffer management complexity reduces as the threshold increases, while the performance increases as the threshold decreases.

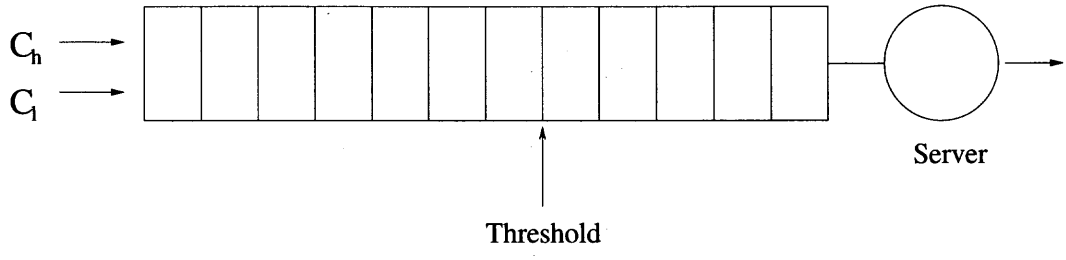

Figure 1.9 Push-out Scheme with Threshold

The second scheme, [21] [22], has an arrangement shown in Fig. 1.10, where high and low priority packets are denoted H and L, respectively. The shared buffer arrangement is used and there are two pointers which denote the ends of the logical high and low priority queues. Arriving packets are stored in accordance with the pointers until the buffer becomes full and the pointers overlap. Under such a circumstance, if the number of high priority packets in the buffer is greater than a the threshold, S, any arriving high priority packet is discarded and any arriving low low priority packet is stored by pushing out a high priority packet. If the number of high priority packets in the buffer is less than the threshold, then the operation is reserved, that is , arriving low priority packets are discarded and arriving high priority packets are stored by pushing out a low priority packet in the buffer.

Since the buffer used is not FIFO a packet scheduling algorithm is required, and the one used is that all buffered high priority packets are served before any buffered low priority ones. Referring to Fig. 1.10, this means that arriving high priority packets have a guaranteed maximum delay through the buffer of S time slots.

Owing to the fact there are only two pointers involved with the operation of this scheme it is simpler to implement than pure selective push-out. The results obtained with this scheme show, as one might expect, that as the threshold increases the loss rate of the low priority class increases, giving the high priority class more priority. However, while the results indeed show that applications using the high priority class would achieve a greater QoS and that the performance of the two classes can be altered through the threshold setting, the nature of the scheme means that the high priority packets have both a superior loss and delay performance and that altering the threshold directly effects both.

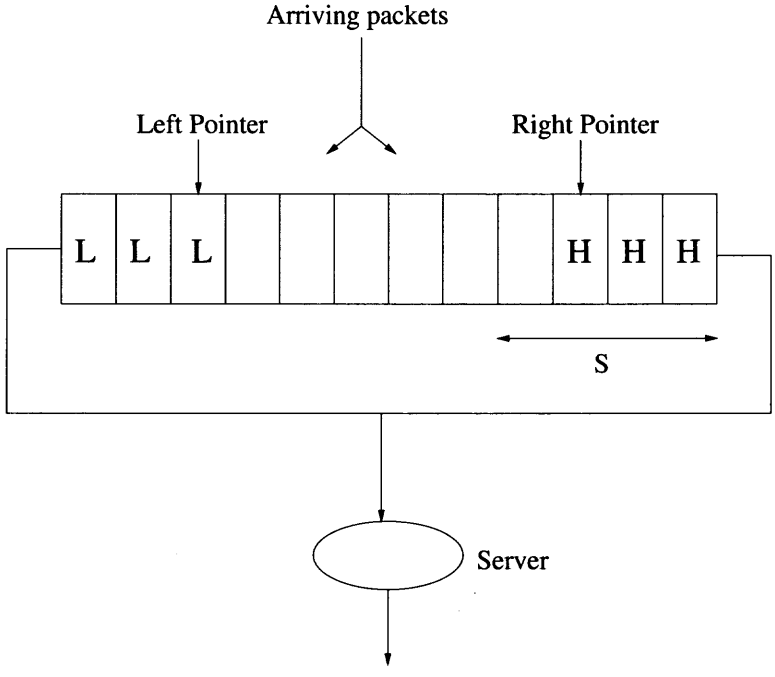

Figure 1.10 Buffer Structure of Push-out Scheme with Threshold

1.5 Statement of the Problem

The user traffic in 3G wireless networks is generated by multimedia or multiple class applications, which are typically bursty. The impressive growth of cellular mobile

telephony as well as the number of Internet users promises an exciting potential for a market that combines both innovations: cellular wireless data services. Within the next few years, there will be an extensive demand for wireless data services. In particular, high-performance wireless Internet access will be requested by users. The ability of packet switching to spread information over the network and to give priority to segments of the same stream makes it possible for QoS adaptation by degrading service, rather than denying service, in overloaded cells. When a cell is overloaded, some packets use adjacent cells or the packets rate is systematically reduced, rather than having service disconnected.

Traffic management is crucial in wireless network. At the BS, packets destined for MTs are transmitted on the forward channels. It is important to gaurantee the QoS for each kind of traffic in the wireless network. Various buffering schemes can be used at the BSs, and packets arriving from a switch will be served in several service disciplines across the BS. Packets from the switch, however, can arrive in bursts with a much higher rate than that being served over the radio link. This fact explains the requirements of buffering at the BSs. Each burst can cause queueing of packets, and is the main cause of packet loss rate.

One of the major issues that must be addressed to enable wireless network guaranteeing a high quality of service is the efficient allocation of bandwidth to the various mobile terminals during handoffs and for new calls. Unlike in wired networks where we can deploy more lines when an increased capacity is required, wireless networks have a fixed capacity due to the limited spectrum availability. This calls for efficient utilization and hence management and allocation of the radio resources. It is important to support as many ongoing calls as possible at any instant while guaranteeing the required QoS.

Wireless Communication Service is expected to provided low-power, high-quality wireless access to the wired network. When a user moves from one cell to

another during a call, a handoff to the new cell is required to maintain the call quality. The forced termination of an ongoing call due to handoff blocking is considered less desired than blocking the initial access of a new call. In a cellular circuit-switched network, several prioritizing schemes have been proposed and studied to reduce the forced termination probability; cut-off prioritization and queueing handoff request. The existing handoff protection schemes were designed primarily for voice, data or mixed form of voice and data, and performance analyses of handoff were obtained by using fixed bandwidth circuit switching. That is, the bandwidth of each connection is equal to that needed to transport a digital voice or data signal, and each connection is given exclusive use of a small portion of the wireless bandwidth for the entire duration of that connection; per-call resource allocation.

While significant work has been done on handoff mechanisms in circuit-switched mobile networks, there is not much literature available on handoff in packet-switched mobile networks. It is critical to handle QoS parameters such as packet loss probability or packet delay, as well as new call blocking and handoff failure probability for analyzing handoff performance, because multimedia or multi-class traffics are characterized by bursty sources. How can the handoff be handled in wireless packet communications?

CHAPTER 2

A NOVEL THROUGHPUT IMPROVEMENT SCHEME FOR MULTI-CLASS SERVICES IN WIRELESS PACKET SYSTEMS

2.1 Introduction

Wireless communication networks have been growing rapidly in recent years. In the wireless network, there are various mobile devices, such as mobile handset, personal digital assistant (PDA) and portable computer, used to transmit voice , video, and data. It implies that there are various services with different transmission rates and qualities in the wireless communication network. As a results, it is important to find out the ways to guarantee the Quality of Service (QoS) for each kind of traffic in the wireless network.

Since the existing wireless networks are bandwidth-limited, it is difficult to support multiple services with different QoS requirements. As QoS guarantee is provided for each kind of traffic in the wired network, a promising solution for integration of multiple services over wireless network should be provided. With the advent of the World Wide Web, the Internet has grown beyond imagination and Enhanced General Packet Radio Services (EGPRS) [3] is developed to support the high data rate traffic in the air.

Several traffic management schemes are proposed in the Wireless ATM network [25][26]. The traffic management scheme proposed herein integrates flow control and buffer management for the downlink traffic (from the source to the mobile terminal) based on TDMA protocol. It can provide the services with real-time constraint and QoS-guarantee for both type of real-time and non-real time traffic.

The focus of this chapter is buffer management. Many buffer management schemes have been proposed, such as partial buffer sharing and selective discarding. In general, these schemes are protective of high priority packets. However, this

performance gain is always achieved only at the cost of a significant performance degradation for low priority packets.

In this chapter, a throughput improvement method is proposed in a Wireless Packet Network. Specially, it is proposed to regulate the buffer by the "Selective-delay Push-in" scheme, which is applicable to scheduling delay tolerant non-real time traffic (NRTT) and delay sensitive real time traffic (RTT). Simulation results show that the performance observed by real time traffic (e.g., voice and video) is improved as compared to the existing partial buffer sharing scheme in term of packet loss probability.

2.2 System Description

EGPRS is one of the proposals submitted to the IMT-2000 initiative of the ITU for third-generation wireless services. EGPRS is also the evolutionary path chosen by the Universal Wireless Communications Consortium, leading toward the convergence of GSM and IS-136 standards for their next-generation wireless systems.

EGPRS permits offering IP-based services such as Internet access in an efficient manner. The network elements are:

- Mobile Terminal (MT), which interfaces to the terminal equipment, and terminates the radio interface,
- Base Station Subsystem(BSS), which constitutes the interface between the network and mobile terminal, and transfers packet and signaling messages between serving GPRS support nodes (SGSNs) and mobile terminal in its coverage area,
- SGSN, a packet switch that routes packets to appropriate mobile terminals within its service area,

- Gateway GPRS support node (GGSN), which acts as the logical interface between the EGPRS network and external packet networks. Its tunnels IP packets from external networks to the SGSN using the GPRS Tunneling Protocol (GTP).

The current phase of EGPRS specifications, which is close to completion, continues to use the GPRS core network and introduces a new air interface, called Enhanced Data rate for GSM Evolution (EDGE), to support higher data rates. This is accomplished mainly by using a higher-level modulation, 8-phase shift keying (8-PSK). With this enhancement the system can provide a data rate over 384 kb/s and spectrum efficiency of 0.5 bps/Hz/base.

2.3 Selective-Delay Push-In (SDPI) Scheme

In order to provide and maintain QoS, the BS is equipped with a buffer manager. If buffer management is assumed to use a single queue approach, arriving packets will be serviced in a first-in-first-out fashion across the BS. Owing to the burstiness of traffic, buffering at the BS is required. Each burst can cause queueing of packets, resulting in Packet Transmission Delay, Packet Delay Variation, and Packet Loss Ratio (PLR).

In general, the traffic can be categorized into two basic classes: real time traffic (RTT) and non-real time traffic (NRTT). RTT has a limitation on the maximum delay time. If an RTT packet is not delivered to its destination within the maximum delay time, it would be dropped. The RTT source may be of voice or video traffic. The NRTT is more tolerant to delay, but has more stringent requirement for packet loss probability. On the scarce wireless bandwidth, to reduce the forced terminations of handoff calls, the delay tolerant and loss tolerant properties of traffic can be exploited at the packet level. Channel utilization can be increased at the expense of QoS degradation such as partial traffic delivery and packet drops in a

buffer. Criteria for such decisions can be based on the application specified quality of multimedia information which the system tries to satisfy. Packets from different mobile terminals are delivered to the buffer of BS by statistical multiplexing and FIFO service discipline.

First, threshold-based discarding scheme is considered, which is called partial buffer sharing scheme. Priority cell discarding is a popular congestion control technique in high-speed networks that allows network resources to be used more efficiently, thereby making it easier to satisfy QoS requirements of different classes of traffics. As shown in Fig. 2.1, the buffer is partitioned by n thresholds, $S_1, \ldots, S_n$, corresponding to n priority classes, where S_n is the buffer size.

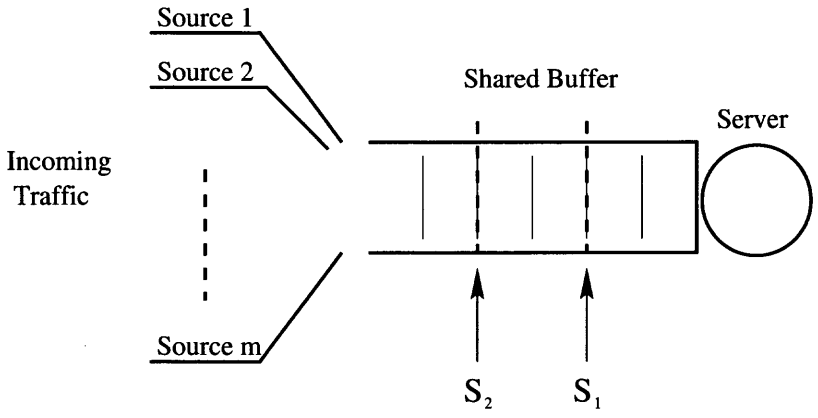

Figure 2.1 The Threshold-Based Discarding scheme

Priority class i cells can be buffered up to threshold level S_i. Once the buffer level exceeds S_i, arriving class i cells are dropped. Note that only new arrivals are dropped; class i cells that are already in the buffer are never dropped and are eventually served. In the case that two kinds of traffics (i.e., real time and non-real time traffic) are considered, non-real time traffic such as data is given priority over real time traffic such as voice and video on this scheme. It is assumed that the buffer

size and the threshold are decided according to the QoS requirement of non-real time traffic (i.e., cell loss probability) and the QoS requirement of real time traffic (i.e., maximum cell delay), respectively. So, real time traffic cells are dropped from a buffer when the buffer level exceed the threshold, decided according to its maximum cell delay.

Second, threshold-based discarding schemeis modified by giving other priority to the real time traffic over non-real time traffic selectively, and thus called selective-delay push-in (SDPI) scheme. With this scheme, non-real time traffic cells can be delayed in favor for real time traffic cells. As illustrated in Fig. 2.2, when the buffer level is less than the threshold, the SDPI scheme operates as like the threshold-based discarding scheme.

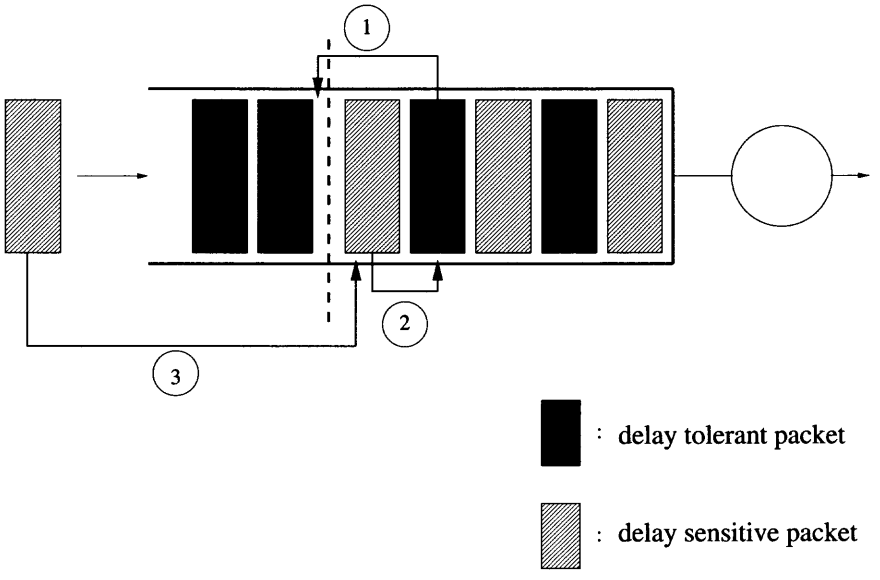

Figure 2.2 The Selective-Delay Push-In scheme

But, when the buffer level is above the threshold, if there exist non-real time traffic cells within the threshold, an arriving real time traffic cell pushes out the latest arrived non-real time traffic cell and positions itself at the end of the buffer

within the threshold. At this moment, the expelled non-real time traffic cell buffers up at the end of the buffer. If no non-real time traffic cell is within the threshold, an arriving real time traffic cell is discarded. When the buffer is full, arriving real time or non-real time traffic cells are just discarded. The threshold is set according to maximum cell delay of real time traffic to satisfy its delay requirement, as like the threshold-based discarding scheme. When the buffer level is above the threshold, if there exist non-real time traffic cells within the threshold, an arriving real time traffic cell is survived in SDPI scheme, but, it is not in threshold-based discarding scheme.

2.4 Traffic Source Models for Multi-class Services

A wireless network is expected to provide a seamless connection to a mobile terminal so that multimedia applications including video, voice and data can be serviced even at the mobile part. For simplification, three kinds of traffics are used, each of which we call as service class 1, 2 and 3, respectively. In addition, since all users that are connected to the same base station are sharing the air interface as the only medium in a wireless environment, resources or channels for wireless transmission are less sufficient than in wired networks.

For the proposed buffer management, three individual traffic source models are considered. First one is the ON-OFF source model [32] for service class 1 as in Fig. 2.3. The ON-OFF source model is commonly used not only for the source which is multiplexed from multiple independent and identical sources but also for the CBR (Constant Bit Rate) traffic source. In the ON state, the source generates packets with a constant bit rate, r_{11}, and does not in the OFF state. p_{11} is the state transition probability from ON state to OFF state, and p_{12} is the reverse probability. The time staying is either state is exponentially distributed. Thus in a steady state, the probability that a source is in either state can be defined as follows:

$$\pi_{11} = \frac{1/p_{11}}{1/p_{11} + 1/p_{12}} = \frac{p_{12}}{p_{11} + p_{12}} \tag{2.1}$$

$$\pi_{12} = \frac{1/p_{12}}{1/p_{11} + 1/p_{12}} = \frac{p_{11}}{p_{11} + p_{12}} \tag{2.2}$$

Thus, the average packet generation rate for this service class is defined by

$$r_1 = r_{11}\pi_{11} + 0 \cdot \pi_{12} = r_{11}\pi_{11} \tag{2.3}$$

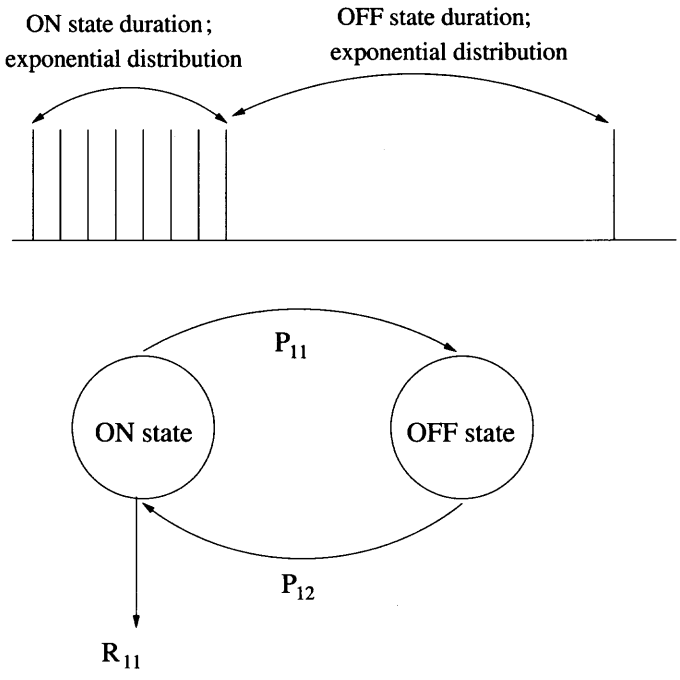

Figure 2.3 ON-OFF source model for service class 1

The second one is the Poisson process model for service class 2, where packets are generated whenever users have any data to transmit.

The third one is the IPP (Interrupted Poisson Process) model for service class 3 as in Fig. 2.4. The IPP model is much similar to the ON-OFF model except that

in state ON, packets are generated by the Poisson distribution with the mean value, r_{31}. For the IPP model, the following parameters are defined.

- r_{31}, r_3: the mean value of packet generation rate for state 1 and overall, respectively. $r_3 = r_{31}\pi_{31}$
- p_{31}, p_{32}: the state transition probability from state 1 to state 2 and from state 2 to state 1, respectively.
- π_{31}, π_{32}: the steady state probability that a source is in state 1 and 2, respectively. $\pi_{31} = p_{32}/(p_{31} + p_{32})$, $\pi_{32} = p_{31}/(p_{31} + p_{32})$

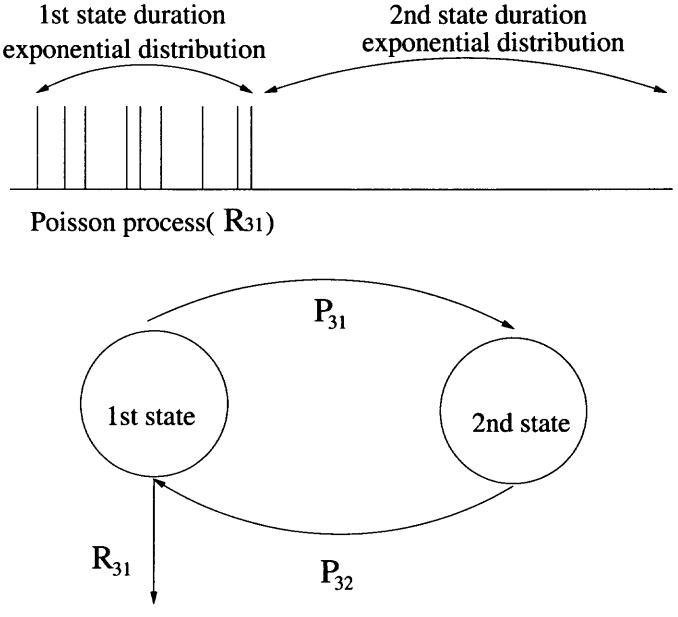

Figure 2.4 IPP source model for service class 3

2.5 Simulation Study

Computer simulations are performed to evaluate the performance of the proposed space priority scheme (i.e., SDPI) in EGPRS network. The key system parameters considered in this chapter are listed in Table 2.1. Class type 1 and 2 have the

maximum allowed transmission delays and can be examples of real time traffic , which is delay sensitive traffic. Class type 3 doesn't allow the packet loss as like non-real time traffic, which is loss sensitive traffic. The threshold of class type 1 and 2 are fixed according to their maximum allowed transmission delay. However, the buffer size is assumed to be infinite, because of the property of loss sensitive and delay tolerant traffic. The fraction of total traffic of each class type is assumed to be 50%, 10%, and 40%. In hopes of improving performance, two schemes are investigated: TBD and SDPI. In the TBD scheme, the priority is given according to loss sensitivity. In the SDPI scheme, the priority is given according to delay sensitivity, based on the TBD scheme.

In Fig. 2.5, packet loss probabilities are plotted as a function of the mean offered load. The thresholds of class type 1 and 2 are 24 and 48, respectively.

Table 2.1 System Parameters of Simulation

Parameter Name	Value
channel capacity	2.4 *Mbps*
packet length	128 *bytes*
average peak rate of class1	32 *Kbps*
average length of ON state for class 1	1.0 (s)
average length of OFF state for class 1	1.35 (s)
maximum allowed packet loss probability for class 1	10^{-2}
maximum allowed transmission delay for class 1	10 *ms*
average packet generation rate for class 2	320 *Kbps*
maximum allowed packet loss probability for class 2	10^{-4}
maximum allowed transmission delay for class 2	20 *ms*
average peak rate of class 3	128 *Kbps*
average length of ON state for class 3	0.2 (s)
average length of OFF state for class 3	1.0 (s)
maximum allowed packet loss probability for class 3	0

These threshold values are based on the maximum allowed transmission delay for each traffic. There is performance improvement for class type 1 and 2 with SDPI

scheme, compared to the TBD scheme. When the buffer occupancy is above the threshold, if there exist non-real time traffic such as class 3 within the threshold, an arriving real time traffic such as class type 1 or 2 is survived at the buffer with SDPI scheme, but, it is not with TBD scheme. At this point, we can get the improvement.

In Fig. 2.6, the throughput is plotted as a function of the mean offered load. Since the performance improvement for packet loss probability is shown in Fig. 2.5, the better throughput can be obtained in SDPI scheme, compared to TBD scheme.

In Fig. 2.7, the effect of the fraction of class 3 on packet loss probability is shown. the offered load is fix at 0.8. The fraction of class 3 increases from 20% to 80%. The ratio of class 1 to class 2 is always 0.8, even though the fraction of class 3 is changed. There is no change in TBD scheme, because, when the threshold is occupied , an arriving packet is discarded, no matter what is within the threshold. However, in SDPI scheme, as the fraction of class 3 increases, packet loss probabilities for class type 1 and 2 decrease. As the amount of class 3 within the threshold increases, an arriving packet of class 1 or class 2 has more chance to see and push out the class 3 packets within the threshold in SDPI scheme, compared to TBD scheme. Therefore, by adjusting the parameter between delay sensitive traffics and delay tolerant traffics without violating the QoS requirement, more efficient channel utilization can be achieved.

In Fig. 2.8, the effect of buffer size on the packet loss probability for each class at offered load 0.8 is shown. The threshold for class 1 and 2 are fixed at 24 and 48, respectively. Buffer size is changed from 60 to 120. In TBD scheme, there is no change for class 1 while buffer size is increased. That is, the packet loss probability for class 2 is decreased gradually, but, it is not much. In SDPI scheme, increasing buffer size does not affect the performance of class 1 much, but, for class 2 the effect is not little. Due to the property of SDPI (that is, the pushed out packet is just delayed, not discarded if the buffer is not full), the buffer is occupied faster than

TBD. Therefore, class 2 is affected more than class 1. Then, the performance of class 3 is better in TBD than in SDPI.

In Fig. 2.9, packet loss probabilities are plotted as a function of number of class 1 users. It is assumed that the number of class 2 and class 3 users are 4 and 12, respectively and the number of class 1 users increases. In the comparison of two schemes, the number of class 1 users satisfying the maximum allowed packet loss probability, 10^{-2}, are 37 in TBD and 49 in SDPI at the fixed number of class 2 and 3 users. It means that 49 class 1 users are supported simultaneously while the packet loss probability less than 1% in the SDPI scheme, but 37 users in the TBD scheme.

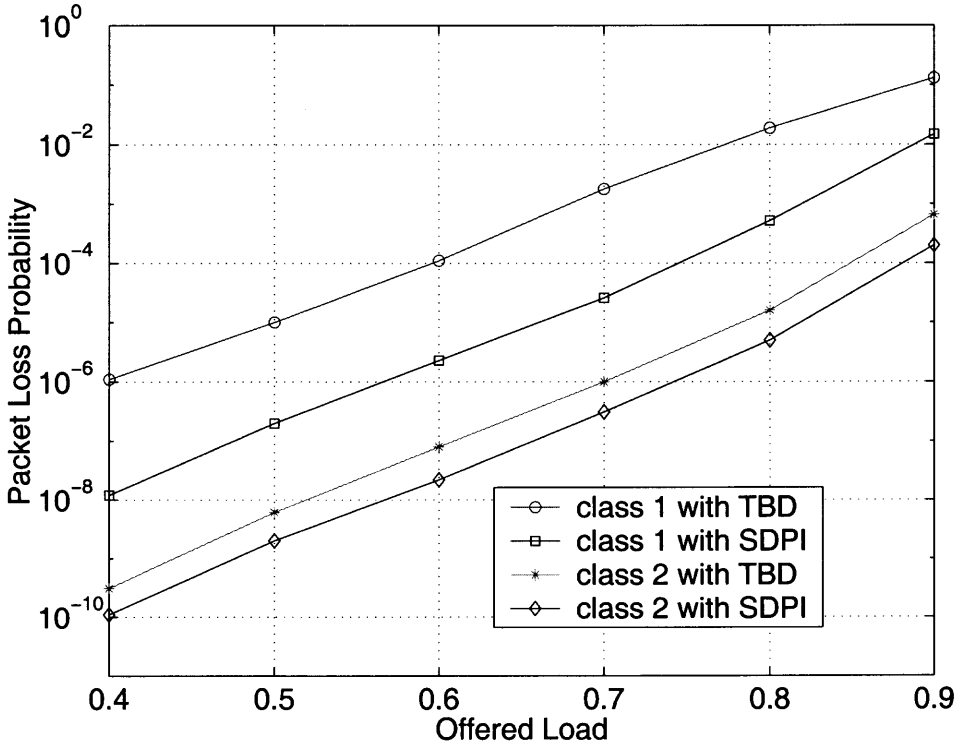

Figure 2.5 Offered load vs. packet loss probability

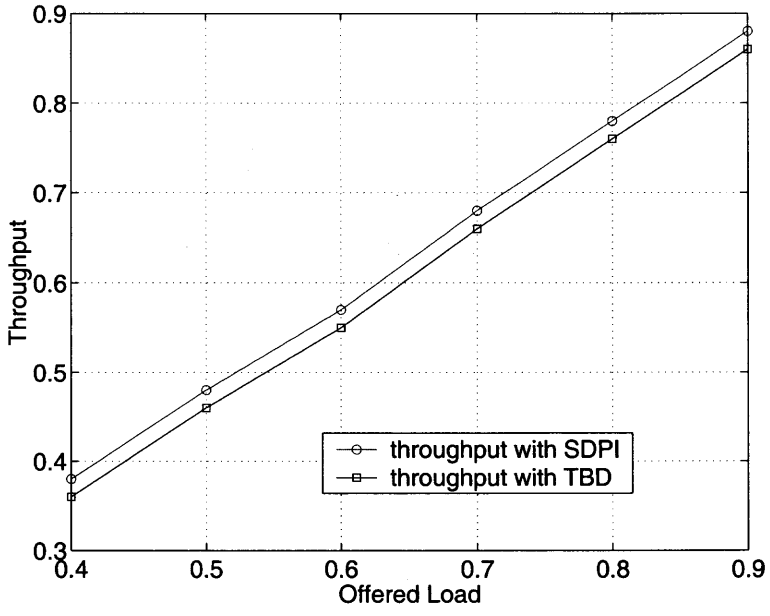

Figure 2.6 Offered load vs. throughput

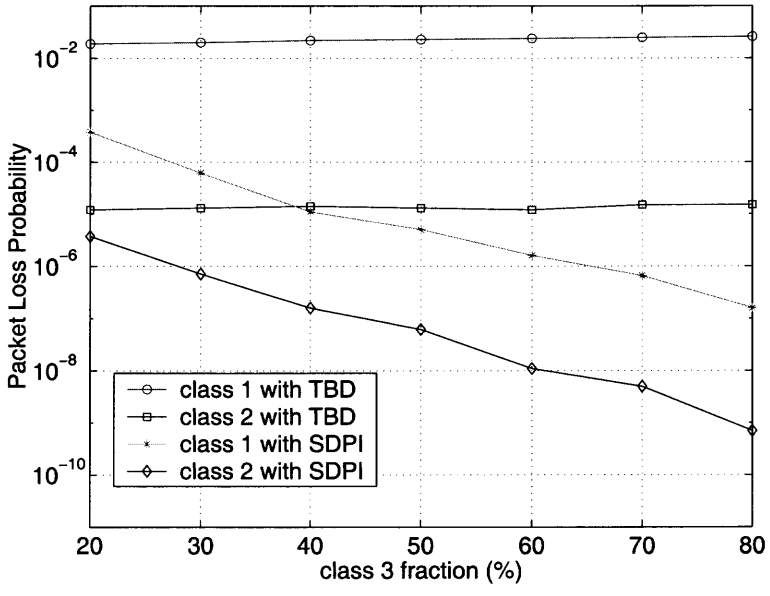

Figure 2.7 Class 3 traffic fraction

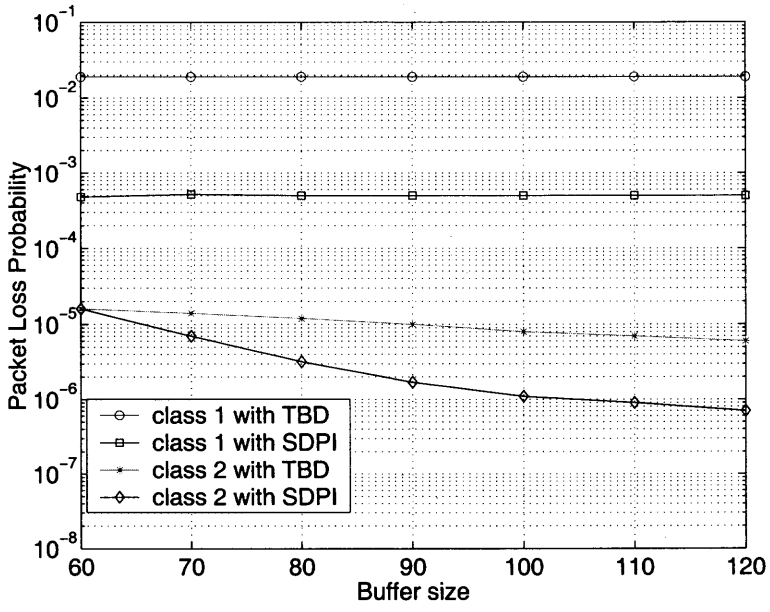

Figure 2.8 Buffer size vs. packet loss probability

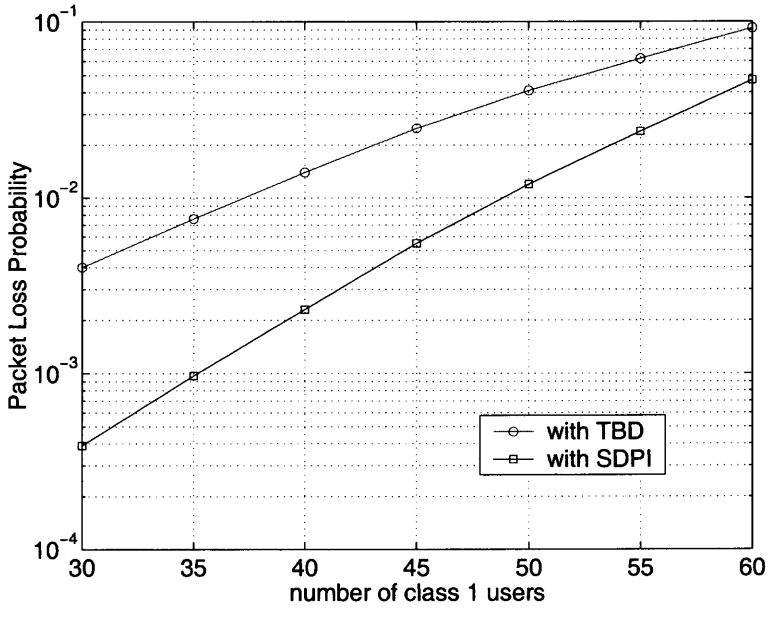

Figure 2.9 Number of class 1 users vs. packet loss probability

CHAPTER 3

PERFORMANCE ANALYSIS OF AN ATM MUX WITH SELECTIVE-DELAY PUSH-IN SCHEME UNDER ON-OFF ARRIVAL PROCESSES

3.1 Introduction

ATM network provides a great variety of services with widely differing bandwidth and QoS requirements. The major characteristics of an ATM-based B-ISDN include: high flexibility of network access, dynamic bandwidth allocation on demand with a fine degree of granularity, flexible bearer capacity allocation, and independence of the means of transmission at the physical layer. However, diverse traffic types and hence different QoS requirements make traffic control of ATM networks an essential and critical challenge. ATM provides the cell transfer for all services, and the ATM adaptation layer (AAL), sitting on top of the ATM layer, provides service-dependent functions to the higher layers. Much research has been concerned with the problem of effectively adapting the quality of the ATM bearer service to the diverse user QoS requirements. One approach to the problem is to support a single ATM cell transfer service by carefully dimensioning the network to satisfy the most demanding QoS requirement imposed. Such an approach leads to poor utilization of network resources and suffers from a lack of flexibility. A more flexible approach is to provide some priority handling mechanism inside the network. The AAL or the end users themselves can make use of this priority handling capacity to derive different QoS while maintaining efficient network use [23].

Several special mechanisms for buffer access have been proposed. They have been used to adapt the cell loss probability of a given class of traffic to the restrictions of the QoS needs of the corresponding service. These mechanisms allow a selective access to the buffer depending on the traffic class. In [15][16][19], the authors proposed a mechanism, called Push-Out, which guarantees the buffer access to a certain class of traffic if the queue is not full, and when it is full, the arriving cell can

replace one with lower priority. The selection of the lowest priority cell to be rejected is done according to the chosen replacement algorithm. Other mechanisms proposed have lower performance but simpler buffer management, called Partial Buffer Sharing [14][15][16][17][19][24], which guarantees the buffer access to a class i cell if the buffer occupancy is less than a threshold, say, S_i. Hence, the highest priority class will be able to access the whole buffer.

The higher bandwidth promised by broadband integrated services digital networks (BISDN) have made applications with real-time constraints possible, such as control, command, and interactive voice and video communications. Excessive delay renders real-time traffic useless, but a certain degree of loss can be tolerated without objectionable degradation in the grade of service. Real-time packets are lost for several reasons. The packet may arrive at the receiver after the end-to-end deadline has expired after having suffered excessive waiting times in the intermediate nodes. Also, intermediate nodes may shed load by dropping packets as an overload control measure. It is natural to engineer communication networks that support real-time traffic, so that delays are bounded at the expense of some loss. However, the magnitude of this loss determines the quality of service and, hence, it is critical to predict this loss accurately in order to provide an acceptable grade of service. Given the fixed length packets and FCFS service at a multiplexer, imposing a buffer size of K is essentially equivalent to imposing a time constraint of Kd, where d is the fixed transmission time of a packet. A broadband network has to guarantee end-to-end delay. The network, in order to meet the delay requirements, forces each node to bound its maximum cell delay.

In this chapter, thorough study of the proposed space priority mechanism is made for the case of bursty traffic. The bursty source is modeled by the Markov Modulated Poisson Process (MMPP), because it is analytically tractable and possesses properties suitable for the approximation of complicated non-renewal

processes. The rest of the paper is organized as follows. Section 2 describes the modeling and analysis of the space priority mechanism; Section 3 presents performance results; finally, some conclusions are drawn in Section 4.

3.2 Threshold-based Priority Scheme

The threshold-based cell discarding is considered. Note that priority cell discarding is a popular congestion control technique in high-speed networks that allows network resources to be used more efficiently, thereby making it easier to satisfy QoS requirements of different classes of traffics. In general, loss-sensitive traffic such as data is given priority over loss-tolerant traffic such as voice and video. RTT ATM cells are dropped from a buffer when the buffer occupancy reaches the threshold. In this work, we consider a simple threshold-based discarding (TBD) scheme. As shown in Fig. 2.1, the buffer is partitioned by n thresholds, $S_1, \ldots, S_n$, corresponding to $n+1$ priority classes. Cells of priority class i can be buffered up to threshold level S_i. Once the buffer level exceeds S_i, arriving cells of class i are dropped. Note that only new arrivals are dropped; class i cells that are already in the buffer are never dropped and are eventually served [16].

A probability vector $\Pi = (\pi_0, \pi_1, \ldots, \pi_S)$ is defined, whose kth component π_k is the probability that a departing packet leaves k packets behind in the system. According to the previous definition of the partial buffer sharing policy, cells of RTT and NRTT are able to join the queueing system if the system state is less than or equal to S_1 and S, respectively. Representing the state transitions of the embedded Markov chain by a transition matrix $\mathbf{Q}$ of size $N \times N$, the following equation system can be stated, describing the stationary characteristics of the system just after a departure instant:

$$\mathbf{\Pi} = \mathbf{\Pi Q} \tag{3.1}$$

Since there is a maximum of one cell served between successive embedded points, transitions from k to level $j < k - 1$ are not possible. Transitions between levels $k \leq S_1$ and $j \leq S_1$ occur with the total arrival rate λ (i.e., $\lambda_1 + \lambda_2$). The corresponding transition probabilities are denoted by the variables $q_1(n)$ given that n arrivals occur between two successive embedded points. The transitions between levels $k > S_1$ and $j > S_1$ depend only on the arrival rate λ_1 of data traffic, since the shared part of the buffer is completely occupied under this condition. These transition probabilities will be denoted by $q_2(n)$, where n describes the number of NRTT cells arriving during one service time. Finally, the remaining transitions consisting of n_1 arrivals with arrival rate λ and n_2 arrivals with arrival rate λ_1 occur with the probability $q_{12}(n_1, n_2)$. Using these notation, the following transition matrix can be established.

The transition probability $q_1(n)$ is given by

$$q_1(n) = \frac{(\lambda \Delta t)^n}{n!} e^{-(\lambda \Delta t)}. \tag{3.2}$$

The transition probability $q_2(n)$ depends on the arrival rate of NRTT

$$q_2(n) = \frac{(\lambda_1 \Delta t)^n}{n!} e^{-(\lambda_1 \Delta t)}. \tag{3.3}$$

For transitions from states $k \leq S_1$ to states $j > S_1$, the arrival rate is reduces from λ to λ_1 when state $S_1 + 1$ is reached because all cells of RTT are discarded in the overload states. The transition probabilities for these transitions can be computed from probability distribution function of the time interval containing n_1 arrivals with arrival rate λ and n_2 arrivals with arrival rate λ_1. Therefore, a different approach is used to derive numerically stable expressions for the transition probabilities. Assuming a constant arrival rate λ during the whole service time. n cells

will arrive with probability $q_1(n)$. After the first n_1 arrivals, each new cell belongs to voice traffic with probability λ_2/λ and will be discarded, because system state $S_1 + 1$ is exceeded. Therefore, the transition probability $q_{12}(n_1, n_2)$ is given by the following equation:

$$q_{12}(n_1, n_2) = \sum_{n=n_1+n_2}^{\infty} q_1(n) \binom{n - n_1}{n_2} (\frac{\lambda_1}{\lambda})^{n_2} (\frac{\lambda_2}{\lambda})^{n-n_1-n_2}. \tag{3.4}$$

The summation can be stopped after a few steps, since the services converges very rapidly. Finally, the probability π_0 is deduced from the probability normalizing condition

$$\sum_{k=0}^{S} \pi_k = 1. \tag{3.5}$$

Steady-state probabilities are of

$$p_k = \begin{cases} \dfrac{\pi_k}{\pi_0 + \lambda\Delta t}, & if \quad 0 \le k \le S_1 \\ \dfrac{\lambda}{\lambda_1} \dfrac{\pi_k}{\pi_0 + \lambda\Delta t}, & if \quad S_1 < k \le S \\ 1 - \dfrac{1}{\pi_0 + \lambda\Delta t}[1 + \dfrac{\lambda_2}{\lambda_1} \sum_{j=S_1+1}^{S} \pi_j], & if \quad k = S + 1 \end{cases} \tag{3.6}$$

The loss probabilities are given as follows:

$$B_1 = p_N \tag{3.7}$$

$$B_2 = \sum_{k=S_1+1}^{N} p_k. \tag{3.8}$$

3.3 SDPI Mechanism

3.3.1 Source Model

The MMPP has been extensively used for modeling arrival rates of point processes because it qualitatively models the time-varying arrival rate and captures some of the important correlations between the interarrival times while still remaining analytically tractable. The accuracy of MMPP in modeling an arrival process depends on which statistics of the actual process are used to determine its parameters. 2-state MMPP models [36][37][38][39] and 4-state MMPP models [40] have been used to approximate the superposition of ON-OFF sources. In [41], the superposition of ON-OFF sources is approximated by means of a 2-state MMPP using the Average Matching Technique. This technique provides good accuracy as compared to simulation results. In particular, the method weakly depends on the number of sources.

At first, assume that the superposition of N independent and homogeneous sources, each characterized by: 1) the peak bit rate, F_p; 2) the activity factor, p; 3) the mean burst length, L_B. With reference to the ATM MUX, denote C as the net output capacity, and thus $M = \lfloor C/F_p \rfloor$ indicates the maximum number of sources that can be accommodated in the MUX, assuming a peak bandwidth assignment. The superposition of N such sources results in a birth-death process. The states of this process are divided into two subsets [38]: 1) an overload (OL) region, comprising the states $M+1,\ldots,N$, where the cell emission rate exceeds the capacity C; 2) an underload (UL) region, consisting of the remaining states $0, \ldots, M$. Therefore, the two states of the approximated MMPP can be chosen so that one of them, called OL state, corresponding to the OL region, and the other, called UL state, associated with the UL region. Let π_j be the limiting probability that the number of active sources is j. Then π_j is given by the binomial distribution.

$$\pi_j = \binom{N}{j} p^j (1-p)^{N-j}$$

where p is the activity factor of a source. Using the average matching procedure, the expression for the four parameters characterizing the MMPP can be determined.

This Average Matching Technique can be adopted for the superposition of independent heterogeneous ON-OFF, consisting of RTT and NRTT. In our case, the finite capacity can be shared by two kinds of traffic. A threshold is defined to separate the two state (Low and High) for each class of traffic. Let N_1 be the set of RTT with peak bit rate, $F_p(1)$, and N_2 be the set of NRTT with peak bit rate, $F_p(2)$. M_1 denotes the threshold which distinguishes the two states (low and high load) for RTT, and similarly, M_2 denotes the threshold which distinguishes the two states (low and high load) for NRTT.

$$M_1 = \left\lfloor \frac{N_1 C}{N_1 F_p(1) + N_2 F_p(2)} \right\rfloor \tag{3.9}$$

$$M_2 = \left\lfloor \frac{N_2 C}{N_1 F_p(1) + N_2 F_p(2)} \right\rfloor \tag{3.10}$$

Thus, two states can be divided for each traffic. That is,

- For RTT

  low load region (Low(1)): [0, 1, ...,M_1]

  high load region (High(1)): [M_1+1, ...,N_1]

- For NRTT

  low load region (Low(2)): [0, 1, ...,M_2]

  high load region (High(2)): [$M_2 + 1$, ...,N_2]

Four parameters are required to represent the 2-state MMPP source of each traffic, as shown in Fig. 3.1, where $\gamma_{L1}(\gamma_{H1})$ is defined as the mean transition rate out of the Low load (High load) state, and $\lambda_{L1}(\lambda_{H1})$ is the mean arrival rate of the Poisson process in the Low load (High load) state for RTT, respectively. Similarly, $\gamma_{L2}(\gamma_{H2})$ is defined as the mean transition rate out of the Low load (High load) state, and $\lambda_{L2}(\lambda_{H2})$ is the mean arrival rate of the Poisson process in the Low load (High load) state for NRTT, respectively.

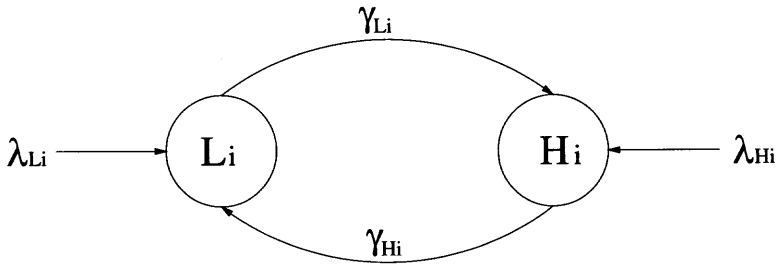

Figure 3.1 2-state MMPP models for RTT and NRTT

3.3.2 SDPI Analysis

The multiplexer is modeled as a finite capacity single server queue where the arrival process is MMPP, and the service is deterministic. In our analysis, the similar assumptions are made as in [40], which deals with the analysis of only one traffic type, that significantly reduce the computational complexity involved in obtaining the stationary distributions at departure points: 1) the probability that the MMPP goes through multiple state transitions between successive departures is negligible, and 2) the state transitions occur at departure epochs, i.e., if a departure leaves the MMPP in state i, the cell arrival rate until the next departure is λ_i. Consider a queue

using SDPI where the MMPP consists of K states denoted by i ($0 \leq i \leq K$-1), and the arrival rates and mean state durations are denoted by λ_i and μ_i, respectively. The characteristics of this system will be determined using an imbedded Markov chain approach. As in the ordinary M/G/1 queueing system, the service completion instants are the imbedded points of the underlying Markov chain. Therefore, a probability vector $\mathbf{\Pi}$ consists of $\pi_i(n_1, n_2)$ ($0 \leq n_1 \leq S_1, 0 \leq n_2 \leq S_2$, where S_2 is the total buffer size) which is defined by the probability that a departing cell leaves n_1 RTT cells and n_2 NRTT cells in the system while the MMPP is in state i. The total transition probability matrix of the imbedded Markov chain, denoted by $\mathbf{Q}$, is formed with K MMPP finite states and F finite buffer states. For example, consider the traffic shown in Fig. 1, where the RTT and NRTT can be aggregated resulting in a 4-state MMPP process (in this case, K=4). The K=4 states are $\{(L_1,L_2), (L_1,H_2), (H_1,L_2), (H_1,H_2)\}$. For a buffer with S_1=3 and S_2=6, there are F=22 finite buffer states corresponding to $\{\{n_1, n_2\} \mid n_1 + n_2 \leq 6 \text{ and } n_1 \leq 3\}$. Thus,

$$\mathbf{Q} = \begin{bmatrix} Q_{0,0} & Q_{0,1} & \cdots & Q_{0,K-1} \\ Q_{1,0} & Q_{1,1} & \cdots & Q_{1,K-1} \\ \vdots & \vdots & \cdots & \vdots \\ Q_{K-1,0} & Q_{K-1,1} & \cdots & Q_{K-1,K-1} \end{bmatrix}, \tag{3.11}$$

where $Q_{j,i}$ is a submatrix, and each element of the submatrix, $Q_{j,i}((n_1, n_2), (n_1', n_2'))$ ($0 \leq j, i \leq K$-1, $0 \leq n_1, n_1' \leq S_1$, $0 \leq n_2, n_2' \leq S_2$) corresponds to a state transition probability. That is,

$$Q_{j,i}((n_1, n_2),\ (n_1', n_2')) = P\{(n_1', n_2'), j \mid (n_1, n_2), i\},$$

where i is the present MMPP state, j is the next MMPP state, (n_1, n_2) is the present buffer state, and (n_1', n_2') is the next buffer state. The submatrix $Q_{j,i}$ can be obtained as follows. Denote A_i as the buffer state transition probability matrix

of the departure point of our system at MMPP state i (with arrival rate λ_i and service time Δt). The transition probability submatrix $Q_{j,i}$ can be simply obtained by multiplying A_i by the probability that the MMPP will not change its state in Δt if $j = i$, or by the probability that the MMPP will change its state from j to i in Δt if $j \neq i$. Define $q_i(k,l)$ as the transition probability that k RTT cells and l NRTT cells can be positioned in the buffer during the service time (Δt) while the MMPP is in state i. Denote $q_i^1(k)$ as the probability of k arrivals of traffic type 1 (i.e., RTT) and $q_i^2(l)$ as the probability of l arrivals of traffic type 2 (i.e., NRTT) during the service time, respectively. Define $q_i^*(k,l)$ as the transition probability that more than k RTT cells and more than l NRTT cells are inserted to the buffer, but only k RTT cells and only l NRTT cells can be positioned in the buffer during the service time (Δt) due to the SDPI mechanism. Thus,

$$q_i(k,l) = q_i^1(k) q_i^2(l),$$

where

$$q_i^1(k) = \frac{(\lambda_i^1 \Delta t)^k}{k!} e^{-(\lambda_i^1 \Delta t)},$$

$$q_i^2(l) = \frac{(\lambda_i^2 \Delta t)^l}{l!} e^{-(\lambda_i^2 \Delta t)},$$

and λ_i^1, λ_i^2 are the arrival rates for traffic type 1 and 2, respectively, and $\lambda_i = \lambda_i^1 + \lambda_i^2$.

Since at most one cell is served between successive imbedded points, transitions from n_1 to $n_1' < n_1 - 1$, from n_2 to $n_2' < n_2 - 1$, and from $n_1 + n_2$ to $n_1' + n_2' < n_1 + n_2 - 1$ are not possible.

Transitions to $n_1' + n_2' < S_2$ and $n_1' < S_1$:

$$q_i(k,l) = q_i^1(k) q_i^2(l) \tag{3.12}$$

Transitions to boundaries:

1. at $n_1' + n_2' < S_2$ and $n_1' = S_1$,

$$q_i^*(k,l) = \sum_{n=k}^{\infty} q_i^1(n) q_i^2(l) \tag{3.13}$$

2. at $n_1' + n_2' = S_2$ and $n_1' = S_1$,

$$q_i^*(k,l) = \sum_{n=k}^{\infty} q_i^1(n) q_i^2(l) + \sum_{n=k}^{\infty} \sum_{m=l+1}^{\infty} q_i^1(n) q_i^2(m) \frac{\binom{n}{k}\binom{m}{l}}{\binom{n+m}{n}} \tag{3.14}$$

3. at $n_1' + n_2' = S_2$ and $n_1' < S_1$,

$$q_i^*(k,l) = \sum_{n=k}^{\infty} \sum_{m=l}^{\infty} q_i^1(n) q_i^2(m) \frac{\binom{n}{k}\binom{m}{l}}{\binom{n+m}{n}} \tag{3.15}$$

The transition probabilities (3.12) denoted by $q_i(k,l)$ implies that exactly k arrivals of traffic type 1 and exactly l arrivals of traffic type 2 occur in any order during the service time. The transition probabilities (3.13) imply that more than k arrivals of traffic type 1 and exactly l arrivals of traffic type 2 occur in any order during the service time. Since the present state $n_1' = S_1$, even though there are more than k arrivals of traffic type 1, only k cells can be positioned in the buffer. According to the SDPI mechanism, an arriving cell is dropped when the buffer is full. Thus, the transition probabilities (3.14) consist of two terms. The first term represents that more than k arrivals of traffic type 1 and exactly l arrivals of traffic type 2 occur. The second term means that more than k arrivals of traffic type 1 and more than l arrivals of traffic type 2 occur. The fraction in the second term represents the probability that k out of n traffic type 1 and l out of m traffic type 2 are the first arrivals. The transition probabilities (3.15) represent that more than k

arrivals of traffic type 1 and more than l arrivals of traffic type 2 occur, as like the second term of the probabilities (3.14).

Define the stationary probability vector $\mathbf{\Pi}$ as

$$\begin{aligned}\mathbf{\Pi} &= \{\pi_0(0,0),...,\pi_0(S_1,S_2-S_1),\pi_1(0,0),...,\pi_1(S_1,S_2-S_1),\\ &\quad ...,\pi_{K-1}(0,0),...,\pi_{K-1}(S_1,S_2-S_1)\}\end{aligned}$$

Then, these stationary probabilities can be obtained as follows:

$$\mathbf{\Pi} = \mathbf{\Pi Q}, \qquad \sum_{i=0}^{K-1}\sum_{n_1}\sum_{n_2}\pi_i(n_1,n_2) = 1.$$

To derive the loss probabilities, it is necessary to determine the probability distribution of the system length (n_1+n_2+1, including the server) from the arrival viewpoint, which is equivalent to the steady-state probability distribution $p_i(n_1,n_2)$ [42]. The probabilities must be different from the former departure-point probabilities $\pi_i(n_1,n_2)$, because the state space is enlarged by the state $G = S_2+1$, where the "1" accounts for the server. Asymptotically, the number of arriving ATM cells equals the number of departing cells. Hence, the departure rate must be equal to the effective arrival rate of ATM cells which are able to join the system.

$$\begin{aligned}&\frac{1-p_i(0,0)}{\Delta t}\\ &= \lambda_i^2\{1-\sum_{n_1+n_2=G}\sum p_i(n_1,n_2)\}+\lambda_i^1\{1-\sum_{n_2=0}^{S_2-S_1} p_i(S_1+1,n_2)-\sum_{n_2=0}^{S_2-S_1} p_i(S_1,n_2+1)\frac{1}{S_1+1}\}\end{aligned} \tag{3.16}$$

where $p_i(n_1,n_2)$ is the steady state probability that an arriving cell sees n_1 RTT cells and n_2 NRTT cells in the system while the MMPP is in state i (i.e., from an arrival point of view). $\frac{1}{S_1+1}$ is the probability that the non-real time traffic cell is

being served, when $S_1 + 1$ cells (i.e., S_1 real time cells and 1 non-real time cell) are within the threshold including the server).

In general, the arrival point queue length distribution of a single server queue is identical to the departure point queue length distribution, given that arrivals and departures occur singly, i.e., $\pi_i(n_1, n_2)$ is the state probability seen by a cell who joins the queueing system [43][44]. Therefore, the following equation holds for the state probabilities just after a departure.

$$\pi_i(n_1, n_2) =$$

$$\begin{cases} \dfrac{p_i(n_1, n_2)}{1 - \dfrac{\lambda_i^2}{\lambda_i} \sum_{n_1 + n_2 = G} \sum p_i(n_1, n_2) - \dfrac{\lambda_i^1}{\lambda_i} \{ \sum_{n_2=0}^{S_2 - S_1} p_i(S_1 + 1, n_2) + \sum_{n_2=0}^{S_2 - S_1} p_i(S_1, n_2 + 1) \dfrac{1}{S_1 + 1} \}}, \\ \quad \text{for } n_1 + n_2 \leq S_1 \text{ or } n_1 < S_1 \text{ and } n_1 + n_2 \leq S_2 \\ \\ \dfrac{\dfrac{\lambda_i^2}{\lambda_i} p_i(n_1, n_2)}{1 - \dfrac{\lambda_i^2}{\lambda_i} \sum_{n_1 + n_2 = G} \sum p_i(n_1, n_2) - \dfrac{\lambda_i^1}{\lambda_i} \{ \sum_{n_2=0}^{S_2 - S_1} p_i(S_1 + 1, n_2) + \sum_{n_2=0}^{S_2 - S_1} p_i(S_1, n_2 + 1) \dfrac{1}{S_1 + 1} \}}, \\ \quad \text{for } n_1 = S_1 \text{ and } n_1 + n_2 \leq S_2 \end{cases} \tag{3.17}$$

The following steady-state probabilities can be obtained by combining (3.16) and (3.17)

$$p_i(n_1,n_2) = \begin{cases} \dfrac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t}, & for \quad n_1+n_2 \le S_1 \ or \ n_1 < S_1 \ and \ n_1+n_2 \le S_2 \\ \dfrac{\lambda_i}{\lambda_i^2}\dfrac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t}, & for \quad n_1 = S_1 \ and \ n_1+n_2 \le S_2 \\ 1-\sum\limits_{\{n_1,n_2\}\in B_1}\sum \dfrac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t} - \sum\limits_{\{n_1,n_2\}\in B_2}\sum \dfrac{\lambda_i}{\lambda_i^2}\dfrac{\pi_i(n_1,n_2)}{\pi_i(0,0)+\lambda_i\Delta t}, & \\ \quad for \ n_1+n_2 = G, & \\ \quad where \ B_1 = \{n_1,n_2 \,|\, n_1+n_2 \le S_1 \ or \ n_1 < S_1 \ and \ n_1+n_2 \le S_2\}, & \\ \qquad\quad B_2 = \{n_1,n_2 \,|\, n_1 = S_1 \ and \ n_1+n_2 \le S_2\} & \end{cases} \tag{3.18}$$

The cell loss probabilities are then given as follows:

a) CLP for NRTT

$$CLP_{NRTT} = \sum_{n_1+\ n_2=G}\sum p(n_1,n_2). \tag{3.19}$$

b) CLP for RTT

$$CLP_{RTT} = \sum_{n_2=0}^{S_2-S_1-1} p(n_1=S_1+1,n_2) + \sum_{n_2=0}^{S_2-S_1-1} p(n_1=S_1,n_2+1)\frac{1}{S_1+1} + CLP_{NRTT}. \tag{3.20}$$

3.4 Results and Discussion

The performance of the SDPI scheme is evaluated for two kinds of traffics. source parameters are chosen which are characterized by the peak bit rate F_p, the activity factor p, and the mean burst length L_B. Assume that the superposition of such

heterogeneous ON-OFF source are offered to an ATM MUX with the net output link capacity C. The performance of the MUX is evaluated by the queueing model with MMPP source and the SDPI priority scheme. The constant service time of the MUX is given by θ=53 bytes/C. The net link capacity is assumed to be 150Mbps.

Some simulation results are reported to evaluate the accuracy of cell loss probability by using the SDPI scheme. The simulations have been performed on SUN SparcStation 60. The source parameters used in our simulations and numerical analysis, which are the same as in [45], are tabulated in Table 3.1. These source parameters are used for each user.

Table 3.1 System Parameters

class	F_p	p	L_B
real time traffic	32Kbps	0.35	1400
non-real time traffic	128Kbps	0.1	1600

In Fig. 3.2, cell loss probabilities are plotted as a function of the mean offered load (real time traffic and non-real time traffic). Note that the simulation results are sufficient reliable, since the 95% confidence intervals range within 10% of the estimated cell loss probability. The threshold and buffer size are assumed to be 10 and 30, respectively. In Fig. 3.3, the comparison between SDPI and threshold-based discarding scheme is shown. It is intuitive to see that SDPI achieves the performance improvement for real time traffic (which is more critical) at the expense of non-real time traffic. As it is mentioned before, when the occupancy is above the threshold, if there exist non-real time traffic cells within the threshold, an arriving real time traffic cell is survived in SDPI scheme, but, it is not in the threshold-based discarding scheme. At this point, there is the improvement for real time traffic with SDPI scheme; that is, the SDPI scheme compensates for the disadvantage for real time traffic of threshold-based discarding scheme, under the circumstance that the threshold is fixed due to the maximum cell delay of real time traffic.

In Fig. 3.4, the cell loss probabilities as a function of real time traffic offered load with a fixed total offered load at 0.9 are shown. There is the improvement for real time traffic with SDPI, compared to the threshold-based discarding scheme, as like Fig. 3.3. As the real time traffic offered load increases, there is no improvement for real time traffic with SDPI at threshold 10 and buffer size 40. As real time traffic is increased and non-real time traffic decreases, the possibility that non-real time traffic is within the threshold decreases and the possibility that arriving real time traffic cells are dropped when the buffer level exceed the threshold increases. In Fig. 3.5, the cell loss probabilities are plotted against the offered load of non-real time traffic. The offered load of real time traffic is fixed at 0.3. As the offered load of non-real time traffic increases, performance for real time traffic in SDPI scheme is getting better, but, performance for non-real time traffic is worse constantly, compared to threshold-based discarding scheme, due to the same reason as in Fig. 3.4.

In Fig. 3.6, cell loss probabilities are plotted as a function of the buffer size. As the buffer size increases while holding the threshold fixed, cell loss probabilities for real time traffic remain constant, but cell loss probabilities for non-real time traffic decreased. Thus, SDPI outperforms threshold-based discarding scheme for accommodating real time traffic, and SDPI may reach comparable performance as threshold-based discarding scheme for accommodating non-real time traffic by increasing the buffer size at the fixed threshold due to the maximum cell delay of real time traffic. In Fig. 3.7, the effects of traffic characteristics on the individual cell loss probabilities are shown. As the activity for non-real time traffic changes, cell loss probability for each traffic is affected. In Fig. 3.8, the cell loss probabilities are plotted as the thresholds are changed. Cell loss probabilities for non-real time traffic is almost not changed, but cell loss probabilities for real time traffic increases as the threshold reaches the buffer size.

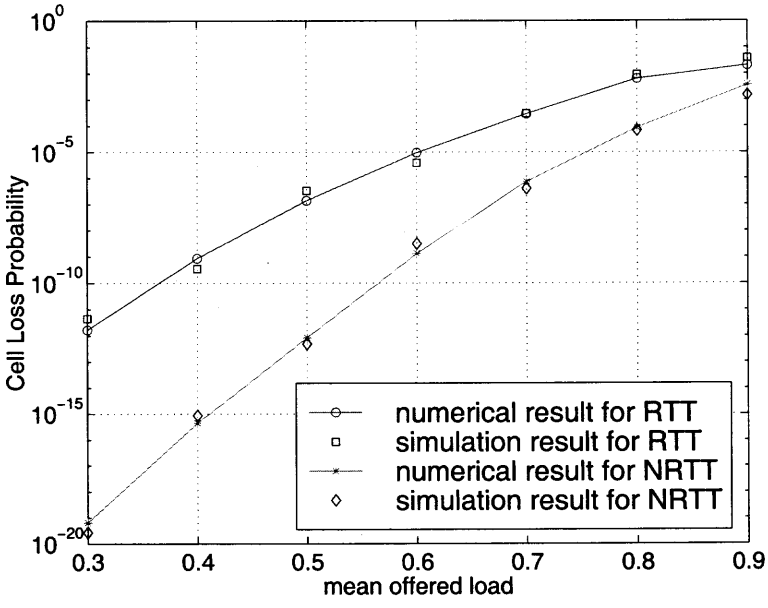

Figure 3.2 Cell loss probability versus mean offered load (comparison among simulation and analytical approaches)

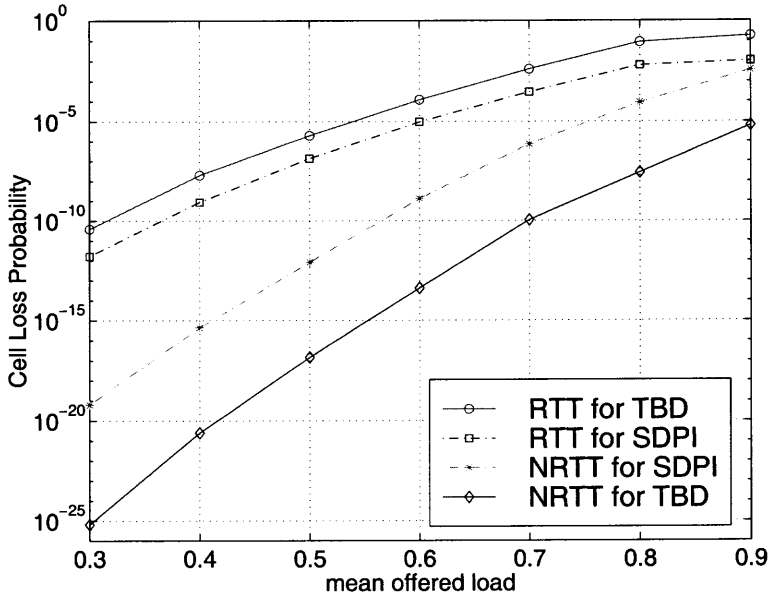

Figure 3.3 Cell loss probability versus mean offered load (comparison among SDPI and TBD scheme)

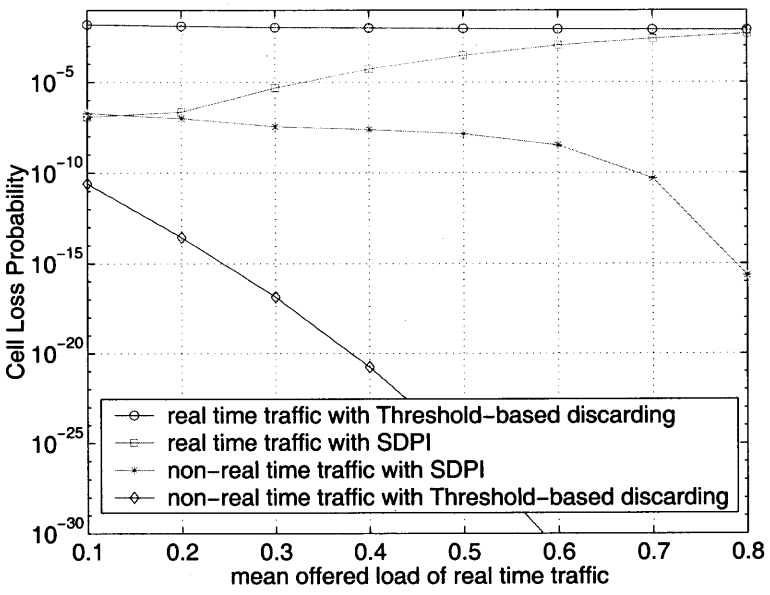

Figure 3.4 Cell loss probability versus mean offered load of real time traffic (comparison between threshold-based discarding scheme and SDPI scheme) (fixed total offered load=0.9, threshold=10, buffer size=40)

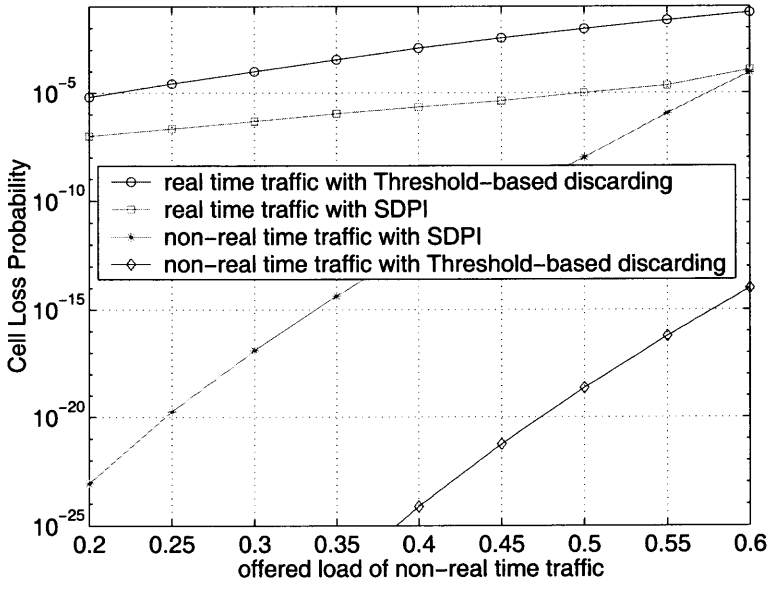

Figure 3.5 Cell loss probability versus mean offered load of non-real time traffic (comparison between threshold-based discarding scheme and SDPI scheme) (offered load of real time traffic is fixed at 0.3, threshold=10, buffer size=40)

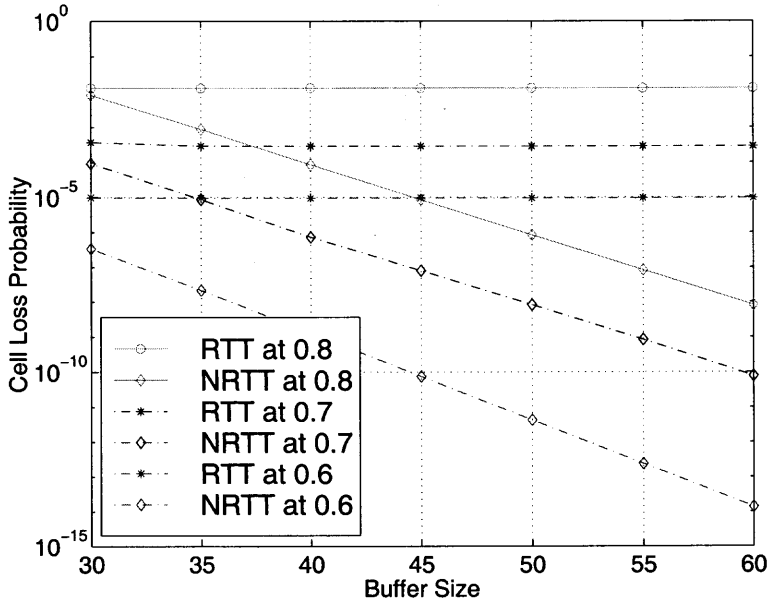

Figure 3.6 Cell loss probability versus buffer size: threshold is fixed (20)

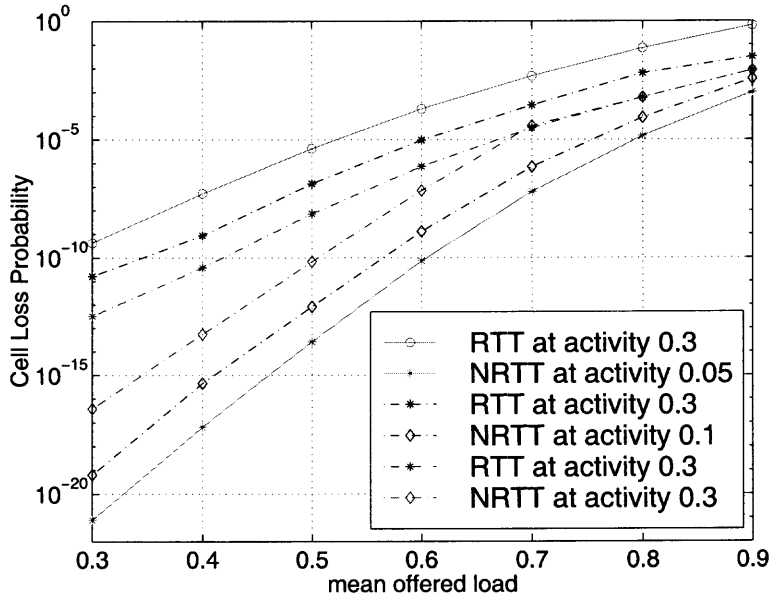

Figure 3.7 Cell loss probability versus mean offered load: different data activity

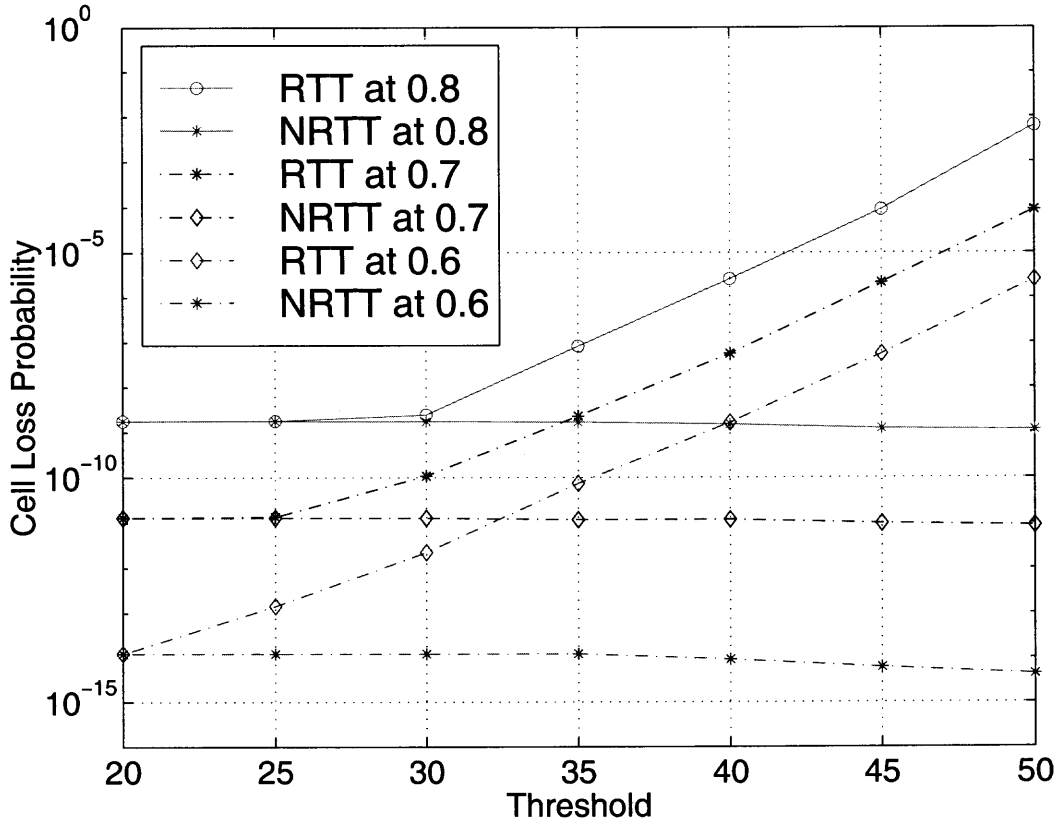

Figure 3.8 Cell loss probability versus threshold: buffer size is fixed (60)

CHAPTER 4

A NOVEL CELL SWITCHING MANAGEMENT SCHEME FOR WIRELESS PACKET COMMUNICATIONS

4.1 Introduction

The user traffic in high-speed wireless networks is generated by multimedia, or multi-class applications, which are typically bursty. The characteristics of packet switched networks can reduce the work required for handoffs relative to circuit switched networks. That is, packet addresses are used to reduce the work associated with changing cells by routing individual packets rather than setting up and tearing down circuits. The ability of packet switching to spread information over the network and give priority to segment of the same stream makes it possible for QoS adaptation by degrading service, rather than denying service, in overloaded cells. When a cell is overloaded, some packets use adjacent cells or the packets rate is systematically reduced, rather than disconnecting sources.

The impressive growth of cellular mobile telephony as well as the number of Internet users promises an exciting potential for a market that combines both innovations: cellular wireless data services. Within the next few years, there will be an extensive demand for wireless data services. In particular, high-performance wireless Internet access will be requested by users. The GPRS is a new bearer service for GSM that greatly improves and simplifies wireless access to packet data networks, e.g., to the Internet. GPRS, EDGE (Enhanced Data rate for GSM Evolution), and UMTS (Universal Mobile Telecommunications Services) are all being developed to accommodate data users in wireless networks. EGPRS/EDGE will evolve to third generation (3G) mobile communications while UMTS will make resolution way for third generation mobile communications [49][50].

Traffic management is crucial in wireless networks. At the base station (BS), packets destined for mobile terminals (MTs) are transmitted on the forward channels.

It is important to find out ways to guarantee the QoS for each kind of traffic in the wireless network. It is non-trivial to support multi-class services with different QoS requirements under limited bandwidth. In order to provide and maintain QoS, the wireless equipment must be equipped with packet buffer. Various buffering schemes can be used at the BSs and packets arriving from the switch will be serviced in several service disciplines across the BS. The maximum radio link throughput is limited and can be expected to be lower than the servicing wired link throughput from the switch. The actual bandwidth allocation must be set according to the wireless link. Packets from the switch however, can arrive in bursts with a much higher rate than that being serviced over the radio link. This fact explains the requirements of buffering at the BSs. Each burst can cause queueing of packets, and is the main cause of packet loss rate (caused by buffer overflow). General buffer management schemes for congestion control have been proposed; e.g., partial buffer sharing and push-out [46][47][48]. The well-known partial buffer sharing scheme is used in this paper, because different priorities must be given to multi-class traffics.

The focus of this chapter is resource allocation for cell switching at BS to satisfy QoS requirements. The QoS requirements are expressed in terms of the packet loss probability and average packet delay for mobile connection. The Enhanced General Packet Radio Service (EGPRS) network [50] which is a TDMA-based approach is considered. A technique to define the *new packets* and *handoff packets* for each type of class is proposed to give the priority at the buffer of BS; there are two classes of packets for each traffic type. The method to examine the effect of packet priority scheme at cell switching (e.g., packet tagging and partial buffer sharing scheme) is proposed. Using the MMPP model for the aggregate ON-OFF traffic streams, the packet loss probability and the average packet delay are computed. The performance of proposed scheme is evaluated by simulation and numerical analysis.

The procedure to find the optimal buffer thresholds that simultaneously satisfy the QoS requirements for multiple types of classes is presented.

4.2 System Description

EGPRS is one of the proposals submitted to the IMT-2000 initiative of the ITU for third-generation wireless services. It uses a TDMA-based packet-switched radio technology and an evolved, packet-switched GPRS core network. The architecture enables the network to provide various packet access services for real-time traffic such as voice and video or non-real-time traffic such as interactive or World Wide Web and related Internet applications. It allows statistical multiplexing of traffic and sharing of physical resources by many users to improve utilization.

There are N types of traffic, labeled $n = 1, 2, 3, ..., N$. Different traffic types are defined by their QoS requirements. It is assumed to be identically distributed packet sizes. Each type of traffic has two priority classes; new packet and handoff packet. These have different priorities at the buffer. There are $I = 2N$ of total priority classes. The QoS requirements for each priority class are assumed to be packet loss probability, PLP_i, and average packet delay, D_i.

Hard handoff is assumed. Handoff decision is based on received power level.

4.3 Packet Tagging

Each cell, served by a BS, is divided into two zones based on the thresholding the received power from the MT at the BS; zone A_1 and zone A_2. The MT in zone A_1 starts to communicate by sending new call request and sends the handoff call request to the BS when it is across zone A_2 during communication. Therefore, zone A_1 is the area for new call generation and zone A_2 is the area for handoff call generation based on received signal strength. The basic traffic model assumes that the new call origination rate and handoff call rate are uniformly distributed over zone A_1 and

zone A_2, respectively. The average rate of new origination in zone A_1 by Λ_n and the average rate of handoff in zone A_2 by Λ_h are denoted. Here, call arrivals are assumed to be Poissonian.

Note that our scheme can easily be generalized to irregular geometric layout, determined solely by received power level. A circular cell representation for ease of demonstration is considered.

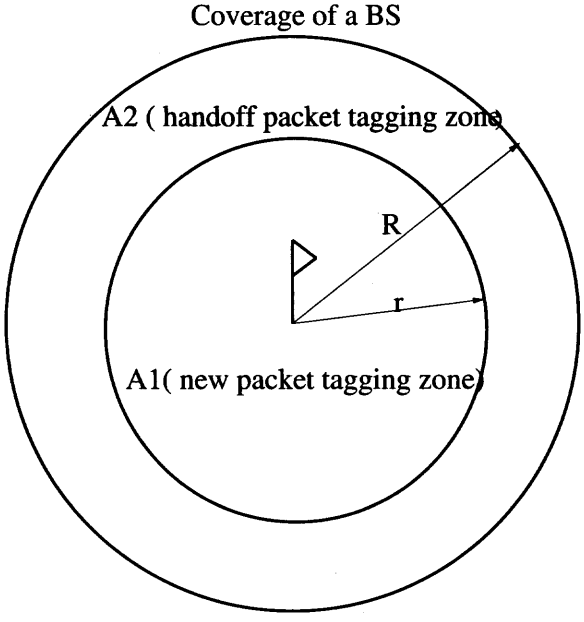

Figure 4.1 New and Handoff Packet Tagging Zone

Handoff packets higher priority over new packets are assigned at the BS's buffer. In order to give priority to handoff packets over new packets, new and handoff packets have to be differentiated. In this case, QoS parameters of new packets and handoff packets could be criteria for call acceptance. How can new and handoff packets are defined? New packets and handoff packets can be defined by the generation areas (e.g., zone A_1 and zone A_2). That is, MT tags new packets and handoff packets based on the coverage zone, as shown in Fig. 4.1; within zone A_1 (e.g., coverage radius

r), new packets; within zone A_2 (e.g., coverage between radius R and r), handoff packets. Note that the areas of zones A_1 and A_2 are

$$S_{A_1} = \pi r^2, \qquad S_{A_2} = \pi(R^2 - r^2).$$

Assumptions for packet tagging are following in EGPRS system;

- The incremental reestablishment scheme in section 1.3.5 is used as a rerouting scheme for handoff. Therefore, the SGSN of EGPRS operates as a crossover switch.

- The MT that starts to transmit packets in zone A_1, tags packets as new packets. When the MT is across the zone A_2, it tags packets as handoff packets.

- The MT that starts to transmit packets in zone A_2, tags packets as handoff packets. When the MT moves into the zone A_1 of same cell, it tags packets as new packets. But, when the MT moves into the zone A_2 of an adjacent cell, it continues to tag packets as handoff packets and transmits handoff packets to new BS, when the handoff occures.

- MT knows where it is, based on the received signal strength from the BS.

- Uniform traffic distribution is considered over the service area. A given packet tagged in a cell belongs to zone A_1 with probability p_1 and to zone A_2 with probability p_2, where $p_1 = S_{A_1}/(S_{A_1} + S_{A_2})$ and $p_2 = 1 - p_1$.

4.4 Cell Switching

A general partial buffer sharing scheme [46][47][48] is considered. In general, loss-sensitive traffic such as data is given priority over loss-tolerant traffic such as voice and video. Real time packets are dropped from a buffer when the buffer occupancy

reaches the threshold. In this work, a threshold-based discarding (TBD) scheme is considered.

As shown in Fig. 4.2, the buffer is partitioned by the i thresholds, S_0,...,S_{I-1}, (e.g., S_{I-1} is the buffer size), corresponding to i priority classes. Packets of priority class i can be buffered up to threshold level S_i. Once the buffer level exceeds S_i, arriving packets of class i are dropped. Note that only new arrivals are dropped; class i packets that are already in the buffer are never dropped and are eventually served.

In order to give priority to handoff packets, some buffer space is reserved for handoff packets of the each type of traffic. The thresholds S_0 and S_1 are for new packets and handoff packets of traffic $n = 1$, respectively; the thresholds S_2 and S_3 are for new packet and handoff packet of traffic $n = 2$, respectively, and so on.

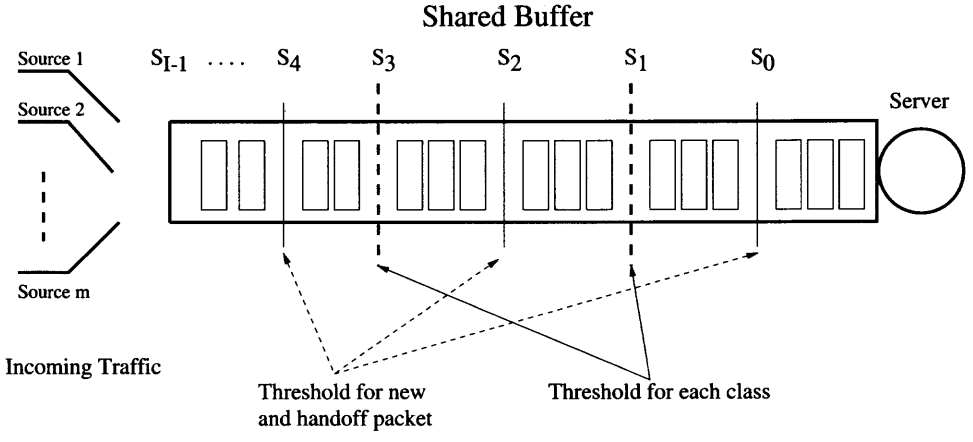

Figure 4.2 Threshold-Based Discarding Scheme handling New and Handoff Packets

4.5 Performance Analysis

The Markov Modulated Poisson Process (MMPP) has been commonly used for modeling arrival rates of point processes. The accuracy of MMPP in modeling an arrival process depends on which statistics of the actual process are used to determine

its parameters. In [41] [51], the superposition of ON-OFF sources is approximated by means of a 2-state MMPP for each traffic.

Four parameters are required to represent the 2-state MMPP source of each traffic, as shown in Fig. 3.1, where $\gamma_{L_i}(\gamma_{H_i})$ is defined as the mean transition rate out of the Low load (High load) state, and $\lambda_{Li}(\lambda_{H_i})$ is the mean arrival rate of the Poisson process in the Low load (High load) state for priority class type i.

The stochastic integral technique proposed in [52] is used to obtain loss probabilities for Markov Modulated Arrival (MMA) streams. In the following, a brief overview of this technique is presented. Consider an arrival process to a finite buffer queueing system. Assume that the buffer size is S_{I-1}. Let $N(t)$ be the number of arrivals in $[0,t]$ and $Z(t)$ denote the number of packets in the queue at time t. Let $U(t)$ be an indicator function for the times at which the buffer is full, namely, $U(t) = 1$ if and only if $Z(t-) = K$, and $U(t) = 0$ otherwise. Then, the packet loss probability P_{loss} is given by:

$$P_{loss} = \lim_{t\to\infty} \frac{1}{N(t)} \int_0^t U(s)dN(s). \tag{4.1}$$

The analysis is based on the following observation. Many stochastic processes such as MMA's have an associated *compensator* $\Lambda(t)$ such that the process $M(t) = N(t) - \Lambda(t)$ is a martingale. Then, under some regularity conditions, as shown in [52], the stochastic integral,

$$R(t) = \int_0^t U(s)dM(s), \tag{4.2}$$

is also a martingale, with the property that $\lim_{t\to\infty} R(t)/t = 0$. The regularity conditions can be shown to hold for MMA's. Given this limiting ratio and by rearranging terms in equation (4.1), there is an expression as following

$$P_{loss} = \frac{\phi}{r}, \tag{4.3}$$

where $r = \lim_{t\to\infty} N(t)/t$ is the arrival rate and $\phi = \lim_{t\to\infty}(1/t)\int_0^t U(s)d\Lambda(s)$.

Let $Y(t)$ be the underlying Markov process which modulates the arrival process of an MMA. Let's assume that the process $Z(t)$ is also Markovian, and let π denote the limiting distribution of the Markov process $\{Y(t), Z(t)\}$. For an MMA, the limit ϕ depends on π. Expressions for the packet loss probability and delay are provided by computing r and ϕ. Note that when the arrival process is Poisson with rate λ, the compensator is $\Lambda(t) = \lambda t$. Then $P_{loss} = \lim_{t\to\infty}(1/t)\int_0^t U(s)$, which gives the well-known results equating the loss probability with the probability that the buffer full.

4.5.1 Computation of packet loss probabilities

The multiplexing of I heterogeneous class types (with different parameter values) is considered. Consider a single queueing system driven by I 2-state MMPP arrival processes. The queueing system has a finite buffer space of size S_{I-1} packets. Service times from I kinds of class sources are exponentially distributed with rate μ. A 2-state MMPP is characterized by a Markov process that alternates between two states, spending an exponentially amount of time in each. Packets are generated in each state according to Poisson process with a rate that is state-dependent.

The generation of packets when the MMPP is in state v_i follows a Poisson process with rate λ_{v_i} for $v_i = \{\lambda_{L_i}, \lambda_{H_i}\}$, i=0, 1, 2, ...,$I-1$. Define the following indicator functions:

$$E_{v_i}(t) = \begin{cases} 1 & \text{if } Y_i(t) = v_i \\ 0 & \text{otherwise.} \end{cases}$$

Then, the aggregate arrival rate at time t is $\sum_{i=0}^{I-1}[E_{L_i}(t)\lambda_{L_i} + E_{H_i}(t)\lambda_{H_i}]$. Let $Z(t)(0 \leq Z(t) \leq S_{I-1})$ denote the system state (the number of packets in the system) at time t. Define the following indicator function for system state q:

$$U_q(t) = \begin{cases} 1 & \text{if } Z_i(t-) = q \\ 0 & \text{otherwise.} \end{cases}$$

Let $N(t) = \sum_{i=0}^{I-1} N_i(t)$ be the cumulative number of arrivals in the time interval [0, t], and let $\Lambda_i(t)$ denote the compensator for i=0, 1, 2, ..., $I-1$, respectively. It is well-known that the compensator for $N_i(t)$ is given by [52].

$$\Lambda_i(t) = \int_0^t (E_{L_i}(s)\lambda_{L_i} + E_{H_i}(s)\lambda_{H_i})ds, \qquad for \;\; i = 0, 1, 2, ..., I-1. \tag{4.4}$$

Finally, the following limiting probabilities are defined. Let $\pi(v_0, v_1, ..., v_{I-1}, q)$ $(0 \leq q \leq S_{I-1})$ be the limiting distribution for the Markov process $\{Y_0(t), Y_1(t), ...,$ $Y_{I-1}(t), Z(t)\}$. Note that $\sum_{\{v_1,v_2,...,v_{I-1}\}} \pi(v_0, v_1, ..., v_{I-1}, q) = \pi(v_0, q)$ and $\sum_{\{v_0,v_2,...,v_{I-1}\}} \pi(v_0, v_1, ..., v_{I-1}, q) = \pi(v_1, q)$ and so on. In this analysis, we obtain the following probabilities.

- the probability $P_i(q)$ that an arrival from priority class i sees the system in state q.
- the probability $P(q)$ that an arbitrary arrival sees the system in state q.

From these probabilities, the packet loss probabilities for each priority class can be easily obtained. First, calculate the probability $P_i(q)$ for an arrival from class i source to see the system state q. From equation (4.3),

$$P_i(q) = \frac{1}{r_i} \lim_{t\to\infty} \frac{1}{t} \int_0^t U_q(s)d\Lambda_i(s). \tag{4.5}$$

Here, the arrival rate from priority class i, r_i, is

$$r_i = \lim_{t\to\infty} \frac{N_i(t)}{t} = \frac{\lambda_{L_i}\frac{1}{\gamma_{L_i}} + \lambda_{H_i}\frac{1}{\gamma_{H_i}}}{\gamma_{L_i} + \gamma_{H_i}} = \frac{\lambda_{L_i}\gamma_{H_i} + \lambda_{H_i}\gamma_{L_i}}{\gamma_{L_i} + \gamma_{H_i}}. \tag{4.6}$$

Next,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t U_q d\Lambda_i(s) = \lambda_{L_i}\pi(L_i, q) + \lambda_{H_i}\pi(H_i, q). \tag{4.7}$$

From equation (4.5), (4.6) and (4.7),

$$P_i(q) = \frac{(\gamma_{L_i} + \gamma_{H_i})(\lambda_{L_i}\pi(L_i, q) + \lambda_{H_i}\pi(H_i, q))}{\lambda_{L_i}\gamma_{H_i} + \lambda_{H_i}\gamma_{L_i}}, \qquad for \quad i = 0, 1, 2, .., I-1. \tag{4.8}$$

Next, the probability $P(q)$ of an arbitrary arrival seeing the system state q is computed. From equation (4.2),

$$P(q) = \frac{1}{\sum_{i=0}^{I-1} r_i} \lim_{t\to\infty} \frac{1}{t}\int_o^t U_q(s) d\Lambda(s), \tag{4.9}$$

where $\Lambda(s) = \sum_{i=0}^{I-1} \Lambda_i(s)$. From equations (4.6), (4.7) and (4.9),

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t U_q d\Lambda_i(s) = \sum_{i=0}^{I-1}(\lambda_{L_i}\pi(L_i, q) + \lambda_{H_i}\pi(H_i, q)). \tag{4.10}$$

From equations (4.6), (4.9) and (4.10),

$$P(q) = \frac{1}{\sum_{i=0}^{I-1} r_i}[\sum_{i=0}^{I-1}(\lambda_{L_i}\pi(L_i, q) + \lambda_{H_i}\pi(H_i, q))]. \tag{4.11}$$

Recall that $P_i(q)$ denotes the probability that an arrival from priority class i sees the system in state q.

$$P_i(loss) = \sum_{q=S_i}^{S_{I-1}} P_i(q), \qquad for \quad i = 0, 1, 2, .., I-1. \tag{4.12}$$

4.5.2 Computation of average packet waiting time

For M/G/1 queueing system, the average number of queue, N_Q, and the mean residual time, R, are noted to find the average packet waiting time. These two random variables mean the average number of queue and mean residual time seen by an outside observer at a random time [55]. This concept can be adapted to our system. The following parameters are defined.

- N_{Q_i} is defined as the average number of packet in queue seen by a packet from the priority class i.
- R_i is defined as the average residual time seen by a packet from the priority class i.
- $\overline{X}$ is defined as the average packet transmission time.
- $\overline{X^2}$ is defined as the second moment of $\overline{X}$.

From equation (4.8),

$$N_{Q_i} = \sum_{q=1}^{S_i} (q-1) P_i(q), \qquad for \quad i = 0, 1, 2, .., I-1. \tag{4.13}$$

From equation (4.6),

$$\begin{aligned} R_i &= \frac{\{r_i(1 - P_i(loss))\}\overline{X^2}}{2}, \qquad for \quad i = 0, 1, 2, .., I-1, \\ &where \;\; r_i = \frac{\lambda_{L_i}\gamma_{H_i} + \lambda_{H_i}\gamma_{L_i}}{\gamma_{L_i} + \gamma_{H_i}}. \end{aligned} \tag{4.14}$$

Therefore, the average packet waiting time is,

$$W_i(delay) = R_i + N_{Q_i}\overline{X}. \tag{4.15}$$

4.6 Optimizing Threshold Values

Now, the problem of finding the thresholds is considered to satisfy the PLP_i and D_i (i.e., packet loss probability requirement and average packet delay requirement for each priority class, i, where $i = 0, 1, 2, ..., I-1$). At first, initial threshold for each priority class is assumed to be arbitrary and small. The thresholds are increased by the program until the QoS requirements are simultaneously satisfied for all priority classes. At each step, packet loss probabilities, average packet delays and normalized difference values (i.e., $(P_i(loss)$ - $PLP_i)/PLP_i$) are calculated. When all the QoS requirements are not satisfied, the maximum difference value is found. And then, the threshold for the priority class with maximum difference value is increased by one. When all the QoS requirements for packet loss probability and average packet delay are satisfied, the search procedure is terminated. Buffer size is decided according to QoS requirements of the highest priority class. The search procedure is the following:

Step 1) Initialize $S_i^k = i + 1$ for all i and k=0.

Step 2) $k = k + 1$ where k is number of iterations

Calculate $P_i^k(loss)$ and $W_i^k(delay)$ for all i

If $P_i^k(loss) \leq PLP_i$ and $W_i^k(delay) \leq D_i$ for all i, then terminate.

S_i^k are optimal thresholds.

Otherwise, $S_i^k = S_i^k + 1$, where k= the index i for

which $(P_i^k(loss)$ - $PLP_i)/PLP_i$ is maximum. Then, go to Step 2.

4.7 The Simulation Model and Results

The ON/OFF model to describe the multi-class sources is used. In this simulation, 3 types of traffics (i.e., 6 types of priority classes; two priority classes for each traffic) are used. The type 1 traffic with 32 *Kbps* rate is modeled with the parameter values; mean ON period (=1.0 *s*) and mean OFF period (=1.35 *s*). The type 2 traffic with 320 *Kbps* rate is modeled as the superposition of multiple identical ON/OFF source;

that is, one source is achieved by the superposition of 15 ON/OFF sources , each characterized by the mean ON period (=33 ms) and the mean OFF period (=67 ms) [54]. The type 3 traffic with 128 $Kbps$ rate is modeled with the parameter values; mean ON period (=0.1 s) and mean OFF period (=0.8 s). It is assumed that packet length is exponentially distributed with mean 1024 bytes and system capacity is 4.8 $Mbps$ [56].

Computer simulations are conducted to investigate the performance of the handoff prioritization scheme. New arrivals and handoff arrivals follow independent and identical Poisson distribution. The fraction of total traffic due to each traffic type is fixed (e.g, the arrival fraction of each traffic type is 46%, 8% and 46%). Also, the fraction of each traffic due to handoffs is kept fixed while the total offered traffic is varied (e.g., the fraction of handoff packet for each traffic is fixed).

In Fig. 4.3, packet loss probabilities are plotted as a function of the mean offered load. The simulation results are in close agreement with our numerical analysis. The thresholds and buffer size are assumed to be 11, 15, 18, 20, 27 and 29 for priority class type 0, 1, 2, 3, 4, and 5, respectively. In order to get this result, the fraction of new call and handoff call for each traffic is fixed in the zone A_1 and A_2. That is, the fraction of new packet and handoff packet (e.g., 50% new packets and 50% handoff packets) is proportional to the ratio of zone A_1 and A_2 (e.g., A_2=A_1). When a call is originated in the zone A_2, we assume that a mobile tags handoff packets, rather than new packets.

In Fig. 4.4, the effect of the fraction of handoff packet for each traffic on packet loss probability is shown. We compare the fraction (e.g., 50% new packets and 50% handoff packets) and the fraction (e.g., 70% new packets and 30% handoff packets). At the fraction of 30% handoff packet, packet loss probabilities is decrease compared to the fraction of 50%, even though the fraction of new packet is increased.

In the Table 4.1 and Table 4.2, the optimal thresholds satisfying the QoS requirements (e.g., packet loss probability and average packet delay) are obtained at the offered load of 0.8 and 0.9 with the fraction of 50% handoff packets, respectively. In this simulation, the packet loss probability requirements are assumed to be 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-10} and 10^{-12} for priority class 0, 1, 2, 3, 4, 5, respectively. The average packet delay requirements are assumed to be 68, 76, 85, 93, 110 and 120 (ms). The searching procedure is started at the thresholds (10, 15, 20, 25, 30, 35) and terminated at the thresholds (41, 46, 49, 52, 60, 62) and (39, 46, 50, 53, 62, 65) in the Table 4.1 and Table 4.2 with meeting the QoS requirements, respectively.

In Fig. 4.5 and Fig. 4.6, packet loss probabilities and thresholds are plotted a function of number of iteration to show the searching procedure for optimal thresholds, respectively. Because initial threshold values were assumed to be small, at first, packet loss probability for each priority class is much greater than its QoS requirement. At each iteration, thresholds are increased, and packet loss probabilities are observed to decrease towards meeting the QoS requirements. In this simulation, initial thresholds are assumed to be (10, 15, 20, 25, 30, 35). In Fig. 4.7, packet delay is plotted a function of number of iterations.

Buffer size vs. packet loss probability for a scheme without thresholding is shown, in Fig. 4.8. The buffer size of about 62 is required to satisfy all QoS requirements with partial buffer sharing scheme. However, if threshold scheme is not used, the buffer size of about 175 is required to meet all requirements. Therefore, by using the threshold scheme, resource utilization can be improved while satisfying all QoS requirements.

Thresholds	Packet Loss Probabilities
(10, 15, 20, 25, 30, 35)	$(5.3 \times 10^{-2}, 5.1 \times 10^{-3}, 1.7 \times 10^{-4}, 1.4 \times 10^{-6}, 1.6 \times 10^{-09}, 5.2 \times 10^{-14})$
(14, 18, 22, 25, 33, 35)	$(2.2 \times 10^{-2}, 3.1 \times 10^{-3}, 1.9 \times 10^{-4}, 1.0 \times 10^{-5}, 2.2 \times 10^{-10}, 2.9 \times 10^{-12})$
(16, 21, 24, 26, 35, 37)	$(1.4 \times 10^{-2}, 1.3 \times 10^{-3}, 1.5 \times 10^{-4}, 2.0 \times 10^{-5}, 1.2 \times 10^{-10}, 1.5 \times 10^{-12})$
(18, 23, 26, 29, 37, 39)	$(8.9 \times 10^{-3}, 8.4 \times 10^{-4}, 1.0 \times 10^{-4}, 5.3 \times 10^{-6}, 1.1 \times 10^{-10}, 1.5 \times 10^{-12})$
(20, 25, 28, 31, 39, 42)	$(5.6 \times 10^{-3}, 5.3 \times 10^{-4}, 6.4 \times 10^{-5}, 3.3 \times 10^{-6}, 7.2 \times 10^{-11}, 1.3 \times 10^{-13})$
(22, 27, 31, 33, 42, 44)	$(3.6 \times 10^{-3}, 3.5 \times 10^{-4}, 2.1 \times 10^{-5}, 2.9 \times 10^{-6}, 2.6 \times 10^{-11}, 7.7 \times 10^{-13})$
(25, 29, 33, 35, 44, 46)	$(1.8 \times 10^{-3}, 2.7 \times 10^{-4}, 1.6 \times 10^{-5}, 2.2 \times 10^{-6}, 1.2 \times 10^{-11}, 1.7 \times 10^{-13})$
(27, 32, 35, 38, 47, 49)	$(1.2 \times 10^{-3}, 1.1 \times 10^{-4}, 1.3 \times 10^{-5}, 7.1 \times 10^{-7}, 4.0 \times 10^{-12}, 5.4 \times 10^{-14})$
(29, 34, 38, 40, 49, 51)	$(7.6 \times 10^{-4}, 7.3 \times 10^{-5}, 4.5 \times 10^{-6}, 6.1 \times 10^{-7}, 3.4 \times 10^{-12}, 4.6 \times 10^{-14})$
(32, 37, 40, 42, 51, 53)	$(3.9 \times 10^{-4}, 3.7 \times 10^{-5}, 4.3 \times 10^{-6}, 5.8 \times 10^{-7}, 3.3 \times 10^{-12}, 4.4 \times 10^{-14})$
(34, 39, 42, 45, 53, 56)	$(2.5 \times 10^{-4}, 2.3 \times 10^{-5}, 2.8 \times 10^{-6}, 1.5 \times 10^{-7}, 3.2 \times 10^{-12}, 5.6 \times 10^{-14})$
(36, 41, 44, 47, 56, 58)	$(1.6 \times 10^{-4}, 1.5 \times 10^{-5}, 1.8 \times 10^{-6}, 9.5 \times 10^{-8}, 5.4 \times 10^{-13}, 7.2 \times 10^{-15})$
(41, 46, 49, 52, 60, 62)	$(5.2 \times 10^{-5}, 4.9 \times 10^{-6}, 5.9 \times 10^{-7}, 3.1 \times 10^{-8}, 6.6 \times 10^{-13}, 8.6 \times 10^{-15})$

Table 4.1 Finding the optimal thresholds (50% new packets and 50% handoff packets at offered load 0.8)

Thresholds	Packet Loss Probabilities
(10, 15, 20, 25, 30, 35)	$(1.3 \times 10^{-1}, 2.2 \times 10^{-2}, 1.2 \times 10^{-3}, 1.4 \times 10^{-5}, 3.1 \times 10^{-8}, 7.6 \times 10^{-12})$
(11, 18, 22, 25, 34, 37)	$(1.9 \times 10^{-1}, 1.6 \times 10^{-2}, 1.3 \times 10^{-3}, 7.4 \times 10^{-5}, 1.2 \times 10^{-9}, 2.6 \times 10^{-12})$
(14, 20, 24, 27, 36, 39)	$(8.4 \times 10^{-2}, 9.8 \times 10^{-3}, 9.7 \times 10^{-4}, 6.4 \times 10^{-5}, 1.6 \times 10^{-9}, 2.7 \times 10^{-12})$
(16, 23, 27, 30, 39, 41)	$(6.8 \times 10^{-2}, 5.8 \times 10^{-3}, 5.7 \times 10^{-4}, 4.0 \times 10^{-5}, 6.4 \times 10^{-10}, 1.2 \times 10^{-11})$
(19, 25, 29, 32, 42, 44)	$(4.4 \times 10^{-2}, 5.2 \times 10^{-3}, 5.1 \times 10^{-4}, 3.4 \times 10^{-5}, 1.4 \times 10^{-10}, 3.3 \times 10^{-12})$
(21, 28, 32, 34, 44, 46)	$(3.4 \times 10^{-2}, 3.7 \times 10^{-3}, 3.4 \times 10^{-4}, 5.8 \times 10^{-5}, 2.3 \times 10^{-10}, 4.4 \times 10^{-12})$
(24, 30, 34, 37, 47, 49)	$(2.7 \times 10^{-2}, 3.1 \times 10^{-3}, 3.0 \times 10^{-4}, 2.1 \times 10^{-5}, 1.8 \times 10^{-10}, 1.4 \times 10^{-12})$
(26, 33, 37, 40, 49, 51)	$(2.3 \times 10^{-2}, 1.7 \times 10^{-3}, 1.5 \times 10^{-4}, 1.6 \times 10^{-5}, 2.2 \times 10^{-10}, 3.1 \times 10^{-12})$
(29, 35, 39, 42, 51, 54)	$(1.8 \times 10^{-2}, 1.8 \times 10^{-3}, 1.8 \times 10^{-4}, 1.3 \times 10^{-5}, 2.0 \times 10^{-10}, 5.3 \times 10^{-13})$
(31, 38, 42, 45, 54, 56)	$(1.0 \times 10^{-2}, 1.1 \times 10^{-3}, 1.1 \times 10^{-4}, 7.2 \times 10^{-6}, 1.3 \times 10^{-10}, 2.2 \times 10^{-12})$
(34, 40, 44, 47, 56, 59)	$(9.5 \times 10^{-3}, 1.0 \times 10^{-3}, 1.0 \times 10^{-4}, 7.6 \times 10^{-6}, 1.4 \times 10^{-10}, 3.4 \times 10^{-13})$
(36, 43, 46, 49, 59, 61)	$(7.6 \times 10^{-3}, 6.0 \times 10^{-4}, 1.3 \times 10^{-4}, 7.6 \times 10^{-6}, 3.6 \times 10^{-11}, 6.6 \times 10^{-13})$
(39, 46, 50, 53, 62, 65)	$(5.6 \times 10^{-3}, 4.2 \times 10^{-4}, 4.8 \times 10^{-5}, 3.9 \times 10^{-6}, 5.3 \times 10^{-11}, 1.3 \times 10^{-13})$

Table 4.2 Finding the optimal thresholds (50% new packets and 50% handoff packets at offered load 0.9)

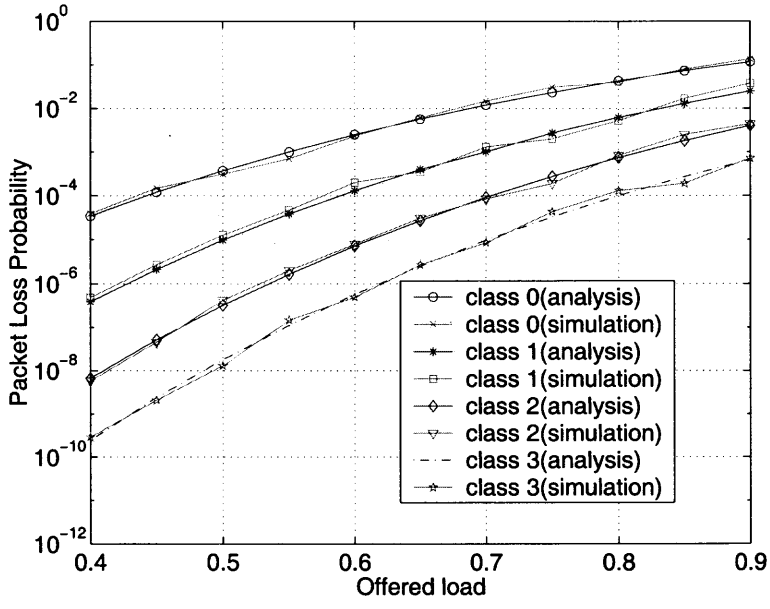

Figure 4.3 Packet loss probabilities vs. offered load (comparison between analysis and simulation)

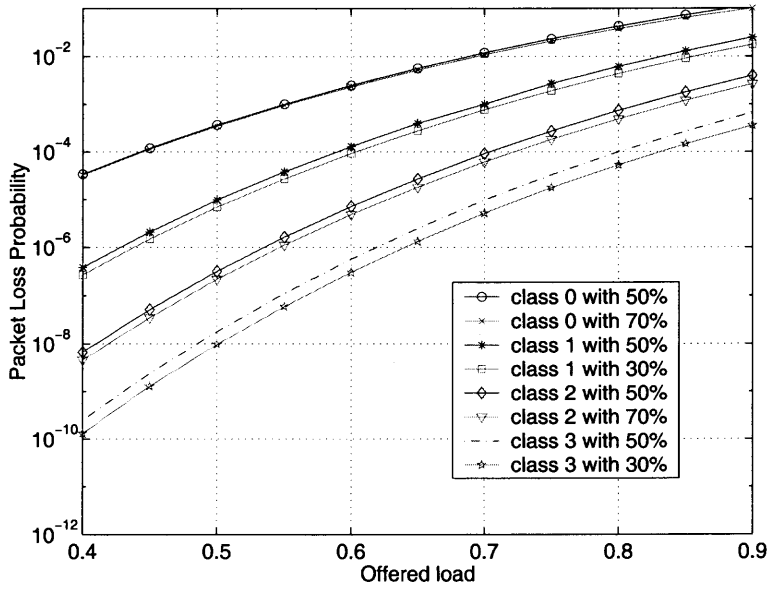

Figure 4.4 Packet loss probabilities vs. offered load (different fractions of handoff packet)

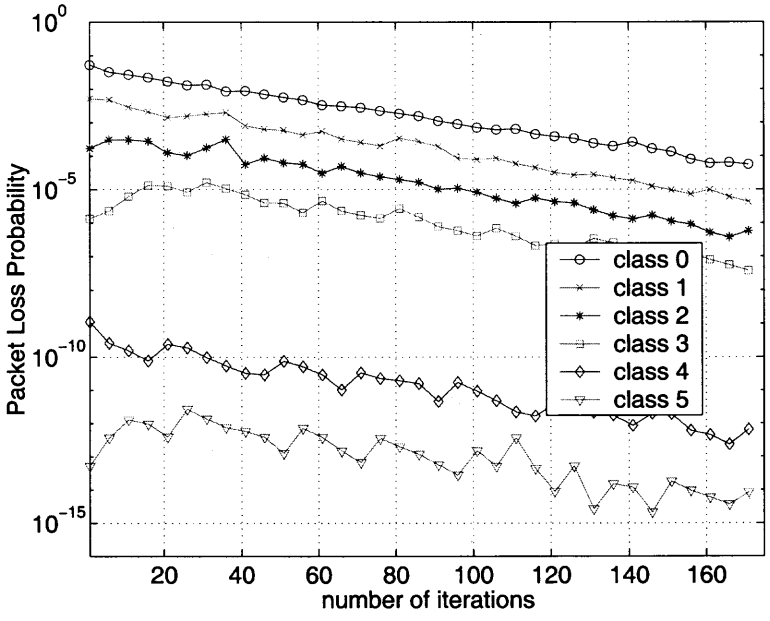

Figure 4.5 Packet loss probabilities vs. number of iterations (offered load 0.8)

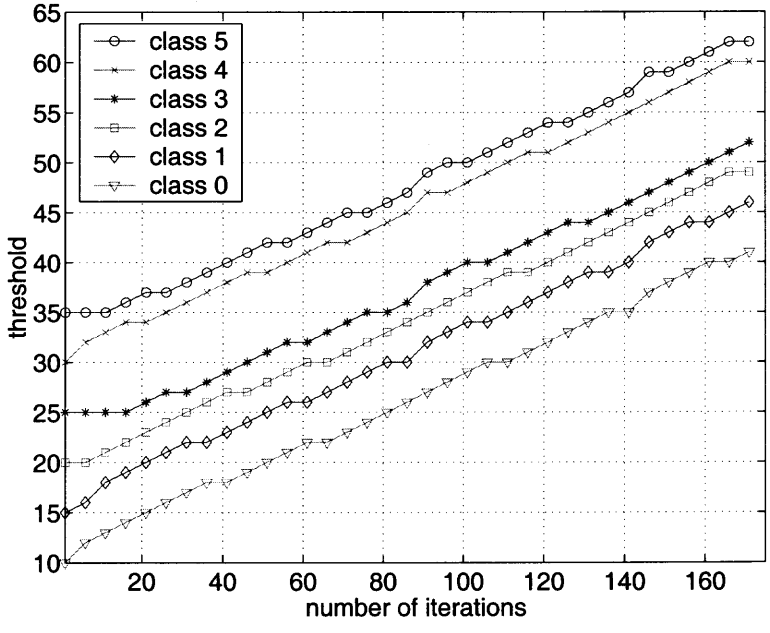

Figure 4.6 Thresholds vs. number of iterations (offered load 0.8)

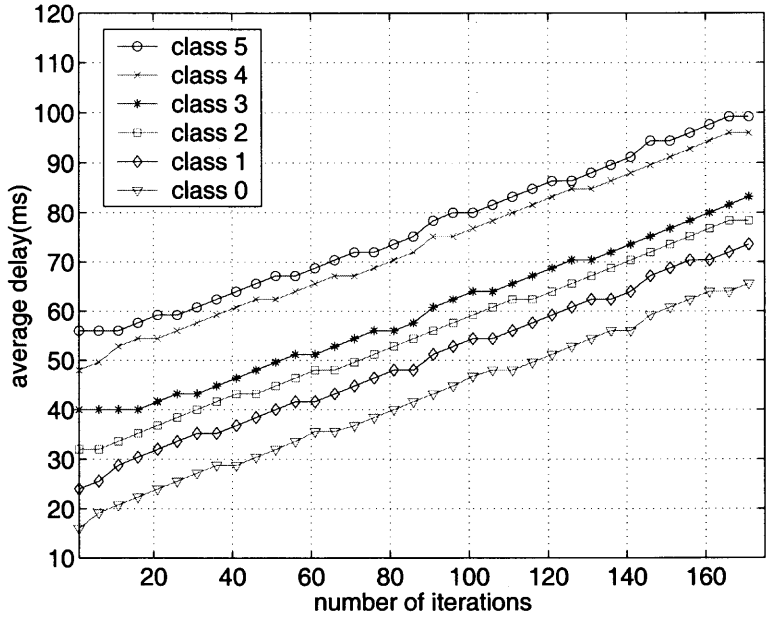

Figure 4.7 Average delay vs. number of iterations (offered load 0.8)

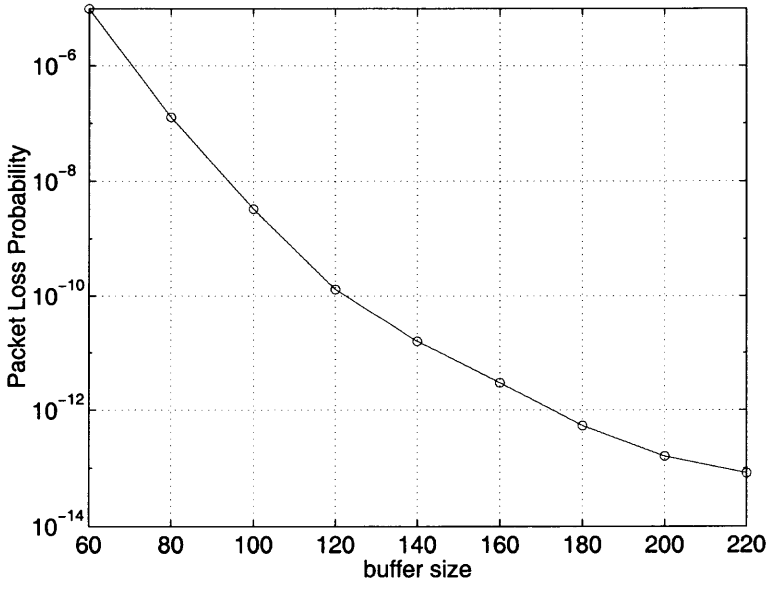

Figure 4.8 Buffer size vs. packet loss probability (for a scheme without thresholding)

CHAPTER 5

CONCLUSIONS

This dissertation studied the buffer management scheme for increasing the channel utilization, and made the cell switching management scheme in Wireless Packet Communications.

First, buffer management scheme has been proposed to improve the performance of wireless packet network. Based on the existing partial buffer sharing scheme, a new space priority scheme for real time traffic was proposed. Through the simulation results, the performance improvement has been shown, compared to the existing scheme under the condition that the threshold is fixed according to the maximum packet delay for real time traffic. However, the complexity for implementation has been experienced.

The second chapter has studied the cell loss performance of an ATM MUX loaded with a traffic stream from the superposition of multiple ON-OFF sources in the two-class environment using the proposed buffer management scheme. By modeling each type of traffic by a 2-state MMPP, the CLP of the respective traffics (i.e., real time traffic and non-real time traffic) using the proposed SDPI space priority scheme could be derived. This scheme is applicable to schedule delay-tolerant non-real time traffic and delay-sensitive real time traffic. That is, by delaying the non-real time traffic cells and pushing in the real time traffic cells selectively, more real time traffic can be accepted within the acceptable QoS requirement (e.g., CLP). By provisioning additional priority to real time traffic, SDPI compensates for the disadvantage of threshold-based discarding (TBD) scheme which favors non-real time traffic at an expense of real time traffic, under the circumstance that the threshold is fixed due to the maximum cell delay of real time traffic. Thus, channel utilization is improved for real time traffic. Simulations have also validated numerical analysis.

Finally, a novel cell switching scheme is considered to support QoS guarantees in packet-switched wireless cellular networks. A new method to examine the effect of packet priority scheme (e.g., partial buffer sharing scheme) at cell switching is proposed. That is, using our packet tagging method, packets are differentiated into new packets and handoff packets, and prioritized handoff packets. By modeling each type of priority class by a 2-state MMPP, the packet loss probability and average packet delay of the respective priority classes using the space priority scheme could be derived. Optimal thresholds at a specified offered load can be obtained through the proposed search procedure. The performance of proposed scheme was evaluated by simulation and numerical analysis in terms of packet loss probability and average packet delay.

REFERENCES

1. L. J. Cimini *et al.*, "Advanced Cellular Internet Service (ACIS)," IEEE Commun. Mag., pp. 150-159, Oct. 1998.

2. A. Furuskar *et al.*, "EDGE: Enhanced Data Rates for GSM ann TDMA/136 Evolution," IEEE Pers. Commun. pp. 56-66, June. 1999.

3. K. Balanchadran *et al.*, "GPRS-136: High-Rate Packet Data Service for North American TDMA Digital Cellular Systems," IEEE Pers. Commun., pp. 34-47, June, 1999.

4. G. P. Pollini, "Trends in Handover Design," IEEE Commun. Mag. Mar. 1996.

5. A. K. Salkintzis, "Radio Resource Management in Cellular Digital Packet Data Network," IEEE Personal Communications, pp. 28-36, Dec. 1999.

6. J. Cai and D. J. Goodman, "General Packet Radio Service in GSM," IEEE Communication Magazine, vol. 35. no. 10, Oct. 1997.

7. D. Hong and S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone System with Prioritized and Nonprioritized Handoff Procedures," IEEE Trans. on Vehicular Technology," vol. vt-35, no. 3, pp. 77-92, Aug. 1986.

8. S. Tekinay and B. Jabbari, "A Measurement-Based Prioritization Scheme for Handovers in Mobile Networks," IEEE JSAC vol. 10, no. 8, pp. 1343-50, Oct. 1992.

9. M. Naghshineh, "Distributed Call Admission Control in Mobile/Wireless Networks," IEEE JSAC, vol. 14, no. 4, May, 1996.

10. M. Marsan, C-F Chiasserini and R. Cigno, "Local and Global Handovers for Mobility Management in Wireless ATM Networks," IEEE Personal Communications, pp. 16-24, Oct. 1997.

11. K. Eng, M. Karol and M. Veeraraghavan, "BAHAMA: A Broadband Ad-Hoc Wireless ATM Local Area Network," Proc. Icc'95, pp. 1216-1223, June, 1995.

12. B. Akyol and D. Cox, "Signaling Alternatives in Wireless ATM Network," IEEE JSAC, vol. 15, no. 1, pp. 35-49, Jan. 1997.

13. A. Acampora and M. Naghshineh, "An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks," IEEE JSAC, vol. 12, no. 8, pp. 1365-1374, Oct. 1994.

14. G. Hebuterne and A. Gravey, "A Space Priority Queueing Mechanism for Multiplexing ATM Channels," ITC Specialist Seminar, Adelaide, Sept. 1989.

15. J. Garcia and O. Casals, "Stochastic Models of Space Priority Mechanisms with Markovian Arrival Processes," Annals of Operations Research 35, pp. 271-296, 1992.

16. H. Kroner and G. Hebuterne, "Priority Management in ATM Switching Nodes," IEEE JSAC, vol. 9, no. 3, pp. 418-427, April 1991.

17. F. Bonomi, L. Fratta and S. Montagna, "Priority on Cell Service and on Cell Loss in ATM Switching," in Proc. 7th ITC Seminar, Morristown, NJ, Oct. 1990.

18. A. Choudhury and E. Hahne, "Space Priority Management in a Shared Memory ATM Switch," IEEE Proc. Globecom'93, pp. 1375-1380

19. J-F Chang and C-S Wu, "The Effect of Prioritization of a Concentrator under an Accept, otherwise Reject Strategy," IEEE Trans. Commun., vol. 38, no. 7, pp. 1031-1039, July 1990.

20. J. Hall and P. Mars, "Critical Review of Buffer Allocation Policies in ATM Switch," IEE Fourteenth UK Teletraffic Symposium, March 1996.

21. T-Y Huang and J-L Wu, "Performance Analysis of ATM Switches using Priority Schemes," IEE Proc. Communications, vol. 141, pp. 248-254, Aug. 1994.

22. T-Y Huang, et al., "Priority Management to Improve the QoS in ATM Networks," IEICE Trans. Commun. vol. E76-B, no. 3, pp. 249-257, March 1993.

23. Arthur Y-M Lin and J. A. Silvester, "Priority Queueing Strategies and Buffer Allocation Protocols for Traffic Control at an ATM Integrated Broadband Switching System," IEEE JSAC, vol. 9, no. 9, pp. 1524-1536, Dec. 1991.

24. N. Yin, S-Q Li and T. E. Stern, "Congestion Control for Packet Voice by Selective Packet Discarding." IEEE Trans. Commun. vol. 38, no. 5, pp. 674-683, May 1990.

25. H. Sato and M. Umehira, "A Novel Buffer Contorl Scheme for ATM Cell Transport with Improved Cell Delay Variation for Wireless ATM," Proc. PIMRC'96, pp. 926-932, Oct. 1996.

26. H. Mitts and S. Veikkolainen, "Use of ABR Flow Control in Wireless ATM Systems," Proc. ICUPC'96, pp. 702-706, Oct. 1996.

27. D. Petr and V.S. Frost, "Nested Threshold Cell Discarding for ATM Overload Control: Optimization Under Cell Loss Constraints," Proc. IEEE INFORCOM'91, pp. 12A.4.1-12A.4.10, 1991.

28. G. Gallasi, G. Rigoio and P. Vaccari, "Resource Allocation in ATM Networks," in Proc. Third RACE 1022 Workshop, Paris, France, Oct. 1898.

29. J. Y. Hui, "Resource Allocation for Broadband Networks," IEEE JSAC, vol. 6, no. 9, pp. 1598-1608, Dec. 1988.

30. L. Chang, S. Chang and H. Hughes, "A Connection Admission Control Algorithm based on Empirical Traffic Measurements," IEEE Proc. Icc'95, pp. 793-797, June 1995.

31. M. Neghshineh and M. Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks," IEEE JSAC, vol. 14, no. 4, pp. 711-717, May. 1996.

32. P. T. Brady, " A Model for On-Off Speech Patterns in Two-Way Conversation," Bell Syst. Tech. Journal, 48(7), pp. 2445-2472, Sep. 1969

33. D. Heyman, T. Lakshman, A. Tabatabai and H. Heeke, "Modeling Teleconference Traffic from VBR Video Coders," IEEE Proc. Icc'94, vol. 3, pp. 1744-1748, 1994.

34. D. Heyman, A. Tabatabai and T. Lakshman, "Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM Networks," IEEE Trans. on Circuits and Systems for Video Technology, vol. 2, no. 1, pp. 49-58, March 1992.

35. H. Kroner, "Comparative Performance Study of Space Priority Mechanisms for ATM Channel," IEEE Proc.Infocom'90, San Francisco, pp. 1136-1143, June 1990.

36. H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," IEEE JSAC, vol. SAC-4, no. 6, pp. 856-868, Sept. 1986.

37. R. Nagarajan, J. F. Kurose and D. Towsley, "Approximation Techniques for Computing Packet Loss in Finite-Buffered Voice Multiplexers," IEEE JSAC, vol. 9, no. 3, pp. 368-377, April 1991.

38. A. Baiocchi, N. B. Melazzi and M. Listanti, "Loss Performance Analysis of an ATM Multiplexer Loaded with High Speed ON-OFF Sources," IEEE JSAC, vol. 9, no. 3, pp. 388-393, April 1991.

39. S. Shah-Heydari and T. Le-Ngoc, "MMPP Modeling of Aggregated ATM Traffic," Canadian Conference on Electrical and Computer Engineering (CCECE'98), pp. 129-132, 1998.

40. F. Yegenoglu and B. Jabbari, "Performance Evaluation of MMPP/D/1/K Queues for Aggregate ATM Models," IEEE Proc. Infocom'93, pp. 1314-1319, 1993.

41. S. Kim, M. Lee and M. Kim, "Σ-Matching Technique for MMPP Modeling of Heterogeneous ON-OFF Sources," IEEE Proc. Globecom'94, pp. 1090-1094, 1994.

42. B. Wolff, "Poisson Arrivals See Time Averages," Operational Research, vol. 30, no. 2, pp. 223-231, Apr. 1982.

43. R. Cooper, "Introduction to Queueing Theory," New York: Macmillan, 1972.

44. D. Gross and C. M. Harris, "Fundamentals of Queueing Theory," New York: Wiley, 1985.

45. S. Kowtha and D. Vaman, "A Generalized ATM Traffic Model and its Application in Bandwidth Allocation," IEEE Proc., ICC'92, pp. 1009-1013, 1992.

46. H. Kroner, G. Hebuterne and P. Boyer, "Priority Management in ATM Switching Nodes," IEEE JSAC, vol. 9, no. 3, pp. 418-427, Apr. 1991.

47. M. Krunz, H. Hughes and P. Yegani, "Design and Analysis of a Buffer Management Scheme for Multimedia Traffic with Loss and Delay Priorities," IEEE Proc. Globecom'94, pp. 1560-1564, Nov. 1994.

48. D. W. Petr and V. S. Frost, "Nested Threshold Cell Discard for ATM Overload Control: Optimization under Cell Loss Constraints," IEEE Proc. Infocom'91, pp. 12A.4.1-12A.4.1.10, 1991.

49. S. Faccin *et al.*, "GPRS and IS-136 Integration for Flexible Network and Services Evolution," IEEE Personal Communications, pp. 48-54, June 1999.

50. X. Qiu *et al.*, "RLC/MAC Design Alternatives for Supporting Integrated Services over EGPRS," IEEE Pers. Commun. pp. 20-33, April 2000.

51. J. Bang, N. Ansari and S. Tekinay, "Selective-delay Push-in Buffering Mechanism for QoS Provisioning in ATM Switching Nodes Loaded with ON-OFF Arrival Processes," The 15th International Conference on Information Networking (ICOIN-15), Beppu, Japan, pp. 799-804, Jan. 2001.

52. W. A. Rosenkrantz and R. Simha, "Some Theorems on Conditional Pasta: A Stochastic Integral Approach," Operations Research Letters, vol. 11, pp. 173-177, April 1992.

53. J. B. Kim, R. Simha and T. Suda, "Analysis of a Finite Buffer Queue with Heterogeneous Markov Modulated Arrival processes: A Study of Traffic Burstiness and Priority packet Discarding," Computer Networks and ISDN Systems 28 pp. 653-673, 1996.

54. X. Wu, S. Wu and H. Sun, "Dynamic Slot Allocation Multiple Access Protocol for Wireless ATM Networks," IEEE Proc. Icc'97, pp. 1560-1565, 1997.

55. D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, New Jersey, 1992.

56. GSM 03.60, V5.2.0, "Digital Cellular Telecommunication System (Phase 2); General Packet Radio Service (GPRS); Service Description (Stage 2)," Dec. 1997.