INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Radio Link Control and Transport Layer Protocol Design Issues in Wireless IP Networks

by

Abu Zafar Mohammad Ekram Hossain
M.Sc.(Eng.), Bangladesh University of Engineering and Technology, 1997

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

We accept this thesis as conforming
to the required standard

Prof. V. K. Bhargava, Supervisor, Dept. of Electrical & Computer Engineering

Prof. F. El-Guibaly, Member, Dept. of Electrical & Computer Engineering

Prof. K. F. Li, Member, Dept. of Electrical & Computer Engineering

Prof. S. Dost, Outside Member, Dept. of Mechanical Engineering

Prof. S. Roy, External Examiner

Supervisor: Prof. V. K. Bhargava

ABSTRACT

Packet-switched wireless data networks built upon IP (Internet Protocol)-based infrastructure are being envisioned to provide ubiquitous Internet access to mobile users. Supporting packet-data services along with the cellular voice services in an integrated networking framework gives rise to new network infrastructure and protocol design issues that are to be resolved to facilitate the introduction of the next generation wireless IP networks.

This thesis addresses several protocol design issues in the area of wireless packet data networking, namely, retransmission control design for multichannel protocols, radio link level protocol for dynamic rate and error control, inter-layer protocol dependency, radio link-layer and transport-layer protocol fairness and radio link-level dynamic bandwidth allocation. A retransmission control policy for a multichannel S-ALOHA (slotted ALOHA) protocol in a high speed wireless data network is proposed and analyzed. A sub-optimal dynamic rate adaptation procedure is proposed for uplink data transmission in WCDMA (Wideband Code Division Multiple Access)-based wireless IP networks. The performance of this scheme is analyzed using a novel 'mean-sense' approach for interference calculation in cellular WCDMA environment. The impact of macrodiversity packet combining on transport-protocol throughput performance is analyzed under different link-level retransmission control policies. A unified TDMA (Time Division Multiple Access)-based centralized bandwidth management mechanism is proposed as a link-level solution for providing service fairness among competing users for uplink data transmission in a wireless IP network. TCP (Transmission Control Protocol) performance is evaluated under different transport-level packet scheduling policies in a correlated fading environment and a time frame-based scheduling policy is proposed to provide service fairness among mobile users in the case of downlink transmission. A set of centralized burst-level bandwidth allocation policies are investigated as a means of service integration with QoS (Quality of Service) provisioning in the wireless IP air interface.

Examiners:

Prof. V. K. Bhargava, Supervisor, Dept. of Electrical & Computer Engineering

Prof. F. El-Guibaly, Member, Dept. of Electrical & Computer Engineering

Prof. K. F. Li, Member, Dept. of Electrical & Computer Engineering

Prof. S. Dost, Outside Member, Dept. of Mechanical Engineering

Prof. S. Roy, External Examiner

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to express my profound gratitude and appreciation to Professor Vijay K. Bhargava for his continuous support, inspiration and understanding. I have been very fortunate to work in such a superb intellectual environment under Professor Bhargava whose supervision and mentoring have always been creating a 'morality of aspiration' inside me.

I am very grateful to Professor Fayez El-Guibaly, Professor Kin F. Li and Professor Sadik Dost for serving on my examination committee.

Special thanks goes to Professor Sumit Roy for agreeing to be the external examiner in my Ph.D. oral examination.

Last but not the least, I would like to extend my sincere thanks to all of my colleagues and friends here in Victoria for their friendship and cooperation.

Dedication

To my grand ma late Mrs. Zobeda Khatun and to the members of my family whose blessing. love and affection have been the greatest possession in my life

"And when old words die out on the tongue. new
melodies break forth from the heart: and where the
old tracks are lost. new country is revealed with its wonders."

Rabindranath Tagore. Gitanjali

Chapter 1

Introduction

1.1 Wireless Communications and Networking Technology

Today, wireless communications and networking is the fastest growing segment of the telecommunications industry. Traditional wireless communications and networking systems, which include cordless and cellular phones, paging systems, mobile data networks and mobile satellite systems, have evolved from the enormous research and development efforts in both academia and industry.

Wireless communications and networking systems can be categorized as follows [1]:

- *High Power Wide Area Systems (or Cellular Systems)* that have been designed to support mobile users roaming over wide geographic area
- *Low Power Local Area Systems*, for example, cordless telephone systems that are implemented with relatively simpler technology
- *Low Speed Wide Area Systems* that have been designed for mobile data services with relatively low data rates (e.g., CDPD (Cellular Digital Packet Data) systems)
- *High Speed Local Area Systems* that have been designed for high speed and local communications (e.g., WLAN (Wireless LAN) systems).

The first two categories are voice-oriented systems while the remaining two are data-oriented systems. Developments in voice-oriented systems can be described by referring to three generations of these systems as follows:

- *First Generation Systems* refer to analog cellular phones and simple cordless phones which were based on analog technology with FM modulation. NMT (Nordic Mobile Telephone), AMPS (Advanced Mobile Phone Service) and TACS (Total Access Communications Systems) are the three primary analog cellular radio systems standards. CT (North American Cordless Telephone), CT0 and CT1/CT1+ are the first generation analog cordless standards.
- *Second Generation (2G) Systems* refer to the digital cellular and cordless systems which employ digital modulation and advanced call processing capabilities. GSM (Global System for Mobile Communication)-the pan-European digital cellular radio system, IS-54/IS-136-the TDMA (Time Division Multiple Access)-based North American digital cellular systems, and IS-95-the CDMA (Code Division Multiple Access)-based digital cellular system and the Japanese PDC (Personal Digital Cellular Radio) represent the second generation of wireless cellular systems. CT2/CT2+, DECT (Digital European Cordless Telecommunications), PHS (Personal Handy Phone System) and the cordless in the ISM[1] bands are among the major second generation digital cordless standards.
- *Third Generation (3G) Systems* are envisioned to support multidimensional (multi-information media, multi-transmission media and multilayered networks) high-speed wireless communications [2]. High speed packet data services up to 2 *Mbps* for wireless Internet access is expected to be the main attribute of 3G systems. Studies and standardization efforts on 3G systems are being carried out under the names IMT-2000 (International Mobile Telecommunications in the year 2000), UMTS (Universal Mobile Telecommunication Service) and MBS (Mobile Broadband System). Majority of candidate RTT (Radio Transmission Technology) proposals for IMT-2000 choose DS/CDMA (Direct Sequence Code Division Multiple Access) as the leading multiple access technique. Activities on the harmonization of the technical specifications for the different RTT candidates are being carried out with an aim to achieve a single converged 3G standard.

[1]It refers to the 902-928 *MHz*, the 2400-2483.5 *MHz* and the 5725-5850 *MHz* bands which have been allowed for unlicensed use of spread-spectrum (direct-sequence or frequency-hopping) radios up to 1 *W*.

Although wireless cellular networks have been very successful in providing mobile voice service, the growth of wide area wireless data services has been much slower. Some of the contributing factors are ([3]-[6]): low speed wireless access links, poor link performance, fragmentation of wireless access standards, protocol and software design issues, lack of core packet network infrastructure, lack of synergy with Internet, expensive user devices and battery technology.

But with the explosive demand for Internet access and the development of a huge number of innovative Internet-based applications and services, it is expected that there will be widespread adoption of wireless data services in the near future. Similar to the evolution of wireline Internet, wireless Internet services will evolve to high bandwidth services such as full multimedia email and web browsing, starting with the low bandwidth text-based services such as wireless e-mail and web information retrieval. Evolving wireless packet networking technology would make it possible to provide seamless wide-area Internet service to mobile users.

Current low-bandwidth wireless data systems include

- *ARDIS (Advanced Radio Data Information Service)*: the ARDIS mobile data network formed by IBM and Motorola is providing coverage in over 400 metropolitan areas in the USA. It uses 25 *KHz* channels in FDMA mode in the 800 *MHz* band, and provides 4.8 *Kbps* transmission rates in most areas [7].
- *RAM Mobile Data*: the RAM Mobile Data Network, based on the MOBITEX protocol developed by Ericsson and Swedish Telecomm, uses 12.5 *KHz* channels in the 896-901 *MHz* and 935-940 *MHz* bands and provides 8 *Kbps* transmission rates. It uses S-ALOHA as the distributed medium access control mechanism.
- *CDPD (Cellular Digital Packet Data)*: the CDPD system provides packet data service as a non-interfering overlay to the AMPS analog cellular system using paired 30 *KHz* channels. It provides a transmission rate of 19.2 *Kbps* with the maximum user throughput about half this rate. CDPD uses the DSMA (Digital Sense Multiple Access) technique. Higher layers of the CDPD protocol stack are based on standard ISO and Internet protocols.

Apart from the above systems, 2G cellular systems also support circuit- and packet-switched data. For example, GSM supports circuit-switched data from 2.4-14.4 *Kbps*. GPRS (General Packet Radio Service), which is a separate packet data

network within GSM. is expected to provide data rates up to 170 *Kbps*. IS-136. the TDMA-based North American 2G standard. can support data rates upto 40-60 *Kbps*. IS-95. the CDMA-based North American digital cellular standard. supports circuit-switched data up to 14.4 *Kbps*. IS-95B can support a peak data rate of 76.8 *Kbps*. Although the cordless systems. such as PHS. DECT and PACS. were designed primarily for voice service. with larger bandwidth available in these systems. data services in the 32-500 *Kbps* range could be supported.

3G systems aim to provide data services with rates of 144 *Kbps* in wide-area vehicular environments and up to 2 *Mbps* in indoor and picocellular environments. EDGE (Enhanced Data for GSM Evolution). which is an enhancement to GSM. is expected to increase data rates to over 384 *Kbps*. Peak data rate achievable in systems based on cdma2000. the 3G North American CDMA standard. would be 307.2 *Kbps* in the 1.25 *MHz* bandwidth or 614.4 *Kbps* in the 5 *MHz* bandwidth.

The WLANs. because of their limited transmission range and wider bandwidths (compared to cellular systems). can provide data rates of 1-16 *Mbps*. Three distinct technologies are currently in use for WLAN systems: IR (Infrared WLANs). microwave WLANs operating at 18-19 *GHz* and spread-spectrum WLANs operating in the ISM bands. IR WLANs are of two types–DFIR (Diffused IR) WLAN and DBIR (Directed-beam IR) WLAN [1]. IR systems are relatively simple. are flexible to install. operate at very low power levels and are less likely to interfere with each other. IR signals do not penetrate walls. Data rates of currently available DFIR and DBIR WLANs are 1 *Mbps* and 10 *Mbps*. respectively. Microwave WLAN products from Motorola (e.g.. Altair and Altair Plus systems) provide a data rate as high as 15 *Mbps*. Spread-spectrum WLANs provide larger coverage than IR and microwave WLANs and its achievable data rates are 2-6 *Mbps*. Lucent's WaveLAN supports data rate up to 2 *Mbps*.

New WLAN standards. such as HIPERLAN (High-Performance LAN). aim to deliver data rates of up to 20 *Mbps*.

1.2 IP (Internet Protocol)-Based Wireless Data Networks

Packet-switched wireless data networks built upon IP (Internet Protocol)-based infrastructure can leverage the rapidly evolving Internet technology in the wireless domain so that user expectations can be met in a cost-effective manner. IP-centric solutions for high-speed wide area wireless networking are expected to incur lower network infrastructure cost, provide radio technology independent core network infrastructure, provide flexible service architecture based on client-server paradigm, simple service creation and application development environment and ultimately will pave the way to build a global mobile Internet.

In a wireless IP networking scenario, each PCS (Personal Communications Services) network will consist of one or more IP-based data sub-networks. Each mobile data user that has a unique IP address may home to a sub-network within the PCS network. Data sub-networks within a PCS network will interwork with the Internet through some IWF (Interworking Function) or gateway (e.g., GGSN (Gateway GPRS Support Node) in GPRS networks). Data traffic among users in different PCS networks are transported using TCP/IP protocols and, in an 'all-IP' scenario, data packets within the PCS networks will be also routed by the 'IP-aware' BS (Base Station)/BSC (Base Station Switching Center)/MSC (Mobile Switching Center), based on the IP address of the destination MT (Mobile Terminal). Mobility management can be performed using the MIP (Mobile IP) protocol. MIP hides the movement of the mobile host to upper layer protocols and applications through the use of home and foreign agents that handle the routing of packets to the mobile host. In a wireless IP network, core network functionalities need to be enhanced to handle mobile multimedia networking functionalities (e.g., call control signaling using H.323[2]).

Some infrastructure design alternatives for wireless IP networks were discussed in [8]. With an aim of developing an 'all-IP' core network architecture to deliver 3G services, efforts are currently under way at 3GPP (3G Partnership Project) [9].

[2]Recommendation H.323 developed by the Telecommunications sector of the International Telecommunications Union describes the procedures for point-to-point and point-to-multipoint audio and video conferencing over packet-switched networks.

A brief description of the transmission protocol stack architecture of wireless IP networks follows. A simplified protocol architecture is shown in Figure 1.1.

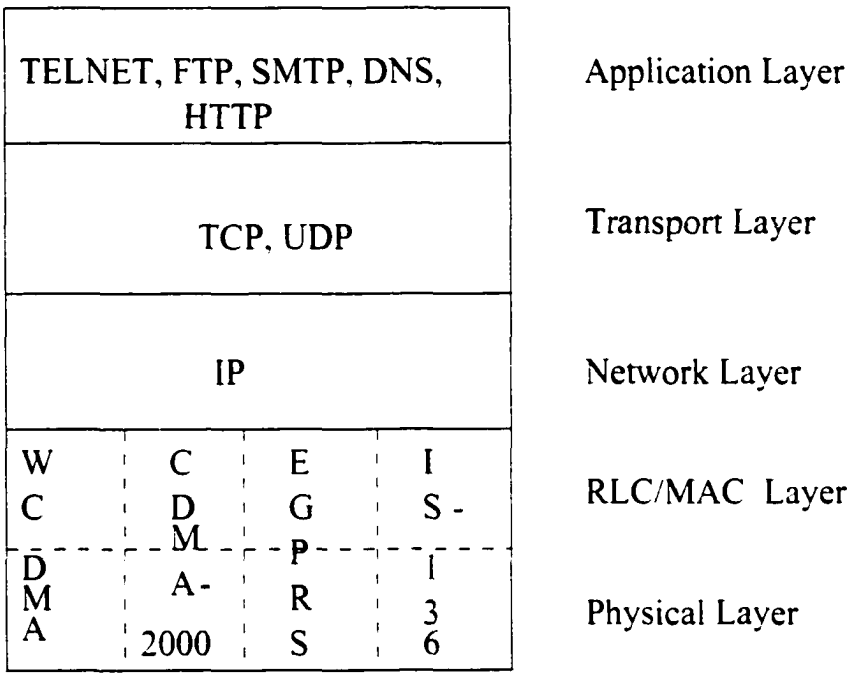

Figure 1.1. *The transmission protocol stack in a wireless IP network.*

Evolving wireless IP networks will have several air-interface technologies (e.g., WCDMA (Wideband Code Division Multiple Access), TDMA-based EDGE). The 'all-IP' core network will provide radio independent networking facility. Robust RLC (Radio Link Control)/MAC (Medium Access Control) protocol would provide efficient transport of real-time and non-real-time data traffic both in the uplink and the downlink.

The IP layer in the wireless IP protocol stack deals with the routing of IP packets. Extensions of MIP or some alternate solution may be required in this layer to handle vehicular mobility of data users in PCS networks. Functions such as TPC (Transmission Power Control) are also to be performed in this layer for CDMA-based physical layer.

In the wireless IP transmission protocol stack, the transport-layer protocol functions include end-to-end transmission control and error recovery (in the case of two-level error recovery), some convergence functions and some optimization functions

(e.g., TCP header compression).

1.3 Motivation and Scope of the Thesis

The motivation for this research flows from one main fact: the potential of a large market that stems from the increasing demand for wireless data services through the Internet. The emergence of private and public micro-cellular wireless networks are envisioned to provide a variety of broadband mobile services, including high-speed Internet access and audio/video delivery. Supporting multimedia applications along with the legacy wireless applications (e.g., cellular voice) within an integrated networking framework gives rise to new network infrastructure and protocol design issues that are to be resolved to facilitate migration to IMT-2000, the third generation wireless scenario and beyond. Achieving IP-centric solutions for wireless multimedia networks to provide reliability, scalability and QoS is a grand technical challenge.

This thesis encompasses several protocol design issues in the area of wireless packet networking technology for multimedia wireless communications. The major areas of this research that have been pursued are RLC/MAC and transport-layer protocols and the interrelations among mechanisms in different protocol layers. The aim of this research has been to accelerate the development of wireless mobile multiservice packet networks and to integrate them efficiently with the present Internet.

Multichannel protocols with advanced modulation and coding techniques are envisioned to provide high throughput wireless data services. Link-layer retransmission control (or access control) is an important issue in designing efficient RLC/MAC protocols (for uplink transmission control) in multichannel systems. Efficient retransmission control schemes that can provide high system capacity over a wide range of system load are to be designed for multichannel random access protocols taking the time varying channel impairments into consideration. This issue is addressed in this work. A retransmission control policy is proposed and analyzed for a multichannel S-ALOHA (Slotted ALOHA) protocol in a high speed wireless data network. With properly chosen protocol parameters, the proposed scheme can provide high system capacity for a wide range of system load.

Optimal dynamic rate allocation among mobile stations for variable rate packet

data transmission in a cellular wireless network is an NP-complete problem; therefore, sub-optimal solutions to this problem are sought for. Again, interference calculation is non-trivial in the case of a packet-switched cellular CDMA network with heterogeneous traffic load in different cells. In this thesis, a sub-optimal two-step dynamic rate selection procedure is proposed for uplink packet data transmission in cellular WCDMA networks. A novel 'mean-sense' approach for inter-cell interference calculation is employed assuming homogeneous traffic load in the different cells. Two different error control alternatives for this variable rate packet transmission environment are presented and their performances are analyzed for three different channel models.

Understanding the impact of different physical layer mechanisms (e.g., macrodiversity packet combining) on higher layer protocol performance would be required for wireless IP transmission protocol stack performance optimization. The relative performance of different link-layer access control schemes, in terms of RLC/MAC-layer and transport-layer throughput under retransmission diversity packet combining at the physical layer and different timer control mechanisms at the transport layer, is investigated for a CDMA S-ALOHA system. All of these enable us to better understand the issues involved in the design of higher layer protocols and to identify the suitable protocol mechanisms for CDMA-based wireless IP networks.

Fairness would be an important issue in the evolving wireless IP networks since wireless subscribers with similar levels of subscription would expect similar service from their WISPs (Wireless Internet Service Providers). For WISPs, this may not be simple due to time and location dependent channel errors in a wireless environment. This issue is addressed in this work. A TDMA-based centralized bandwidth management mechanism is proposed as a RLC/MAC-layer solution for providing fair bandwidth allocation among competing uplink traffic flows in a wireless IP network. It consists of an admission control policy, a traffic conditioning policy and a bandwidth allocation policy. The proposed scheme can be used in an adaptive QoS framework for dynamically adjusting the QoS of flows in order to accommodate wireless channel errors and user mobility.

Improved fairness among mobile subscribers in the case of downlink transmission (e.g., mobile stations downloading web documents) can be provided through

transport-level packet scheduling at the base station. Since TCP (Transmission Control Protocol) is the flagship protocol in today's Internet, the issue here is how to provide improved channel utilization while maintaining throughput fairness in the case of multiple TCP flows traversing the downlink. This issue is addressed through evaluation of the performances of different transport-layer packet scheduling schemes in a correlated fading environment with the revelation that a fair packet scheduling scheme, along with the dynamic adaptation of TCP parameters (e.g., maximum transmission window size) based on network delay and error characteristics, can enhance both channel utilization and throughput fairness in wide area wireless IP networks.

Dynamic bandwidth allocation at the RLC/MAC level for transporting multiservice traffic is one of the primary research issues for the wireless IP network designers. This issue is addressed here by evaluation of the performances of several centralized burst-level packet scheduling schemes for the transmission of voice, video and data traffic over TDMA (Time Division Multiple Access)/TDD (Time Division Duplex) channels. These schemes are based on the 'soft-reservation' of bandwidth for the different traffic types in each TDMA/TDD frame along with their instantaneous bandwidth requirements. Although the simulation experiments are based on ATM (Asynchronous Transfer Mode)-based transmission, the concepts here readily apply to wireless IP networks. Smaller RLC/MAC-layer frame size (e.g., ATM cell) provides fine-grain multiplexing, which leads to improved queueing delays and delay jitter; this is desirable for service integration in the wireless IP air-interface.

The literature review pertinent to each of the problems addressed in this dissertation will be presented in the relevant chapters separately.

1.4 Thesis Outline

Subsequent chapters of this dissertation are organized as follows:

- In Chapter 2, a retransmission control scheme for a multichannel S-ALOHA protocol in a micro-cellular wireless environment is proposed and analyzed. The delay-throughput performance is evaluated using exact Markov analyses for error-free channel and *i.i.d.* (independent and identically distributed) Rayleigh fading channel. For correlated Rayleigh fading channels, computer simulations

are used to evaluate performance.

- In Chapter 3, a sub-optimal dynamic link-level rate adaptation procedure is proposed for transmitting uplink packet data in VSF (Variable Spreading Factor) WCDMA-based cellular wireless IP networks. Two different error control alternatives for this variable rate packet transmission environment are presented and their performances are analyzed for three different channel models.
- In Chapter 4, the performance implications of retransmission diversity packet combining on link-layer and transport-layer protocol performance are investigated for three different heuristic-based link-layer access control schemes and two transport-layer timer control mechanisms in a CDMA S-ALOHA network under frequency selective Rayleigh fading.
- In Chapter 5, a unified TDMA-based centralized wireless access scheme is proposed for performing the statistical multiplexing of bursty data sources in a wireless packet data network. This scheme combines dynamic bandwidth allocation with admission control and packet conditioning (at the mobile stations) to provide fair bandwidth distribution among bursty data flows with different profile rates (or subscription levels) in an error-prone environment. The performance of the scheme is evaluated using computer simulations for different total subscription levels, for different compositions of flows with different profile rates and for different channel quality with different channel-error correlation pattern.
- In Chapter 6, performance evaluation of a number of TCP (Transmission Control Protocol) level packet scheduling policies is carried out in terms of channel utilization and throughput fairness (among similar competing TCP connections) in a wide area wireless IP network.
- In Chapter 7, a set of centralized burst-level cell scheduling schemes, namely, FCFS-FR (First Come First Served with Frame Reservation), FCFR-FR+, EDF-FR (Earliest Deadline First with Frame Reservation), EDF-FR+ and MTDR (Multitraffic Dynamic Reservation), are investigated for transmission of multiservice traffic over TDMA/TDD channels.
- Chapter 8 summarizes the contributions of the thesis and discusses a few directions for future research.

Chapter 2

Binary Feedback-Based Retransmission Control for a Multichannel S-ALOHA Protocol in Fading Channels

2.1 Introduction

Third-generation wide-area wireless data services (e.g., EDGE) are envisioned to provide a data traffic throughput of at least several hundred *Kbps* [4]. Multichannel networks with advanced modulation and coding techniques appear to be possible solutions for providing wireless data services with such a high throughput.

ALOHA-based single-channel wireless packet networks, such as CDPD, offer a throughput of less than 20 *Kbps* due to some intrinsic limitation in carrying high data rate traffic. Since the duration of each packet decreases as the data rate increases, it is necessary to increase the signal power to maintain a constant E_b (Energy per Bit) [10] so that the same BER (Bit Error Rate) can be achieved. Multiple ALOHA channels in parallel can be used to meet the high data rate requirements. To provide a high data rate, GPRS, which has been specified as part of GSM Phase 2+, enables a mobile station to use more than one timeslot from one or more physical channels [5].

In addition to providing a high data rate, a multichannel wireless network can function in a graceful degrading mode in case one or more channels fail due to harsh

channel conditions. In addition, in a distributed environment, prioritized access to different mobile stations can be provisioned by restricting their access to a different number of available parallel channels. For example, depending on the service requirement, a set of mobile stations can be prioritized to have access to a larger number of channels than another set of mobile stations. Network designers and service providers have to evaluate the network performance under different system parameters for proper dimensioning and allocation of system resources.

S-ALOHA has been widely adopted as a media access protocol and as a component in many reservation protocols in wireless networks. Due to the simplicity of the protocol, it is still of great interest to wireless network designers. Unless modified, some of the protocols, such as CSMA, CSMA/CD and their multichannel extensions ([11]-[12]), will not work in a wireless network due to the *'hidden terminal problem'*[1].

Like any other random access protocol, the satisfactory performance of a multichannel S-ALOHA protocol depends on its ability to adjust the protocol parameters, especially when the offered load to the network is near or above the capacity of the channels. Retransmission probability is the most appropriate parameter that can be varied if the network load changes. The selection of the retransmission probability is crucial for the satisfactory performance of the system. The use of an appropriate backoff scheme (local and/or global) can result in an improved delay-throughput performance compared to the performance of non-adaptive systems [13]. The choice of the retransmission probability as a function of the number of busy (i.e., either successful or suffering collision)/idle channels can lead to stable multichannel access control protocols.

A simple retransmission backoff scheme for single channel S-ALOHA networks was proposed in [14] and it was extended for single-hop WDM (Wavelength Division Multiplexing)[2] star-coupler networks [15]. Literature on retransmission probability selection in a multichannel finite population S-ALOHA system is very limited. A multichannel infinite population S-ALOHA protocol with constant probability retransmission control was studied in [16] and an optimizing criteria for stability was

[1]This refers to a situation where two mobile terminals, each of which is within the range of some intended third entity (e.g., BS), are out of range of each other.

[2]It is an approach for very high speed networking which divides the optical spectrum into many different wavelengths assigned to each channel.

formulated.

In a S-ALOHA based wireless network, frame[3] loss may occur due to the simultaneous transmission of more than one frame (unless any one of the simultaneously transmitted frames can 'capture' the receiver) and due to the channel fading. Therefore, in the case of high system load, transmission attempts from the mobile stations need to be controlled by delaying them appropriately in order to improve wireless channel utilization. Retransmission control policies can also exploit the channel fading information so that there are more frequent frame transmissions when the channel state is good and there are fewer transmission attempts when there are errors, which are typically bursty in nature. Delaying retransmissions rather than persisting on retransmissions under correlated channel fading is desirable from a power conservation viewpoint. The performance of a multichannel ISMA (Inhibit-Sense Multiple Access) protocol (where the mobile stations attempt transmissions in the uplink channels based on the busy/idle status of the channels in each slot broadcasted by the BS in the downlink) in the presence of bursty frame losses was studied in [17] for a simple 'persist-until-success' retransmission strategy.

In this chapter, we provide Markov analyses for the steady-state performance evaluation of a multichannel finite population S-ALOHA protocol with a generalized retransmission backoff scheme under *i.i.d.* Rayleigh fading. The retransmission control scheme is based on the binary feedback information about the most recent transmission attempts in the uplink channels which is broadcasted by the BS in the downlink channel(s). The incorporation of a capture model in the analyses enables us to get insight on the effects of large-scale path loss (and hence near-far effect) on the system performance. More interestingly, it makes analyses of a multichannel S-ALOHA system, without capture, simpler. The performance results for the single channel systems with constant probability retransmission control can be found from the general model as special cases. The steady-state performance of the proposed retransmission control scheme in correlated fading channels is evaluated with the aid of computer simulation. The performance results for a more power conservative version of the retransmission control policy based on local adaptation are also derived

[3]It refers to a unit data block at the RLC/MAC level. We prefer to use *frame* rather than *packet* since the latter is more appropriate for referring to higher layer data units.

through simulation. In addition, we discuss some other possible application areas of the backoff scheme and possible extensions of this work.

The organization of the rest of the chapter is as follows. The system model is described in Section 2.2. In Section 2.3, we present Markov analyses for the proposed retransmission control scheme. The performance results, obtained through computer simulation for the proposed scheme in fading channels, are presented in Section 2.4. In section 2.5, we analyze a variant of the proposed scheme based on local adaptation. Finally, in Section 2.6, the main points are summarized.

2.2 System Model and Description

A multichannel S-ALOHA system with N mobile stations and M parallel channels (i.e., carriers) is considered where each mobile station has access to all the channels (i.e., a mobile station can transmit to and receive from any of the channels). The time axis is considered to be divided into equal-length slots. The slots in the different channels operate in parallel. The mobile stations transmit or receive constant length frames and each frame fits into one timeslot. Each mobile station is assumed to have a transmit buffer of one frame size and the station remains blocked until the frame (if any) in the buffer is successfully transmitted through one of the channels in a timeslot.

Due to the *capture effect* in a cellular radio environment, the BS can decode a frame with normal accuracy even in the presence of collisions and channel fading if the signal power corresponding to the frame is sufficiently stronger than that of its contenders. Due to spatial attenuation, multipath reflections, diffractions and different spatial distributions in a mobile radio channel, the mobile stations are dynamically and randomly divided into different classes of access power. Rather than explicitly taking all these factors into consideration, we consider only the received power levels of the transmitted frames at the receiver.

Here, we assume that frames received at the BS can be classified into L classes (class 1 through class L, which are linearly spaced) based on their power levels where class 1 corresponds to the class with the lowest power and class L corresponds to the one with the highest power. The capture model considered here is as follows: during

a timeslot, the BS can decode a frame correctly if the corresponding power level is greater than that of each of the contending frames in that slot in the same physical channel. That is, a frame from mobile i, $i \in \{1, 2, \cdots, N\}$ is captured if $\Gamma_i > \Gamma_j$, $j = 1, 2, \cdots, N$, $j \neq i$, where N is the number of frames simultaneously transmitted in the same timeslot in the same physical channel in which mobile i is transmitting and Γ_k, $k = 1, 2, \cdots, N$ is the received power for the frame transmitted by mobile k. This is a more optimistic capture model than one where capture is assumed to take place if cumulative powers of the contending frames is less than the desired user power, namely, $\Gamma_i > \sum_{j=1, j \neq i}^{N} \Gamma_j$ [18].

We consider here three variations of the retransmission control policy based on the binary feedback about the channel status: *MBFS-1* (Multichannel Binary Feedback-1), *MBFS-2* (Multichannel Binary Feedback-2) and *MBFS-3* (Multichannel Binary Feedback-3). *MBFS-1*, *MBFS-2* and *MBFS-3* use the *success/non-success*, *idle/non-idle* and *collision/non-collision* feedback, respectively. Non-collision feedback corresponds to both success and idle. Each blocked mobile station updates the value of the retransmission probability according to the feedback received in the previous timeslot. A mobile station newly generating frames can synchronize to other mobile stations if we assume that the BS broadcasts the current value of the retransmission probability at the end of each timeslot.

The round trip propagation delay is assumed to be less than the slot duration so that a mobile station can learn the status of a channel (i.e., whether the transmission is successful, a collision has occurred, or the channel has been idle) in a slot before the beginning of the next slot. This is a reasonable assumption for micro-cellular wireless networks. We also assume that the feedback information from the BS is transmitted error-free in the downlink channel(s). This is also reasonable since the small number of bits conveying the feedback information can be adequately protected through channel coding. The performance results obtained with the assumptions of instantaneous and error-free feedback serve as an upper bound for the system performance.

Let $p(t)$ be the retransmission probability in the tth timeslot. For the three variations of the retransmission control scheme, $p(t)$ is updated as follows:

$$\text{(i) } \textit{MBFS-1}\text{: } p(t+1) = \begin{cases} min\,(p_{max},\ p(t)/q), & \text{if } M_{s(t)} \geq \lfloor x \times M \rfloor \\ max\,(p_{min},\ p(t) \times q), & \text{if } M_{s(t)} < \lfloor x \times M \rfloor \end{cases}$$

$$\text{(ii) } MBFS\text{-}2\text{: } p(t+1) = \begin{cases} min\,(p_{max},\ p(t)/q)\,. & \text{if } M_{i(t)} \geq \lfloor x \times M \rfloor \\ max\,(p_{min},\ p(t) \times q)\,. & \text{if } M_{i(t)} < \lfloor x \times M \rfloor \end{cases}$$

$$\text{(iii) } MBFS\text{-}3\text{: } p(t+1) = \begin{cases} max\,(p_{min},\ p(t) \times q)\,. & \text{if } M_{c(t)} \geq \lfloor x \times M \rfloor \\ min\,(p_{max},\ p(t)/q)\,. & \text{if } M_{c(t)} < \lfloor x \times M \rfloor \end{cases}$$

where $0 < p_{max} \leq 1$, $0 < q < 1$, $0 < x \leq 1$, $p_{min} = q^r$ and $r \in I$ $(r > 0)$. For $M = 1$, the value of the parameter x is 1 for all the three cases. Here, $p(t+1)$ is the retransmission probability in slot $(t+1)$. $M_{c(t)}$ is the number of channels suffering collision in slot t. $M_{s(t)}$ is the number of successful transmissions in slot t and $M_{i(t)}$ is the number of channels that remained idle in slot t.

The protocol parameters x and q determine the rate of adaptation in the value of the retransmission probability. In a particular system configuration, for a particular value of x, higher values of q cause slow adaptation whereas lower values of q result in high rate of adaptation.

A DFT (Deferred First Transmission) mode is considered here, i.e., after the generation of a new frame, a mobile station immediately goes into blocked mode. We assume p_{max} to be equal to 1.

2.3 Performance Evaluation for Error-Free Channels by Markov Analysis

2.3.1 Capacity Analysis

The normalized system capacity, C, is defined as the number of successfully transmitted frames per slot per channel when all the mobile stations have frames to transmit during each timeslot. In other words, C is the throughput per channel under full load condition. The system states under full load condition can be described by the discrete-time Markov chain shown in Figure 2.1 where the system state X_r is chosen such that the value of the retransmission probability p_r is q^r in that state.

The state transition probabilities are given by

(i) For *MBFS-1*.

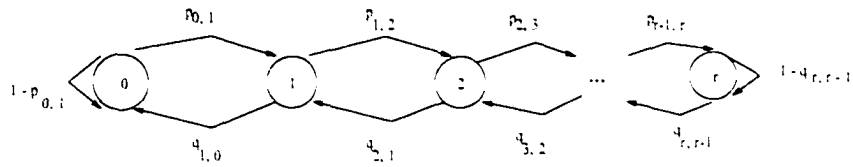

Figure 2.1. *Markov chain showing the system states under full load condition.*

$$p_{i,i-1} = Pr\{M_{s(t)} < \lfloor x * M \rfloor | X_t\} = \sum_{j=0}^{\lfloor x*M-1 \rfloor} \binom{M}{j} \cdot p_{succ(i)}^{j} \cdot (1 - p_{succ(i)})^{M-j} \quad (2.1)$$

where $p_{succ(i)}$ is given as follows:

$$p_{succ(i)} = \sum_{k=1}^{N} \binom{N}{k} \cdot \left(\frac{q^i}{M}\right)^k \cdot \left(1 - \frac{q^i}{M}\right)^{N-k} \cdot r_k. \quad (2.2)$$

In (2.1), r_k is the probability that a frame is successfully decoded at the BS when k frames are transmitted simultaneously in a slot in the same physical channel. This is given by (2.3)

$$r_k = I_A + \frac{k}{L^k} \cdot \sum_{i=1}^{L} (i-1)^{k-1}. \quad L > 0 \quad (2.3)$$

where I_A is an indication function defined as follows:

$$I_A = \begin{cases} 1 & \text{if } L = 1 \text{ and } k = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (2.4)$$

(ii) For *MBFS-2*.

$$p_{i,i-1} = Pr\{M_{i(t)} < \lfloor x * M \rfloor | X_t\} = \sum_{j=0}^{\lfloor x*M-1 \rfloor} \binom{M}{j} \cdot p_{idle(i)}^{j} \cdot (1 - p_{idle(i)})^{M-j} \quad (2.5)$$

where $p_{idle(i)}$ is given as follows:

$$p_{idle(i)} = \left(1 - \frac{q^i}{M}\right)^N. \quad (2.6)$$

(iii) For *MBFS-3*.

$$p_{i,i-1} = Pr\{M_{c(t)} \geq \lfloor x * M \rfloor | X_t\} = \sum_{x*M}^{M} \binom{M}{j} . p_{col(t)}^{j} . (1 - p_{col(t)})^{M-j} \quad (2.7)$$

where $p_{col(t)}$ is given as follows:

$$p_{col(t)} = \sum_{k=2}^{N} \binom{N}{k} . \left(\frac{q^t}{M}\right)^k . \left(1 - \frac{q^t}{M}\right)^{N-k} . (1 - r_k) . \quad (2.8)$$

For all the cases.

$$q_{i,i-1} = 1 - p_{i-1,i}. \quad (2.9)$$

Let Π and $\boldsymbol{P}$ denote the limiting probability vector and the transition probability matrix, respectively. Then, the limiting probabilities can be derived using (2.10).

$$\Pi = \Pi.P \quad and \quad \sum_{0}^{r} \Pi_i = 1. \quad (2.10)$$

The normalized system capacity (C) can be determined using (2.11)

$$C = \sum_{i=0}^{r} \Pi_i . \sum_{k=1}^{N} \binom{N}{k} . \left(\frac{q^t}{M}\right)^k . \left(1 - \frac{q^t}{M}\right)^{N-k} . r_k. \quad (2.11)$$

The delay is the average number of slots that a frame has to wait until it is successfully transmitted. It can be calculated using Little's formula as follows [16]:

$$D = \frac{N}{C.M}. \quad (2.12)$$

The results for the single-channel S-ALOHA can be found as special cases of the above formulation by putting $M = 1$. For the non-adaptive case, with the constant probability of retransmission being equal to p_r, the normalized system capacity can be determined directly from (2.11) by substituting p_r, 1 and 0 for q^t, Π_i and r, respectively. In the above analysis, we avoid using the general formula for determining the probability of s successful captures of channels when i stations are contending for m idle channels, as given by (2.13), which has been used in some previous papers (e.g., [12], [16]). Thus the computations here become relatively simpler.

$$p_s[i, m, s] = \frac{(-1)^s . i! . m!}{m^i . s!} . \sum_{v=s}^{min(i,m)} \frac{(-1)^v . (m-v)^{i-v}}{(v-s)! . (i-v)! . (m-v)!}.$$
$$0 \leq s \leq min(m, s). \quad (2.13)$$

With a fixed value of M, N and r, typical variations of C with q and x for different values of L are shown as contour plots in Figures 2.2, 2.3 and 2.4 for *MBFS-1*, *MBFS-2* and *MBFS-3*, respectively. For brevity, the corresponding delay plots are omitted. It is to be noted that, in a particular system configuration, for a certain value of C, the average frame delay (D) is inversely proportional to C, as given in (2.12).

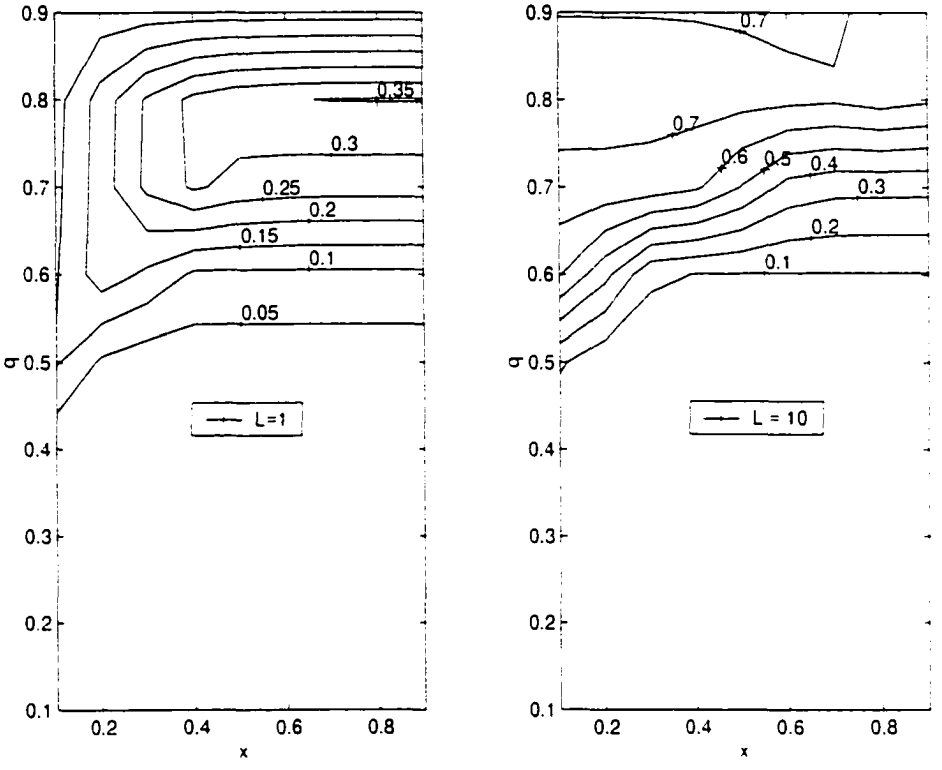

Figure 2.2. MBFS-1: *Variation of normalized system capacity with q and x (for N = 100, M = 10, r = 9).*

For *MBFS-1*, with smaller values of q, C deteriorates significantly regardless of the value of x. With a fixed value of x, as q increases, C first increases until it reaches some threshold, then decreases with increasing q. This threshold depends on the ratio of N/M and L. For larger values of L, the dropoff in C is more graceful with increasing q (Figure 2.3). When N/M is relatively small and L is relatively high, for a certain value of x, C monotonically increases with increasing q. For a fixed value of q, as x increases, C first increases and then decreases with increasing x. For larger values of L, this threshold shifts towards smaller values of x. For very large values of

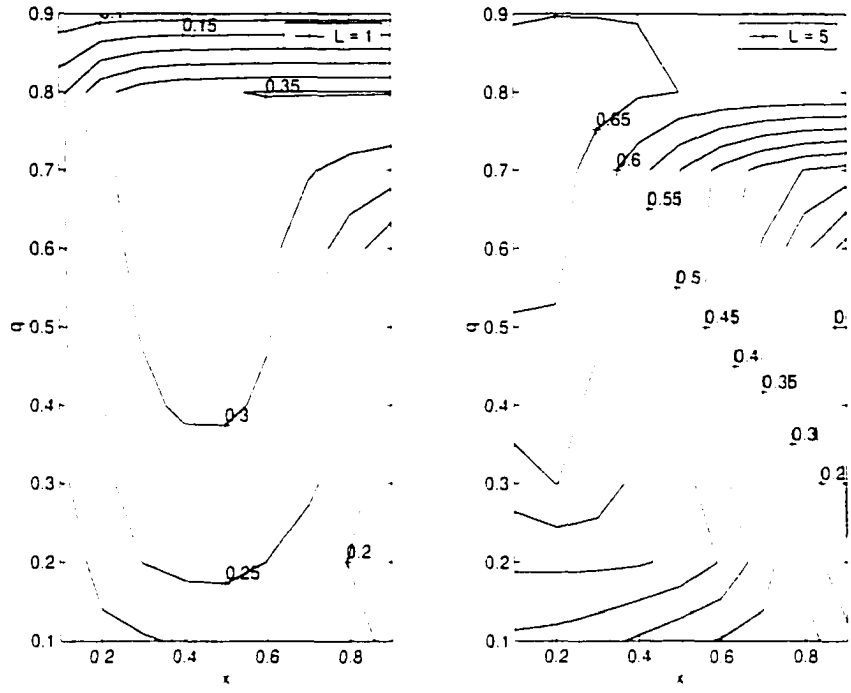

Figure 2.3. MBFS-2: *Variation of normalized system capacity with q and x (for N = 100, M = 10, r = 9).*

q (e.g., $q = 0.9$). C does not change appreciably with a change in the value of x.

For *MBFS-2*, as q increases, with a fixed value of x, C first increases and then decreases with increasing q after C reaches some threshold value. This threshold depends on the ratio of N/M and L. For larger values of L, similar to *MBFS-1*, the dropoff in C is more graceful with increasing q (Figure 2.4) and for relatively small N/M and high L, for a certain x, C monotonically increases with increasing q. For moderately low values of q, as x increases, C first increases, remains almost same for a range of values of x and then decreases with increasing x. For larger values of q and L, this threshold shifts towards smaller values of x; that is, at higher values of L, smaller values of x give better normalized system capacity when N/M is relatively small. However, as x increases, for very high values of q (e.g., $q = 0.9$), initially C increases and then does not change appreciably with increasing x. The variation of normalized capacity with x and q for *MBFS-3* is similar to the variation for *MBFS-2* (Figure 2.4).

For all three cases, the system capacity improves as L increases. For a fixed M,

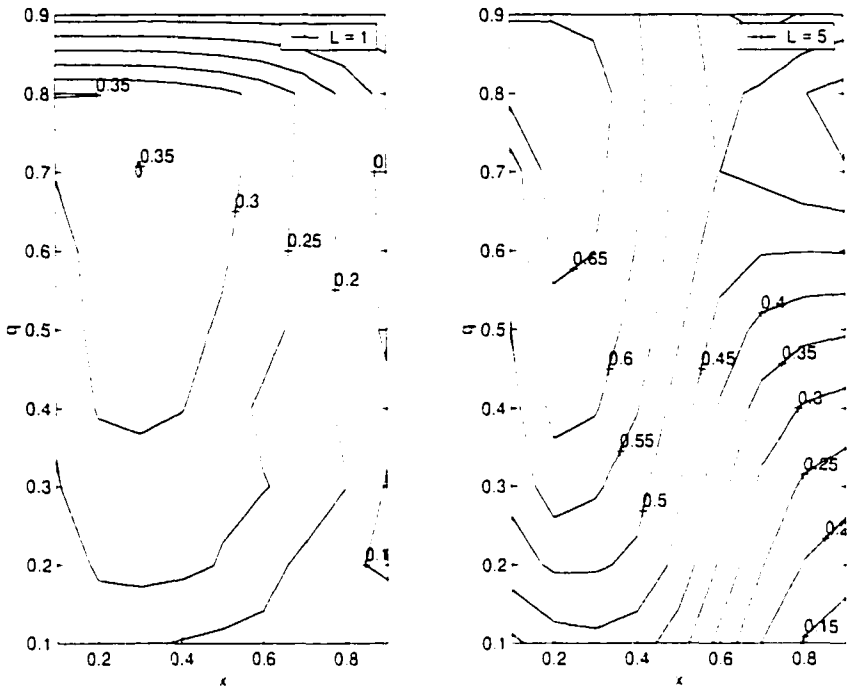

Figure 2.4. MBFS-3: *Variation of normalized system capacity with q and x (for N = 100, M = 10, r = 9).*

higher values of L provide better stability in system capacity when N is large. The rate of improvement in system capacity declines as L increases. For a particular multichannel system configuration, the parameters x and q may be chosen to have a steady-state normalized system capacity as large as 70% for values of L not larger than 10.

For particular values of N/M and L, the values of x and q that lead to maximum normalized system capacity can be found through search. For a given number of parallel channels (i.e., number of parallel timeslots in different careers) the three protocol parameters r, q and x can be chosen such that the channel capacity is stable for a wide range of values of N. For $M = 20$, $r = 5, 9, 15$ and $L = 1$ (i.e., no capture), the values of x and q that lead to the maximum normalized system capacity (i.e., e^{-1}) are listed in Table 2.1, Table 2.2 and Table 2.3 for *MBFS-1*, *MBFS-2* and *MBFS-3*, respectively. For an increase in the number of users, q needs to be decreased (i.e., the rate of adaptation is to be increased) to achieve the maximum system capacity.

The system capacity is found to be more sensitive to the variation in q than to the

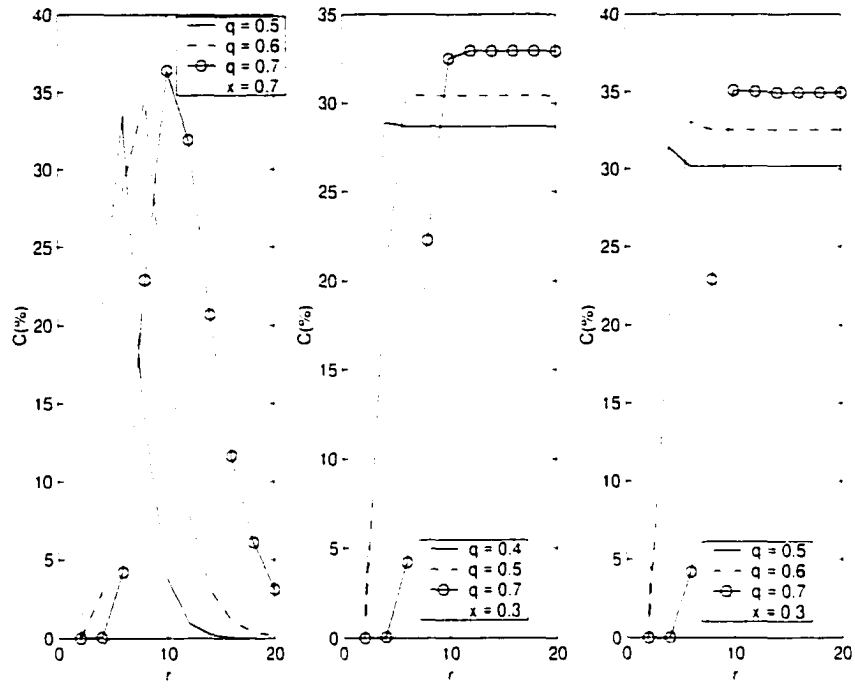

Figure 2.5. MBFS-1, MBFS-2 and MBFS-3: *Effect of r on normalized system capacity (for $N = 400$, $M = 10$).*

variation in x. At maximum system capacity, the difference among *MBFS-1*, *MBFS-2* and *MBFS-3* lies in the selection of x. For particular values of N/M and r, q is almost the same in all three cases for maximum normalized system capacity. If r is changed, the corresponding values of q change but the maximum system capacity can be achieved with the same value of x. For a certain N/M, if r is increased (decreased), q is to be increased (decreased) to achieve the maximum capacity.

For fixed x, q and r, the variations in the normalized system capacity C with the number of active mobiles are shown in Figure 2.7 for the three retransmission backoff schemes. For a fixed value of x, as q increases beyond a threshold, the system tends to become unstable in the sense that, for a fixed number of channels, the system capacity falls off sharply with an increase in the number of active mobiles. The system shows the most stable performance with $q = 0.5$ with normalized system capacity as high as 30% (without capture). The throughput in this case is slightly lower than the maximum system capacity $1/e$. Numerical results show that in a typical micro-cellular wireless network scenario with $N/M \leq 30$, normalized system capacity of

Table 2.1. *The values of q and x for which C is maximum under different N/M (for* MBFS-1 *with $M = 20$, $L = 1$).*

N/M	5	10	15	20	25	30	35	40	45	50
x	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
q ($r = 5$)	0.70	0.65	0.60	0.55	0.53	0.50	0.48	0.47	0.46	0.45
q ($r = 9$)	0.85	0.78	0.75	0.72	0.70	0.68	0.67	0.65	0.65	0.65
q ($r = 15$)	0.90	0.85	0.83	0.82	0.80	0.80	0.78	0.78	0.77	0.77

Table 2.2. *The values of q and x for which C is maximum under different N/M (for* MBFS-2 *with $M = 20$, $L = 1$).*

N/M	5	10	15	20	25	30	35	40	45	50
x	0.9	0.95	0.95	0.9	0.95	0.95	0.9	0.9	0.9	0.9
q ($r = 5$)	0.70	0.65	0.60	0.55	0.53	0.50	0.48	0.47	0.46	0.45
q ($r = 9$)	0.84	0.77	0.74	0.72	0.70	0.68	0.68	0.66	0.66	0.64
q ($r = 15$)	0.90	0.85	0.83	0.82	0.81	0.80	0.78	0.78	0.78	0.77

more than 30% can be achieved for some other value of q (e.g., 0.7) even without capture. As has already been pointed out in [14], choosing the value of q to be 0.5 offers the advantage of implementation simplicity since the value of the retransmission control parameter can then be calculated using a simple shift register.

For fixed N/M, q and x, with *MBFS-1*, as r increases, C increases, reaches some maximum value and then decreases with increasing r (Figure 2.5). As q is increased, this maximum shifts towards higher values of r. But for *MBFS-2* and *MBFS-3*, C becomes almost constant after r reaches some threshold. As q is increased, this threshold shifts towards larger values of r.

2.3.2 Throughput Analysis

Under low load condition, a mobile station, which is not blocked, generates a frame with probability p_g ($0 < p_g < 1$) for transmission during a timeslot (whereas under full load condition $p_g = 1$). The normalized system throughput (S) is defined as the number of frames successfully transmitted per channel during a timeslot. The system

Table 2.3. *The values of q and x for which C is maximum under different N/M (for* MBFS-3 *with $M = 20$, $L = 1$).*

N/M	5	10	15	20	25	30	35	40	45	50
x	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
q $(r = 5)$	0.70	0.65	0.60	0.55	0.53	0.50	0.48	0.47	0.46	0.45
q $(r = 9)$	0.84	0.77	0.74	0.72	0.70	0.69	0.67	0.66	0.66	0.65
q $(r = 15)$	0.90	0.85	0.83	0.82	0.81	0.80	0.78	0.78	0.78	0.77

under low load condition can be described by a Markov process with each state (i, j) chosen such that, in this state, the probability of retransmission is q^i and the number of backlogged mobiles is j. Then, the possible state transitions are the following:

$(i, j) \rightarrow (i + 1, k)$, *for* $0 \leq j, k \leq N, 0 \leq i < r$
$(i, j) \rightarrow (i - 1, k)$, *for* $0 \leq j, k \leq N, 1 \leq i \leq r$
$(i, j) \rightarrow (i, k)$, *for* $0 \leq j, k \leq N, i \in \{0, r\}$.

Let X_i denote the event that the retransmission probability is q^i. The state transition probabilities can then be defined as follows:

(i) $P_{(i,j),(i-1,k)} = p_1 \times p_2$, $0 \leq j, k \leq N$, $0 \leq i < r$.

For *MBFS-1*, p_1 is same as the p_i defined in (2.1), i.e.,

$$p_1 = Pr\{M_{s(t)} < \lfloor x * M \rfloor | X_i\} = \sum_{j=0}^{x*M-1} \binom{M}{j} . p_{succ(i)}^j . (1 - p_{succ(i)})^{M-j}$$

and $p_{succ(i)}$ is given by

$$p_{succ(i)} = \sum_{k=0}^{N-j} \binom{N-j}{k} . p_g^k . (1 - p_g)^{N-j-k} . \sum_{l=1}^{j-k} \binom{j+k}{l} . \left(\frac{q^i}{M}\right)^l . \left(1 - \frac{q^i}{M}\right)^{j-k-l} . r_l. \quad (2.14)$$

For MBFS-2, p_1 is same as the p_i defined in (2.5), i.e.,

$$p_1 = Pr\{M_{i(t)} < \lfloor x * M \rfloor | X_t\} = \sum_{j=0}^{x*M-1} \binom{M}{j} . p_{idle(i)}^{j} . (1 - p_{idle(i)})^{M-j}$$

where $p_{idle(i)}$ is given by

$$p_{idle(i)} = \sum_{k=0}^{N-j} \left[\binom{N-j}{k} . p_g^k . (1-p_g)^{N-j-k} . \left(1 - \frac{q^t}{M}\right)^{j-k} \right]. \tag{2.15}$$

For *MBFS-3*, p_1 is same as the p_i defined in (2.7), i.e.,

$$p_i = Pr\{M_{c(t)} \geq \lfloor x * M \rfloor | X_t\} = \sum_{j=\lfloor x*M \rfloor}^{M} \binom{M}{j} . p_{col(i)}^{j} . (1 - p_{col(i)})^{M-j}$$

where $p_{col(i)}$ is given by

$$\begin{aligned} p_{col(i)} \quad &= \sum_{k=0}^{N-j} \binom{N-j}{k} . p_g^k . (1-p_g)^{N-j-k} . \\ &\sum_{l=1}^{j-k} \binom{j+k}{l} . \left(\frac{q^t}{M}\right)^l . \left(1 - \frac{q^t}{M}\right)^{j-k-l} . (1 - r_l). \end{aligned} \tag{2.16}$$

For all the three cases, $p_2 = Pr\{$*Number of backlogged mobiles changes from* j to $k \mid X_t\}$

$$\begin{aligned} &= \sum_{s=0}^{N-j} \binom{N-j}{s} . p_g^s . (1-p_g)^{N-j-s} . \\ &\left\{ \sum_{l=j-s-k}^{j-s} \binom{j+s}{l} . q^{tl} . (1-q^t)^{j+s-l} . p[l, M, L, j+s-k] \right\} \end{aligned} \tag{2.17}$$

where $p[l, M, L, j+s-k]$ is the probability of $j+s-k$ successful transmission(s) when l mobiles are contending in M channels with L linearly spaced power levels at the receiver. This is given by

$$p[l, M, L, j+s-k] = \binom{l}{j+s-k} . p_{(l,M,L)}^{j-s-k} . \left(1 - p_{(l,M,L)}\right)^{l-j-s}. \tag{2.18}$$

Here, $p_{(l,M,L)}$ is the probability of successful transmission in a channel for L linearly spaced power levels in the receiver when l stations contend in M channels, and it is given by (Appendix A)

$$p_{(l,M,L)} = \frac{1}{L} \cdot \sum_{i=1}^{L} \left[1 + \frac{i - (1+L)}{ML} \right]^{l-1} . \quad L > 0. \tag{2.19}$$

(ii) $P_{(i,j),(i-1,k)} = (1 - p_1) \times p_2$, $0 \leq j,k \leq N$, $0 < i \leq r$, where p_1 and p_2 are as defined above.

(iii) $P_{(0,j),(0,k)} = (1-p_1) \times p_2$ where p_1 is as defined above and p_2 is defined as follows:

$$p_2 = \sum_{l=0}^{N-j} \binom{N-j}{l} . p_g^l . (1-p_g)^{N-j-l} . p[j+l, M, L, j+l-k].$$

(iv) $P_{(r,i),(r,j)} = p_1 \times p_2$, where p_1 and p_2 are as defined above for *MBFS-1*, *MBFS-2* and *MBFS-3*.

Let Π and $\boldsymbol{P}$ denote the limiting probability vector and the transition probability matrix, respectively. Then, the limiting probabilities can be derived using (2.20) and hence the normalized system throughput (S), mean backlog (B) and mean frame delay (D) can be determined using (2.21), (2.22) and (2.23), respectively.

$$\Pi = \Pi . P \quad and \quad \sum_{i=0}^{r} \sum_{j=0}^{N} \Pi(i,j) = 1. \tag{2.20}$$

$$S = \sum_{i=0}^{r} \sum_{j=0}^{N} \Pi(i,j) . \sum_{l=0}^{N-j} \binom{N-j}{l} . p_g^l . (1-p_g)^{N-j-l} .$$
$$\sum_{m=0}^{j-l} \binom{j-l}{m} . \left(\frac{q^i}{M}\right)^m . \left(1 - \frac{q^i}{M}\right)^{j-l-m} . r_m . \tag{2.21}$$

$$B = \sum_{i=0}^{r} \sum_{j=0}^{N} j . \Pi(i,j). \tag{2.22}$$

$$D = \frac{B}{S.M}. \tag{2.23}$$

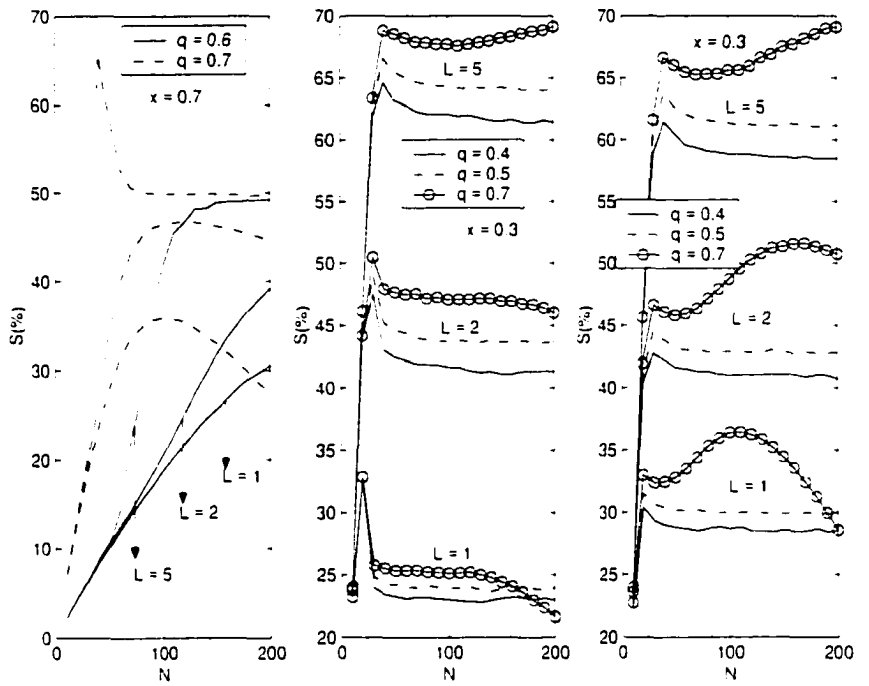

Figure 2.6. MBFS-1, MBFS-2 and MBFS-3: *Variation of normalized system throughput with N (for $M = 5$, $r = 9$, $p_g = 0.1$).*

The results for the single channel case without capture can be obtained by letting $M = 1$ and $L = 1$ in the above formulation. Typical variations in the normalized throughput with the number of mobile stations are shown in Figure 2.6. It is evident from the figure that the capture effect substantially improves throughput. For $N = 200$, $M = 5$, $q = 0.7$ and $x = 0.3$, with $L = 5$, for *MBFS-2* and *MBFS-3*, the normalized throughput is about 69% for frame generation probability of 0.1, whereas without capture (i.e., for $L = 1$), it is only about 21% (for *MBFS-2*) and 27% (for *MBFS-3*). For the same parameters, in the case of *MBFS-1*, $S = 30\%$, 39% and 49% with $L = 1$, 2 and 5, respectively.

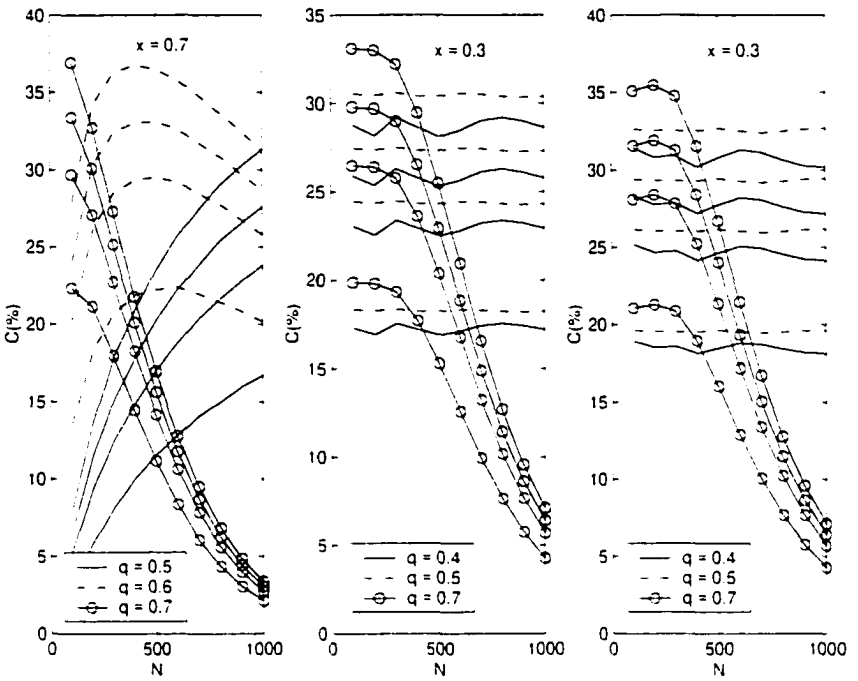

Figure 2.7. MBFS-1, MBFS-2 and MBFS-3: *Variation of C with N (for M = 10, L = 1, r = 9, P_E = 0.0, 0.1, 0.2, 0.4).*

2.4 Performance Evaluation in Fading Channels

2.4.1 Fading Channel Model

For a broad range of parameters, the frame (i.e., a data block of duration T) error process in a mobile radio channel, where the multipath fading can be considered to follow a Rayleigh distribution, can be modeled using a two-state Markov chain with transition probability matrix M_c as follows ([19]-[20]): $M_c = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}$ where p and 1-q are the probabilities that the jth frame transmission is successful, given that the (j-1)th frame transmission was successful or unsuccessful, respectively. This two-state Markov channel model has been used for wireless protocol performance evaluation in fading channels. For example, throughput predictions of protocols such as ARQ (Automatic Repeat Request) and TCP using two-state Markov channel models were found to match quite well with those found by simulations over fading channels [21].

Given the matrix M_c, the channel properties are completely characterized. Markov processes describing different channels will be independent of each other. The steady-state probability of frame error is given by $P_E = \frac{1-p}{2-p-q}$ and the average length of an error burst is $(1-q)^{-1}$. Also, for a Rayleigh fading channel with a *fading margin*[4] F, $P_E = 1 - e^{-1/F}$ and the Markov parameter $q = 1 - \frac{Q(\theta,\rho\theta)-Q(\rho\theta,\rho)}{e^{1/F}-1}$ where $\theta = \sqrt{\frac{2/F}{1-\rho^2}}$. Here, $Q(\ldots)$ is Marcum's Q function, $\rho = J_0(2\pi f_d T)$ is the Gaussian correlation coefficient (where J_0 is the zero order Bessel function of the first kind) of two successive samples of the complex amplitude of a fading channel with maximum Doppler shift f_d (where $f_d = v/\lambda$ = mobile speed/carrier wavelength) taken T seconds apart.

We assume that during the time period T, the fading envelope does not change significantly. For high rate data transmission, this assumption of a *slowly varying channel*[5] is reasonable for the usual values of the carrier frequency (900-1800 MHz). The parameter f_dT is the normalized Doppler bandwidth and it is directly related to the correlation in the fading process. From P_E and q in the above equations, the Markov parameter p can be obtained. Different P_E and f_dT values will result in different degrees of correlation in the fading process. When f_dT is small, the fading process is very correlated (i.e., there will be long bursts of frame errors); on the other hand, for large values of f_dT, the channel error events are almost independent (i.e., there will be short bursts of frame errors). For a certain value of average frame error rate P_E (particularly for a value in the range of $O(10^{-1})$ to $O(10^{-3})$), the error burst length increases (i.e., the fading process becomes more correlated) as user mobility (and hence f_dT) decreases. Other parameters remaining the same, as P_E increases, the error burst length increases rapidly for low-mobility users.

It is assumed that when the channel is in bad state, frames are lost regardless of the corresponding power levels at the receiver. This is a pessimistic assumption since even under some fading condition a frame with higher received power may be decoded correctly by the BS. When the channel is in good state, the frame with the highest received power level is assumed to be successfully decoded by the receiver.

In the case of *MBFS-1*, a channel is declared to be successful by the BS if the channel status is good and transmissions on that channel are not garbled due to

[4]It is the maximum fading attenuation which still allows correct reception of a frame.

[5]It refers to a channel condition where all the bits of the same frame experience approximately the same fading conditions.

multiple simultaneous transmissions (i.e., at least one frame is received with a power level higher than each of the other contending frames). In this way, the channel status information is exploited in the retransmission control policy. Similarly, for *MBFS-3*, a channel is declared to be suffering collision if transmissions on that channel are garbled due to multiple simultaneous attempts and/or the channel is in bad state. But in the case of *MBFS-2* (i.e., for idle/non-idle feedback), channel error status information does not affect the retransmission control directly although the system delay-throughput performance is affected due to the channel impairments. Under these assumptions, the system performance metrics for the three variations of the retransmission control scheme can be evaluated for *i.i.d.* Rayleigh fading by redefining r_k in (2.3) as in (2.24) and by changing (2.2), (2.8), (2.11), (2.14) and (2.16) accordingly.

$$r_k = e^{\frac{-1}{F}} \cdot \left(I_{A} - \frac{k}{L^k} \cdot \sum_{i=1}^{L} (i-1)^{k-1} \right), \quad L > 0 \tag{2.24}$$

where F is the fading margin for the corresponding channel.

For high mobility users, the *i.i.d.* Rayleigh fading assumption is reasonably good [19]. But for high data rate services with low user mobility, the channel fading events are correlated rather than independent. Since deriving closed form equations for performance evaluation in correlated fading channel would be quite involved, we resort to using computer simulation.

A *C*-coded time-driven simulator is used for obtaining the performance results. The simulator run-time is chosen to be 600000 *timeslots*, which corresponds to 5 *hrs* of simulation time assuming the frame size to be 75 *bytes* and channel data rate to be 20 *Kbps* (i.e., slot duration = 0.03 *s*). Simulation statistics are flushed after 100000 *slots* to avoid the transient effects. In the case of correlated fading, all the channels are assumed to be characterized by the same fading parameters but they suffer burst errors independently. We derive simulation results for a wide range of values of normalized Doppler bandwidth to show the effects of correlated fading on protocol performance for different degrees of channel-error correlation.

2.4.2 Performance Results

The variations in normalized system capacity under *i.i.d.* Rayleigh fading without capture are shown in Figure 2.7 for P_E = 0.0, 0.1, 0.2 and 0.4. P_E = 0.1, 0.2 and 0.4 correspond to F = 9.6 dB, 6.6 dB and 2.9 dB, respectively. For *MBFS-2* and *MBFS-3*, even with channel fading margin of only 6.6 dB, the normalized system capacity is about 25% which is 68% of the maximum achievable capacity without capture (i.e., e^{-1}). With M = 10 and N = 400, for *MBFS-1*, the normalized system capacity C is about 28% for the same fading margin.

The performance results for correlated channel fading are shown in Figure 2.8 for P_E = 0.1 and 0.2 with the normalized Doppler bandwidth f_dT equal to 0.0084, 0.0252, 0.5004 and 1.5003. The results for *i.i.d.* fading are also included in the same plot for comparison. With a carrier frequency of 900 *MHz* and frame duration of 0.03 *s* (which corresponds to a frame size of 75 *bytes* for a 20 *Kbps* channel), f_dT = 0.0084, 0.0252, 0.5004 and 1.5003 correspond to an MT speed of 0.336 *km/hr*, 1 *km/hr*, 20 *km/hr* and 60 *km/hr*, respectively. For P_E = 0.1, with these values of f_dT, the average length of error bursts (in frames) are 14.25622, 5.4593, 1.12245 and 1.11486, respectively. For P_E = 0.2, the corresponding error burst lengths are 13.21135, 8.40766, 1.27541 and 1.25840 frames, respectively.

Simulation results show that, for typical values of x and q (e.g., x = 0.7, q = 0.5, 0.6, 0.7 for MBFS-1 and x = 0.3, q = 0.4, 0.5, 0.7 for MBFS-2), system capacity under correlated fading is almost the same as that due to *i.i.d.* fading. Simulation results also show that, for *MBFS-3*, the effects of channel error correlations on the normalized system capacity are small when the protocol parameters (e.g., r, q) are properly chosen. For a certain value of r, the value of the retransmission control parameter q can affect the capacity degradation in the case of correlated fading. For example, it is observed that for MBFS-3, with relatively poorer channel quality (e.g., F = 6.6 dB), as f_dT decreases (i.e., channel error correlation increases), capacity degradation is more graceful for q = 0.5 compared to the degradation for q = 0.4 (Figure 2.8), while for q = 0.7, the values of C are almost the same as those in the case of *i.i.d.* errors.

With properly chosen protocol parameters, all of the three variations of the retransmission control scheme can effectively avoid the channel error correlations and

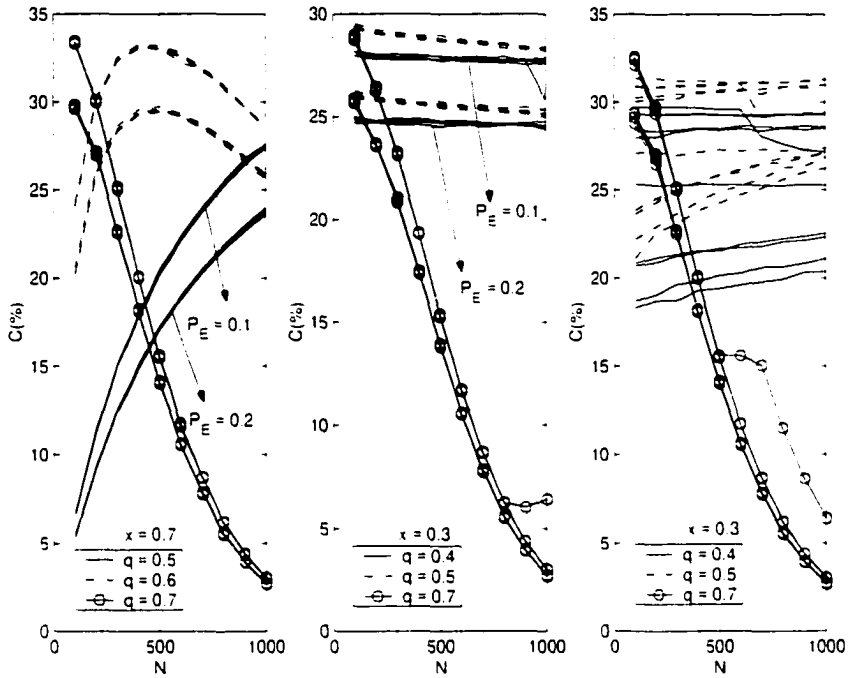

Figure 2.8. MBFS-1, MBFS-2 and MBFS-3: *Variation of C with N under correlated fading (for M = 10, L = 1, r = 9, P_E = 0.1, 0.2 and f_dT = 0.0084, 0.0252, 1.5003).*

provide considerably high capacity even without capture. Moreover, antenna diversity and combining schemes, when employed at the BS, would improve system capacity further.

The results presented above have been derived assuming that, when the channel is in bad state, transmitted frames are lost regardless of the corresponding power levels at the receiver. This assumption, as has been stated earlier, is a bit pessimistic since the receiver may be able to decode a frame correctly even when the channel state is bad, depending on the received power-level, modulation, coding, diversity and interference statistics. Some simulation results are shown in Figure 2.9 to demonstrate how this assumption affects the performance results. Simulation results are obtained for the different values of the parameter p_s, which denotes the probability of successful reception of a frame in the case of channel fading when the corresponding power level is higher than that of each of the contending frames. As is evident from the results, for a particular system configuration, C improves as the value of p_s is increased. This improvement is significant at relatively higher values of p_s (and at higher system load

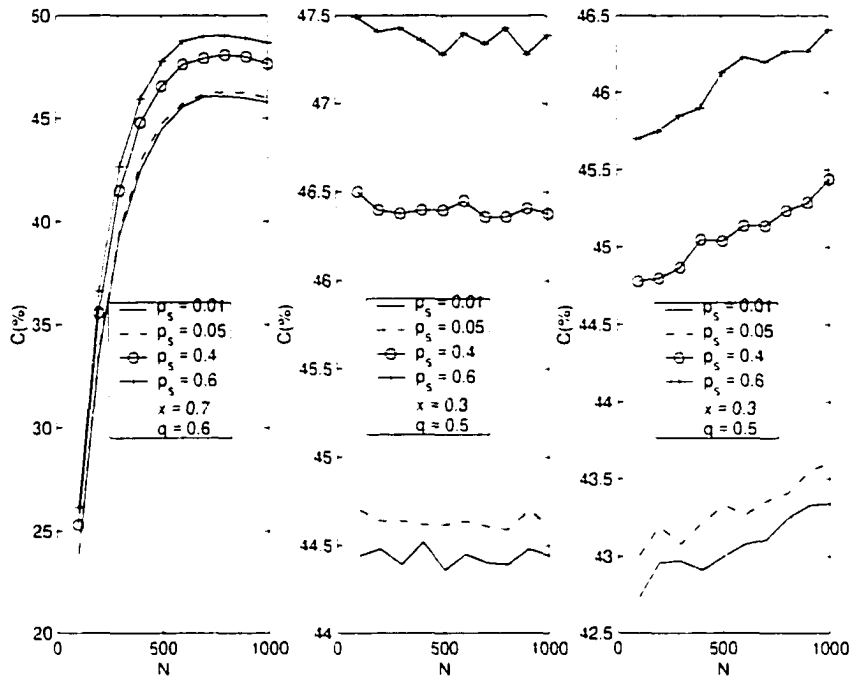

Figure 2.9. MBFS-1, MBFS-2 and MBFS-3: *Variation of C with N under correlated fading for different values of p_s (for $M = 10$, $L = 2$, $r = 9$, $P_E = 0.1$ and $f_dT = 1.5003$).*

in the case of *MBFS-1*). From the simulation results for different values of q, L and for different values of the normalized Doppler bandwidth, we observe that this trend is independent of the values of q, L and channel-error correlation pattern.

2.5 Local Adaptation Based on Binary Feedback

Up to now, we have assumed a global adaptation strategy where all the active mobile stations synchronously adapt the retransmission control probability. In such a case, every active mobile has to 'snoop' the feedback channel at each timeslot to receive the channel status information broadcasted by the BS at the beginning of each timeslot (unless the BS itself computes the value of the retransmission probability and broadcasts it at the beginning of each timeslot). In an energy-constrained environment this mode of operation may not be desirable. Again, if there are errors in the feedback channel, some of the mobiles may not be able to adapt their retransmission control

probability synchronously with other mobiles; therefore, a local adaptation of the retransmission control parameter may be more desirable.

In this section, we analyze the performance of the binary feedback-based backoff schemes described above, assuming local adaptation at each of the mobile stations. In this variation, after the RLC/MAC layer receives a data frame to transmit, p_r is set to 1 (i.e., the mobile station transmits the frame at the beginning of the next slot). This mode of operation is called *IFT* (Immediate First Transmission). If the transmission is not successful, based on the feedback information, it updates p_r according to the backoff scheme (i.e., *MBFS-1/MBFS-2/MBFS-3*) and switches to more power-conservative standby mode. At the same time, it starts a timer based on the current value of p_r. The mobile station switches to transmit mode after the timer expires. While in transmit mode, the mobile station attempts retransmission and the above process repeats.

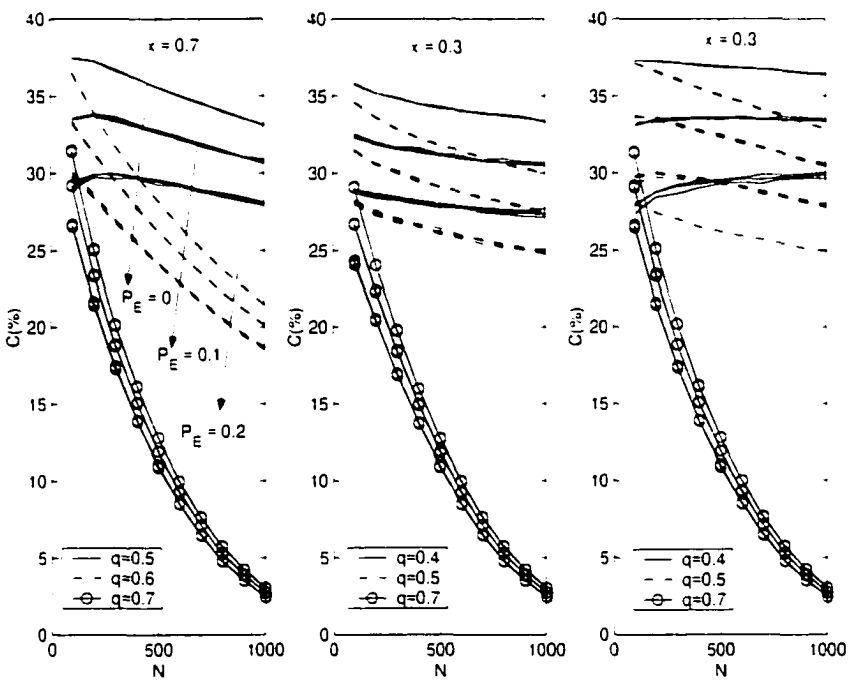

Figure 2.10. MBFS-1, MBFS-2 and MBFS-3: *Variation of C with N in correlated fading channel under local adaptation (for $M = 10$, $L = 1$, $r = 9$, $P_E = 0.0, 0.1, 0.2$, $f_dT = 0.0084, 0.0252, 0.5004, 1.5003$).*

The variations in the normalized system capacity with the number of mobile

terminals (N) are shown in Figure 2.10 for different channel quality with different degree of channel-error correlation. Simulation results reveal that, in the case of local adaptation, all the three variations of the backoff scheme perform better in fading channels for an appropriate choice of protocol parameters. With channel fading margin of 9.6 dB, for all the three cases, $q = 0.5$ provides a normalized channel capacity of more than 30% over a large range of values of N/M. The normalized channel capacity is more than 27% for a channel fading margin as low as 6.6 dB even when normalized Doppler bandwidth is as small as 0.0084. The retransmission backoff scheme can effectively avoid the channel error correlations and provide fairly high system capacity over a large variation of N/M.

2.6 Chapter Summary

In this chapter, we have evaluated the performance of three variations of a global adaptation-based truncated exponential backoff retransmission control policy for a multichannel S-ALOHA protocol in micro-cellular networks in terms of capacity, throughput and delay in Rayleigh fading channels by exact Markov analyses. We have also introduced a capture model in the analysis. The performance results for single channel systems with non-adaptive retransmission control in an error-free channel and an *i.i.d.* Rayleigh fading channel can be found as special cases from the analytical model. System performance results under correlated Rayleigh fading have been obtained through simulation.

For a broad range of system load, a stable system capacity as large as 30% can be achieved even without capture for error-free channels. Simulation results reveal that in a typical micro-cellular environment, where the ratio of the number of active mobiles and the number of available parallel channels is moderately high (e.g., for $N/M \leq 30$), even higher system capacity can be achieved with a retransmission control policy based on idle/non-idle or collision/non-collision feedback. The capture effect can provide better capacity with improved stability. For a particular system configuration, the normalized system capacity without capture has been found to be equal to e^{-1} for proper selection of q and x. For the assumed channel model, the protocol parameters can be chosen to have a stable normalized system capacity as

large as 70% of the maximum achievable capacity for a channel fading margin of 6.6 dB (which corresponds to average frame error rate of 20%) even without capture.

Simulation results show that for properly chosen protocol parameters, normalized system capacities for correlated frame errors are almost same as for *i.i.d.* frame errors. The three variations of the binary feedback-based retransmission control scheme are found to perform equally well (and even better) in the case of local adaptation at the mobile stations. The local adaptation-based retransmission control policy can result in a more power-conservative RLC/MAC protocol.

Non-adaptive CDMA S-ALOHA systems are inherently unstable with a large population of terminals ([37]-[38]). A similar type of retransmission control scheme can be devised for CDMA S-ALOHA networks to combat the throughput degradations at large offered loads. This adaptive retransmission control scheme may be used also for TDD-TDMA/TDD-CDMA-based channel access.

Chapter 3

Integrated Rate and Error Control in WCDMA-Based Cellular Wireless IP Networks

3.1 Introduction

Adaptive packet access schemes based on dynamic rate and error control can provide superior system performance resulting from higher radio resource utilization. WCDMA systems (e.g., ETSI WCDMA, cdma2000) which will be the major radio transmission technologies for IMT-2000, have intrinsic support for dynamic rate transmission, which can be achieved through variable processing gain ([22]-[23]). In a WCDMA network, the transmission rate of a mobile station can be controlled on a frame by frame basis. Fast and simple algorithms for dynamic rate adaptation based on the channel status (which is a function of channel load and error characteristics) are to be developed for cellular WCDMA systems. The role of rate adaptation in a WCDMA system is analogous to the role of dynamic link adaptation in narrowband systems (e.g., EDGE [4])–the aim is to maximize the overall network throughput which is constrained by channel fading, interference and noise. In fact, dynamic rate adaptation can increase the statistical multiplexing gain with a consequent increase in instantaneous system throughput.

A time-slotted system is considered here in which the slot duration is fixed. Depending on the rate selection however a mobile station can transmit a variable number of fixed length RLC/MAC-layer frames within one timeslot, which are derived

from the segmentation of variable-length IP packets. In such an environment, the RLC/MAC-level error recovery subsequent to some transmission failure (in the case of non-transparent mode of operation) may be combined with the rate control mechanism. Again, for a hybrid-ARQ type of error recovery there are several possible design alternatives as will be described later in this chapter.

Dynamic rate and error control can be implemented in a centralized, distributed, or hybrid fashion. In the centralized case the BS will dynamically assign transmission rates to the mobile stations. The crux of the problem in this case is to sense the channel load and channel condition and to subsequently determine the rate combination. In fully distributed dynamic rate adaptation, each mobile station can choose a transmission rate independently; the problem here is to determine proper adaptation criteria and a suitable method of rate adaptation. Hybrid schemes can be devised for uplink packet data transmission where the BS can assist the mobile stations in performing the dynamic rate adaptation. One such scheme will be described later in this chapter.

An adaptive variable transmission rate random access scheme based on DS-CDMA S-ALOHA in an AWGN environment was proposed in [24] although no rate selection procedure (other than exhaustive search) was assumed in the presented analytical model. An adaptive multiple access scheme for uplink admission/congestion control was proposed in [25] for variable spreading gain CDMA systems operating in an AWGN environment. The scheme is based on the adaptation of the retransmission control probability for the users with different but fixed transmission rates. The problem of dynamic rate control was not addressed in [25]. In fact, dynamic rate control and adaptive retransmission control can work complementarily to enhance the uplink radio spectrum efficiency.

The problem of dynamic rate allocation for optimal channel throughput (which will be described later) is NP-complete[1]. Therefore, fast and efficient sub-optimal solutions to this problem are sought for. Here, a sub-optimal two-step rate selection procedure for uplink traffic control is presented, taking into account the interference conditions present in a cellular WCDMA network, for three different channel models.

[1]The family of NP (Non-deterministic Polynomial)-complete problems includes many important optimization problems which are all inherently of exponential time complexity and algorithms for solving them in less than exponential time do not exist.

The performance of the scheme is compared to that of an optimal rate allocation scheme which is based on exhaustive search. A 'mean-sense' approach is used for inter-cell interference calculation under variable rate packet data transmission. A Markov analytical model is developed to evaluate the RLC/MAC-layer performance with dynamic rate selection (in the ideal case) for two different hybrid ARQ-based error control strategies. To this end, a BS-assisted dynamic rate adaptive medium access control scheme is presented.

The rest of the chapter is organized as follows. The dynamic rate adaptation problem is described in Section 3.2. Two alternatives for retransmission-based link-level error control strategies for packet-switched cellular WCDMA systems with variable rate transmission are also presented in this section. The different channel models assumed in this chapter are described in Section 3.3. The proposed two-step rate search procedure is presented in Section 3.4. System performance under dynamic rate and error control is evaluated in Section 3.5 using a Markov model and considering constant probability retransmission control. A BS-assisted distributed medium access control scheme employing dynamic rate adaptation is presented in Section 3.6. Conclusions are stated in Section 3.7.

3.2 Dynamic Rate and Error Control in WCDMA Systems

3.2.1 Dynamic Rate Allocation Problem

In a cellular packet-switched WCDMA environment, variable rate data transmission is expected to utilize the time-varying nature of the traffic intensity in a cell in order to enhance the radio spectrum efficiency. With variable rate packet transmission, the number of RLC/MAC-layer frames to be sent in a timeslot is varied depending on the transmission rate used in that timeslot.

Let us assume that the length of a slot, denoted by T_s, is fixed and that the transmission rates can be selected from the set of rates $\{v_1, v_2, \ldots, v_\varsigma\}$. If it is assumed that $v_m = mv_1$ $(m = 2, \ldots, \varsigma)$ [2], and that only one frame (of fixed length of M

[2]For the rest of the chapter it is assumed that $v_m = mv_1$ so that the normalized value of v_m with

bits) corresponding to the smallest (basic) rate v_1 can be sent in a slot, then for a transmission rate of v_m, m frames can be sent in a timeslot. For this, a variable spreading gain WCDMA system can be used where the basic gain is given by N chips per bit, and for rate v_m, the spreading gain is reduced to N/m chips per bit. Note that the signal energy per bit (E_b) is kept constant to achieve the same bit error rate (BER) performance. The best performance can be obtained if the users, transmitting simultaneously, choose the optimum combination of the transmission rates.

In the case of optimum rate selection, the throughput ($\mathcal{J}$) is maximized, i.e.,

$$\max_{\{n_1,n_2,\ldots,n_r\}} \mathcal{J}(n_1, n_2, \ldots, n_r) \quad \text{subject to} \quad n_1 + n_2 + \ldots + n_r = n \tag{3.1}$$

where n_m denotes the number of users choosing rate v_m ($m = 1, 2, \ldots, r$), given the total number of n simultaneous users in a slot. Here, the throughput ($\mathcal{J}$) can be expressed by

$$\mathcal{J}(n_1, n_2, \ldots, n_r) = \sum_{m=1}^{r} m n_m P_{c,m} \qquad \text{frames/slot} \tag{3.2}$$

where the probability of correct reception of a frame $P_{c,m}$ is dependent on the physical layer BER (Bit Error Rate) performance (i.e., $P_{b,m}$) (and hence on the channel interference and fading conditions) and on the error control protocol employed. Channel interference conditions depend largely on the dynamically selected transmission rates in the different cells.

It is not difficult to observe that the computational complexity of the exhaustive search for determining the optimal rate allocation (i.e., the rate allocation vector $\mathbf{n} = \{n_1, n_2, \ldots, n_r\}$) in a system with available transmission rates $\mathbf{v} = \{v_1, v_2, \ldots, v_r\}$ would be of $O(n^r)$. Since finding the optimal solution requires cycling through all possible assignments which would involve exponential time complexity, the problem of optimal rate selection (which is similar to the *satisfiability problem* ([26], p. 673)) is NP-complete. Therefore, it is unlikely that a true optimal algorithm for this problem (other than the exhaustive search) exists. Hence, in a practical scenario, a fast and efficient procedure may be desirable, which may not necessarily result in true optimal rate allocation but in sub-optimal rate allocation.

---

respect to the basic rate v_1 is m.

3.2.2 Realization of Variable Rate Transmission

Variable rate transmission in a WCDMA system can be realized either by multi-code transmission [27] or by single-code transmission with variable spreading gain. The latter, which is being considered here, is preferable for uplink transmissions due to its simplicity in implementation and potential for low power consumption at the mobile stations [28].

For single-code transmission, where a user changes the spreading factor according to the transmission rate, mutually orthogonal codes for different transmissions are possible if the rates are constrained by

$$r = \frac{r_c}{2^n}, \quad n = 0, 1, \cdots \tag{3.3}$$

where r_c is the highest transmission rate possible, which, presumably, would be limited by some minimum required spreading gain. This results in a variable rate-resolution which becomes very coarse for relatively higher transmission rates. For a constant rate resolution, non-orthogonal spreading codes must be used as a consequence of which the intra-cell interference will be increased.

3.2.3 Error Control Alternatives

Two different error control schemes, referred to as *selective-repeat (SR)* and *Go-Back-m (GBm)* schemes, are presented here for variable rate packet data transmission. In the former, each frame sent in a timeslot is treated independently so that only the frames involved in errors will be retransmitted. Therefore, error recovery in this case is similar to error recovery in the classical selective-repeat ARQ scheme although the transmission protocol here is more complicated due to the variable rate transmission. On the other hand, the latter assumes a whole frame decoding rather than an individual frame decoding within a timeslot. That is, the m frames transmitted in a timeslot (with transmission rate r_m) are treated as one single frame of length mM bits (Figure 3.1). Therefore, when the whole frame is declared to be failed after decoding, all the m frames in the timeslot will be retransmitted. This scheme is different from the classical *go-back-N ARQ* scheme since the number of retransmitted frames is determined only by the rate adopted (i.e., N is variable in this case rather than fixed).

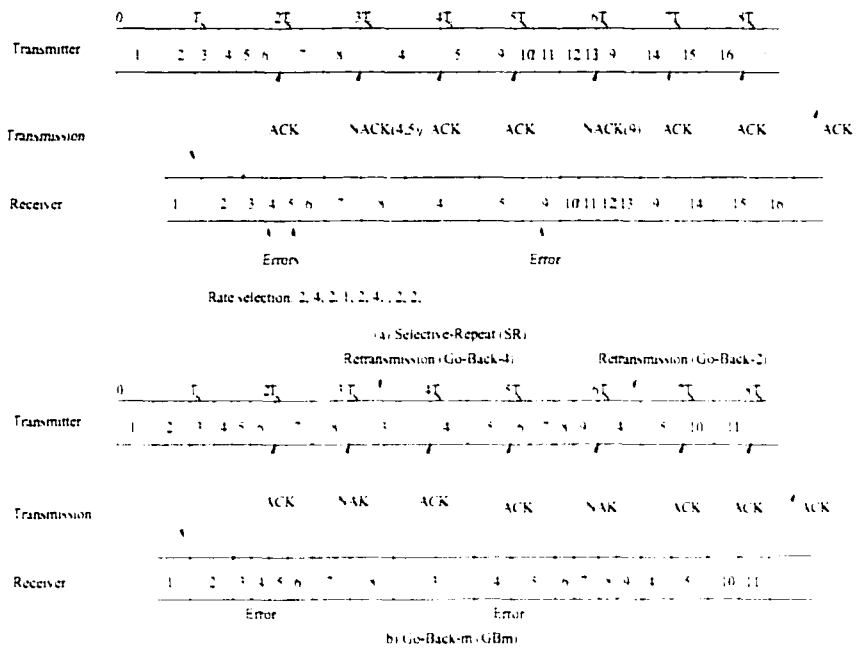

Figure 3.1. *ARQ-based error control in variable rate transmission systems.*

To illustrate, flow diagrams of the proposed error control protocols are depicted in Figure 3.1 for transmissions from a single user assuming that the acknowledgement delay is T_s and that the data rate v_m can be selected from the set $\{v_1, v_2, v_4\}$. Here, a heavy traffic condition is assumed so that newly generated frames are always available which can be transmitted along with the old frames (which are being retransmitted) so that the variable rate transmission can be fully exploited from slot to slot. For a multi-user case, it is necessary that an access control scheme is used to randomize other user interferences, and hence a random delay would be inserted between successive transmissions from a mobile station using some retransmission control scheme.

Under these two error control schemes, the probability of correctly receiving a frame $P_{c,m}$ for a user transmitting at rate v_m can be expressed as

$$P_{c,m} = \begin{cases} \sum_{e=0}^{t} \binom{M}{e} P_{b,m}^{e}(1-P_{b,m})^{M-e} & \text{for } SR \\ \sum_{e=0}^{t_m} \binom{mM}{e} P_{b,m}^{e}(1-P_{b,m})^{mM-e} & \text{for } GBm \end{cases} \tag{3.4}$$

where $P_{b,m}$ is the BER for a user with rate v_m and $\binom{k}{l} \triangleq k!/(k-l)!/l!$. Here, $t(t_m)$-error correction capability is assumed for a frame of length $M(mM)$ bits. To compare

the performances of these two error control protocols fairly, the code rates are to be kept the same. For a certain value of t, the code rate r satisfies the Varsharmov-Gilbert lower bound [29] as given by (3.5)

$$r \geq 1 - H\left(\frac{d_{min} - 2}{M}\right) \tag{3.5}$$

where $H(x) = -x\log_2 x - (1-x)\log_2(1-x)$ and $d_{min} \geq 2t + 1$. Therefore, for a certain minimum value of r (corresponding to $d_{min} = 2t + 1$), d_{min} and hence t_m corresponding to a frame of length mM, can be determined (at least approximately) using (3.5) with M replaced by mM.

3.3 Channel Models

In order to observe the impact of channel quality on the dynamic rate selection and hence on the system performance under different error control alternatives, the following three different channel models are considered here:

- Single path channel with no fading
- Multipath Rayleigh fading channel with equal path gain and
- Multipath Rayleigh fading channel with unequal path gain.

3.3.1 Single Path Channel with No Fading

In this case, the effective signal-to-interference plus noise ratio $(SINR)_{o,m}$ experienced by a user with rate v_m, is [30]

$$(SINR)_{o,m} = \left[\frac{x - m}{3N} + \frac{\eta\alpha G}{3N} + \left(\frac{2E_b}{N_o}\right)^{-1}\right]^{-1}. \tag{3.6}$$

Here, x is the effective number of frames per timeslot and is defined as follows:

$$x \triangleq n_1 + 2n_2 + \ldots + mn_m + \ldots + \tau n_\tau \tag{3.7}$$

where the rates $v_2, \ldots, v_\tau$ are normalized by the basic rate v_1 and η represents the frequency reuse factor (or outer-cell interference coefficient) reflecting the inter-cell interference in cellular WCDMA environment, assuming a homogeneous load across

different cells with an offered average load of G attempts/slot/cell. Here, G can also be interpreted as the mean of the loads across different cells. That is, although the traffic load in different cells may vary, a mean load G (averaged over all the cells) is assumed to be maintained across all cells. This may be achieved through RNC (Radio Network Controller) in a WCDMA network by using a suitable load/power management mechanism across different cells. Note that $\alpha \stackrel{\cdot}{=} \mathbf{E}\{r_m\}/r_1$ ($\mathbf{E}$ means the expectation) accounts for the rate adaptation in other cells and can be expressed as (Appendix B)

$$\alpha \cong \frac{x}{G} \qquad \text{subject to } n = G \tag{3.8}$$

while the effect of traffic channels carrying voice is reflected by the effective bit $(SNR)_o = (2\frac{E_b}{N_o})$. $\frac{E_b}{N_o}$ = ratio of bit energy and AWGN noise spectral density.

The BER $P_{b,m}$ corresponding to the rate r_m is a function of $(SINR)_{o,m}$ and for given values of inter-cell interference and background noise, as expressed by the second and third terms in (3.6), it is denoted by $P_b(x-m)$ here and is given by

$$P_b(x-m) = Q\left(\sqrt{(SINR)_{o,m}}\right) \tag{3.9}$$

for $Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-u^2/2}\, du$.

From now onwards, the probability of the successful reception of a frame $P_{c,m}$ corresponding to $P_{b,m}$ (as in (3.4)) will be denoted by $P_c(x-m)$.

3.3.2 Multipath Rayleigh Fading Channel with Equal Path Gain

In this case, an L-path Rayleigh fading channel with uncorrelated scattering and equal average path power is considered. With this model, the effective per-path $(SINR)_{c,m}$ for a user with rate r_m can be expressed as

$$(SINR)_{c,m} = \left[\frac{x \cdot L - m}{3N} + \frac{\eta \alpha L G}{3N} + \left(\frac{2E_b/L}{N_o}\right)^{-1}\right]^{-1}. \tag{3.10}$$

Here, the BER $P_{b,m}$ is expressed as $P_b(x \cdot L - m)$, and for L-diversity order it is given by [31]

$$P_b(x \cdot L - m) = \left[\frac{1}{2}(1-\zeta)\right]^L \sum_{l=0}^{L-1} \binom{L-1+l}{l} \left[\frac{1}{2}(1+\zeta)\right]^l \tag{3.11}$$

where $\zeta = \sqrt{(SINR)_{c,m}/[2 + (SINR)_{c,m}]}$.

3.3.3 Multipath Rayleigh Fading Channel with Unequal Path Gain

For multipath Rayleigh fading with unequal average path power, the effective $(SINR)_{l,m}$ for the lth path ($l = 1, 2, \ldots, L$) can be expressed by (3.10), with L replaced by $\rho_l = 1/\mathbf{E}\{a_l^2\}$, where a_l is the Rayleigh-faded path gain and $\sum_{l=1}^{L} \mathbf{E}\{a_l^2\} = 1$ (normalized).

In this case, the BER can be expressed as [31]

$$P_b(x \cdot \rho_1 - m, x \cdot \rho_2 - m, \ldots, x \cdot \rho_L - m) = \frac{1}{2} \sum_{l=1}^{L} \sigma_l \left[1 - \sqrt{\frac{\gamma_l}{1 + \gamma_l}} \right] \qquad (3.12)$$

where $\sigma_l = \prod_{\substack{l'=1 \\ l' \neq l}}^{L} \gamma_l/(\gamma_l - \gamma_{l'})$ with $\gamma_l = \frac{1}{2}(SINR)_{l,m}$.

3.4 A Two-Step Rate Selection Procedure

In this section, a two-step sub-optimal rate selection procedure is presented. The performance of this scheme is compared to that of the optimal rate selection scheme. The efficacy of controlled rate selection over uncontrolled random rate selection is also demonstrated.

3.4.1 Description of the Procedure

A two-step rate selection procedure is proposed here for a specified rate vector $\mathbf{v}$ assuming n simultaneous users. The idea here is to determine the value of x (as defined previously) for which the throughput is maximized. Firstly, the value of the rate (e.g., v_m) is determined for which the maximum throughput per timeslot (as given by (3.2)) can be achieved when all of the mobile stations select the same transmission rate. The throughput is calculated here in terms of the number of basic RLC/MAC frames (each of length M bits) successfully transmitted per timeslot.

Using the notations for BER and frame success probability, as introduced previously, the RLC/ MAC layer throughput $\mathcal{J}$ in the case of common rate selection by all

the users can be expressed as follows:

$$\begin{aligned}
n_1 &= n \longrightarrow \mathcal{J}(n_1) = nP_c(n-1) \\
n_2 &= n \longrightarrow \mathcal{J}(n_2) = 2nP_c(2(n-1)) \\
&\cdot \qquad\qquad\qquad\qquad \cdot \\
n_m &= n \longrightarrow \mathcal{J}(n_m) = mnP_c(m(n-1)) \\
&\cdot \qquad\qquad\qquad\qquad \cdot \\
n_r &= n \longrightarrow \mathcal{J}(n_r) = rnP_c(r(n-1)).
\end{aligned}$$

Let us denote m_o as the optimum selection in the first step when all the stations choose the same rate. Therefore, the first step of the rate selection procedure (*course search*) can be described as

Procedure First_Step_Rate_Selection() {
  $m_o = m_- =$ lowest available rate (e.g., r_1)
  select next higher rate m
  while $(P_c(mn^*) \geq \left(\frac{m_-}{m}\right) P_c(m_- n^*))$ {
    $m_o = m$
    select next higher rate m
  }
  return m_o
}

where m_- denotes the rate nearest to m ($m = 2, \ldots, r$) with $m_- < m$ and $n^* = n-1$. Since $mP_c(mn^*)$ increases first with an increase in m and then decreases as m is increased, an optimum selection of m can be made easily by determining the largest m that maximizes $mP_c(mn^*)$. Note that the value of x corresponding to this optimum rate m_o is nm_o. As is evident, the complexity of this step is linear with respect to the size of the rate vector $\mathbf{v}$.

The second-step of the rate selection procedure (*fine search*) is performed based on the following:

$$\max_{\{n_1, n_2, \ldots, n_r\}} \left\{ \mathcal{J}(n_1, n_2, \ldots, n_r) = \sum_{m=1}^{r} mn_m P_c(x-m) \right\} \quad \text{subject to}$$
$$n_1 + n_2 + \ldots + n_r = n \quad \text{and} \quad x \in (m_{o_-} n,\ m_{o_-} n) \qquad (3.13)$$

for the nearest rates $m_{o_-} < m_o$ and $m_{o_-} > m_o$.

This step basically searches for a value of x (where $x \in (m_{o\_}n,\ m_{o\_}n)$) that offers a higher value of $\mathcal{J}$ than that obtained during the *course search* step. This step proceeds by assigning rate $m_{o\_}$ (or $m_{o\_}$ depending on the *search direction*[3] to obtain higher $\mathcal{J}$) to an increasing fraction of stations out of n until $\mathcal{J}$ increases monotonically. The highest achieved value of $\mathcal{J}$ is denoted by $\mathcal{J}^*$. For a fixed $\mathbf{v}$, the complexity of this step is linear with respect to n (i.e., $O(n)$). A C-like implementation code for the second step of the search procedure is given in Appendix C. In this particular implementation, the set of rates chosen is either $\{m_{o\_}, m_o\}$ or $\{m_o, m_{o\_}\}$ although sets with a wider range of rates may be obtained by some additional computations (i.e., repeating the procedure for some other rates) with the order of computational complexity remaining the same.

Since $P_c(x - m) \approx \mathbf{E}\{P_c(x - m)\} \approx P_c(x - \alpha)$ (using the Jensen's inequality [32]) $\mathcal{J}$ in (3.13) can be well approximated to

$$\mathcal{J}(n_1, n_2, \ldots, n_{\_}) \approx xP_c(x - \alpha) \tag{3.14}$$

which implies that, for $x \in (m_{o\_}n,\ m_{o\_}n)$, $\mathcal{J}$ is more likely to assume a global maximum. Therefore, the proposed two-step rate selection algorithm obtains a near-optimum selection of x and hence a sub-optimal solution $\{n_1, n_2, \ldots, n_{\_}\}$ for a given n.

It is to be noted that in the case of the multipath Rayleigh fading channel model with equal path gain, $n^* = n \cdot L - 1$ during *course search* and x would be replaced by $x \cdot L$ during *fine search*. In the case of unequal path gain, the looping condition in the first step of the rate selection procedure would be read as

$$P_c(mn_1^*, mn_2^*, \ldots, mn_L^*) \geq \left(\frac{m\_}{m}\right) P_c(m\_n_1^*, m\_n_2^*, \ldots, m\_n_L^*). \tag{3.15}$$

where $n_l^* = n \cdot \rho_l - 1$. Also, the second step (*fine search*) is based on (3.13) with $P_c(x - m)$ replaced by $P_c(x \cdot \rho_1 - m, x \cdot \rho_2 - m, \ldots, x \cdot \rho_L - m)$.

3.4.2 Performance Results

The variations in the value of the first step rate selection (m_o) with the number of active users (n) for different average offered load (G) are observed for different values

[3]The direction refers to whether decrease or increase x.

of η (e.g., $\eta = 0.0, 0.3, 0.5$). $\frac{E_b}{N_o}$ (e.g., $\frac{E_b}{N_o} = 10, 15, 20\ dB$) and t (e.g., $t = 2, 4, 6$). The value of the parameter α is determined using (3.8) where x is found through the two-step rate selection procedure when $n = G$. Parameters for channel model-C have been based on the vehicular-B model [33] for macro-cell. The set of available rates is assumed to be $\{v_1, v_2, v_3, \ldots, v_8\}$ with $v_m = mv_1$. The assumed values for some of the other parameters are: $N = 128$, $M = 128$ and $L = 2$ (for channel model-B).

Figure 3.2 shows typical variations in m_o with n for *SR*-based error control for the different channel models previously described. Figure 3.3 shows the variation in m_o with n for *GBm*-based error control under similar conditions. While the other parameters remain the same, the curves move downward with an increase in the value of G and/or η and they move upward with an increase in the value of t and/or $\frac{E_b}{N_o}$. This is expected since an increase in the value of G and/or η increases inter-cell interference, which causes the maximum achievable transmisison rate to diminish. On the other hand, increasing the value of t and/or $\frac{E_b}{N_o}$ results in enhanced channel reliability and, consequently, the maximum affordable transmission rate increases. Note that, $\eta = 0.0$ corresponds to the case when there is no inter-cell interference. Therefore, the impact of inter-cell interference on the rate selection can be observed by varying the value of η.

The selection of the rate m_o is affected by the values of the different system parameters. It has been observed that with $G = 20$, $\eta = 0.5$, $\frac{E_b}{N_o} = 10\ dB$, a two-fold increase in the value of $n \times m_o$ occurs when t is increased from 2 to 4 and it doubles again when t is increased to 6. With the assumed values of the parameters, this is observable in the case of channel model-A for $n < 20$, while in the case of channel model-B and channel model-C, it is observable even when n is as large as 15. An increase in the value of $n \times m_o$ ultimately results in higher RLC/MAC-layer throughput.

For the proposed two-step rate selection procedure, the effects of the different system parameters (e.g., η, G, $\frac{E_b}{N_o}$, t) on the variations in RLC/MAC layer throughput with the number of active users are illustrated through Figures 3.4 and 3.5 for channel model-C. It is observed that, among all these parameters, the FEC (Forward Error Correction) parameter t impacts channel reliability (and hence throughput) the most. This is evidenced by the large increase in the value of J^* with an increase in the value

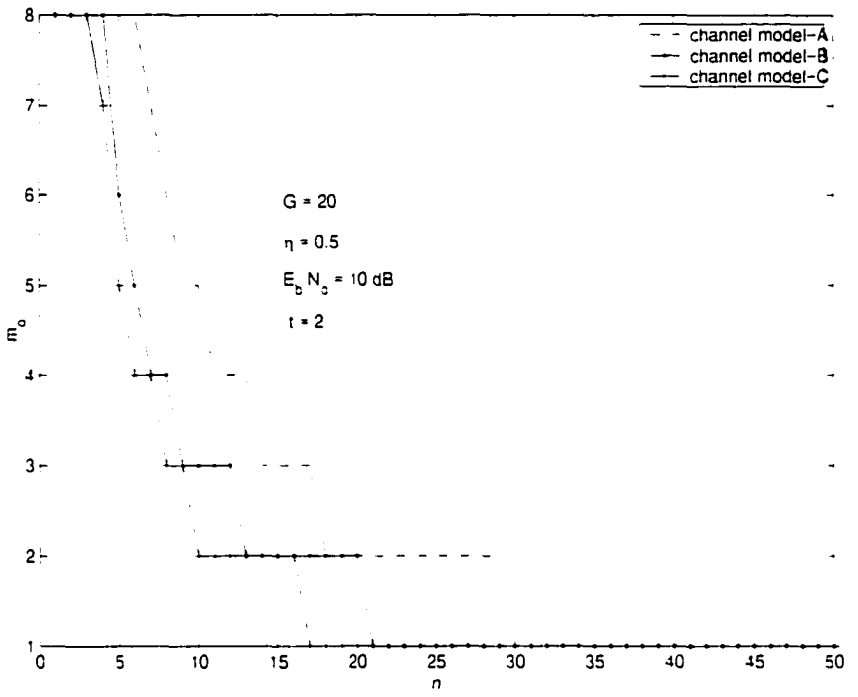

Figure 3.2. *First-step rate selection m_o vs. number of active users n (for* SR*-based error control).*

of t ($t = 2, 4, 6$). Improvement in the value of β^* due to reducing G ($G = 40, 30, 20$) and/or η ($\eta = 0.5, 0.3, 0.0$) is much less compared to the improvement due to increasing t. Again, improvement in the value of β^* for each 5 dB increase in $\frac{E_b}{N_0}$ is more pronounced than that due to a reduction in G or η. Similar results are observed for other channel models as well. The effects of inter-cell interference on throughput can be perceived by comparing the values of β^* for different non-zero values of η with those for $\eta = 0.0$. As the value of the inter-cell interference factor η increases, the throughput performance deteriorates.

For the different channel models, with *SR*-based error control, better throughput is obtained over a range of system load. For a certain average traffic load G, higher throughput performance may be observed with *GBm*-based error control when the number of active mobile stations is relatively small and it becomes more evident with increased channel reliability.

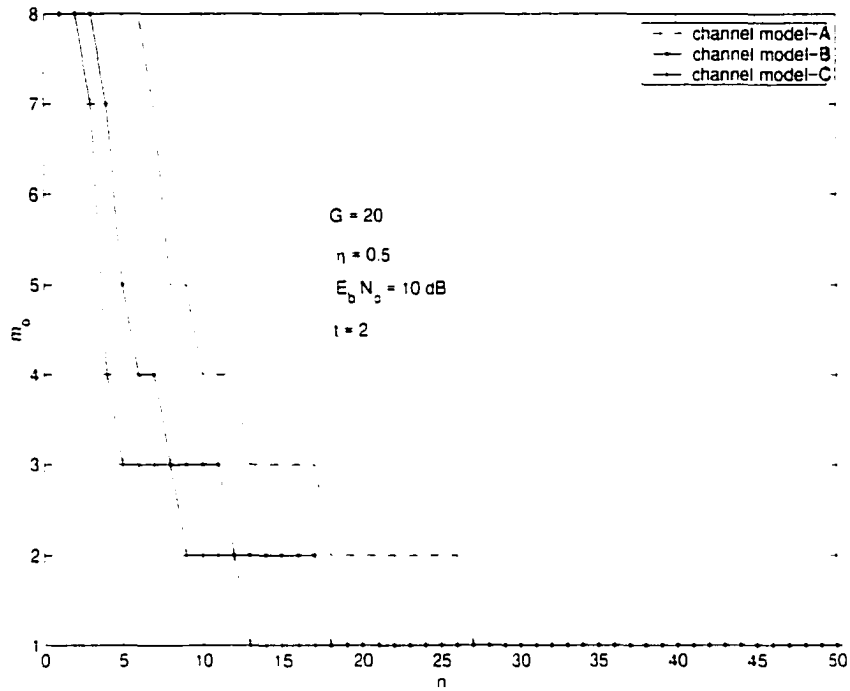

Figure 3.3. *First-step rate selection m_o vs. number of active users n (for* GBm*-based error control).*

3.4.3 Performance Comparison with the Optimal Rate Selection

To observe the performance degradation due to the proposed two-step sub-optimal rate selection, the performance of the proposed rate selection procedure is compared with that of the optimal rate selection (based on exhautive search). The 'mean-sense' approach for inter-cell interference calculation is used in the latter case as well. For a certain average load G, the *optimal throughput* for a certain number of active users (n) is obtained by cycling through all possible combinations of assigned transmission rates to n users.

Some typical results for both the *SR* and *GBm*-based error control are presented in Figures 3.6 and 3.7, respectively, from where the suitability of the proposed sub-optimal rate selection procedure becomes evident. In this particular scenario with $G = 20$, $\eta = 0.5$, $t = 2$ and $E_b/N_o = 20\ dB$, the maximum degradation in throughput for sub-optimal rate selection is found to be less than 7% in the case of *SR*-based

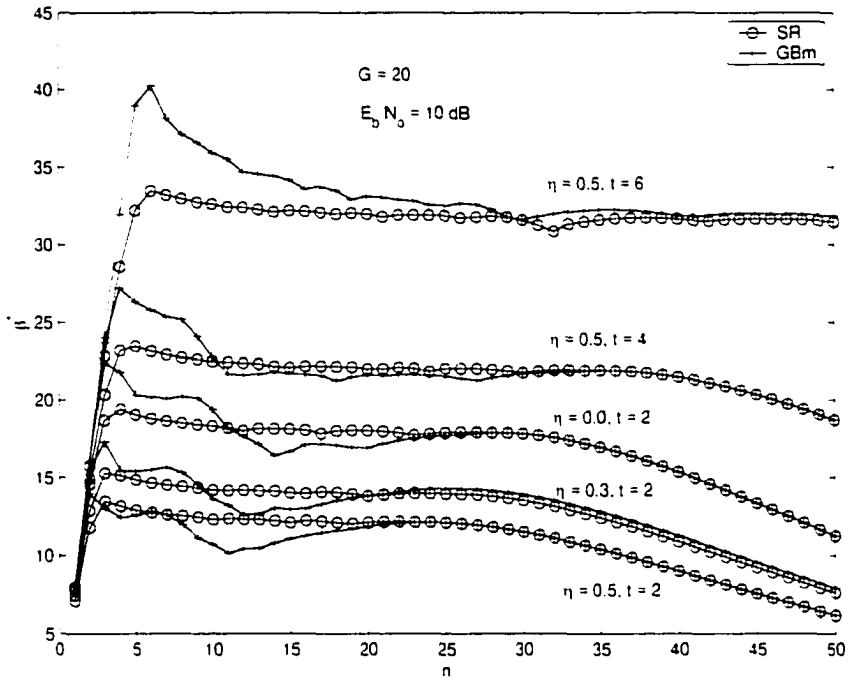

Figure 3.4. *Effects of η and t on variations in β^* with n (for channel model-C).*

error control. Similar results are observed with *GBm*-based error control and for other values of the system parameters (e.g., η, G, E_b/N_o, t, $\mathbf{v}$). The results shown in Figure 3.7 are obtained assuming $\mathbf{v} = \{v_1, v_2, \cdots, v_6\}$.

It is to be noted that, the optimal rate allocation ensures global maximum throughput (which is the throughput summed over all the cells) but it does not guarantee the best 'local-throughput' (i.e., throughput in the 'target cell'). This explains the observable occasional higher throughput in the case of sub-optimal rate selection (e.g., in Figure 3.7 for channel model-A and channel model-B) for some values of n. In this particular case, shown in Figure 3.7, the value of α is 2.35 and 2.55 in the case of the two-step sub-optimal and the exhaustive search-based optimal rate allocation, respectively. Since α is higher (which results in increased inter-cell interference) in the case of the optimal rate selection, it is not unlikely that for some values of n (number of active mobiles), the observed throughput in the 'target cell' falls slightly below that of the sub-optimal case although the former would provide the best performance in the 'sum of throughput' sense over all the cells. This would be due to the higher effective number of active users (as manifested in higher value of α in this case) in

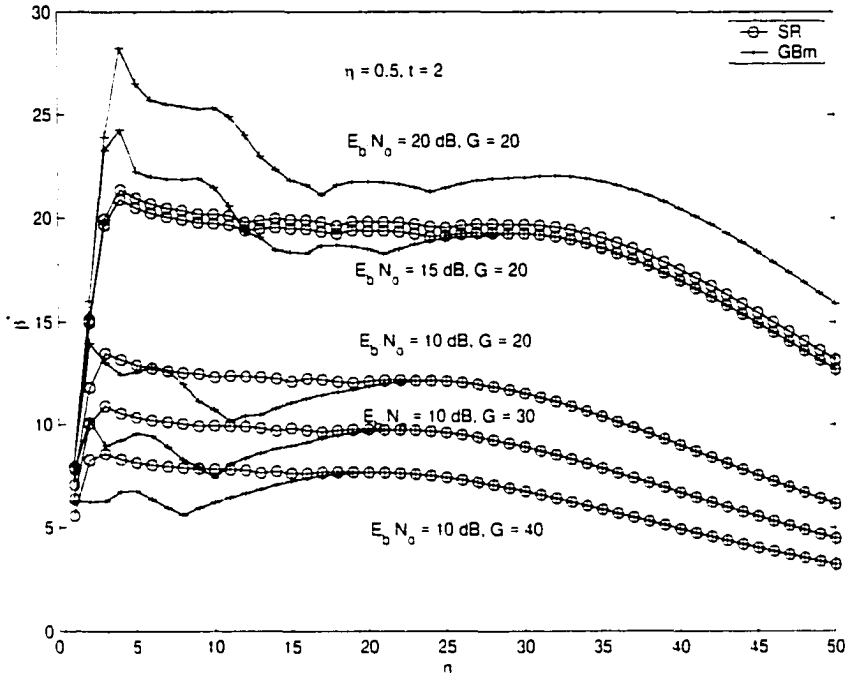

Figure 3.5. *Effects of G and* $\frac{E_b}{N_0}$ *on variations in* β^* *with n (for channel model-C).*

the case of optimal rate selection.

3.4.4 Performance Comparison with Uncontrolled Random Rate Selection

The performance of the controlled rate adaptation through the two-step search procedure is compared to that of an uncontrolled rate selection in Figures 3.6 and 3.7. In the latter case, each mobile independently chooses a transmission rate in a random fashion. The 'mean-sense' approach is used for inter-cell interference calculation as follows. For a particular value of G, a rate is assigned randomly among the mobiles and the corresponding value of α is evaluated using (3.8). This procedure is repeated for a sufficient number of times (e.g., 1000) and the average value of α is determined and then used for calculating the throughput (β) for different number of active users (n).

As is evident from Figures 3.6 and 3.7, with uncontrolled rate adaptation, the RLC/MAC-layer throughput decreases significantly, especially at relatively higher

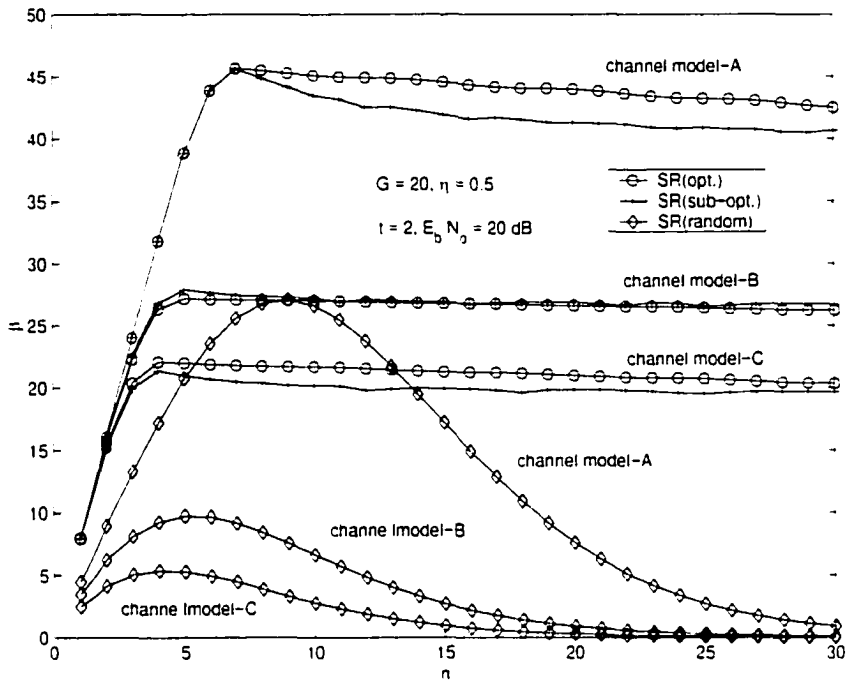

Figure 3.6. *Comparison between optimal and sub-optimal rate selection performance (with* SR*-based error control).*

system loads and/or poor channel conditions (as is observed in the case of channel model-C). On the other hand, controlled rate adaptation can provide more stable throughput performance over a wide range of system load by exploiting the user and channel information. The performance of the uncontrolled packet access in AWGN channel (i.e., channel model-A) is superior to the performance in the case of multi-path fading channel (e.g., channel model-B and channel model-C). This throughput vulnerability in multipath fading channels results from the very low SINR and hence very high BER ($P_{b,m}$).

3.5 A Markov Analytical Model for RLC/MAC-Layer Performance Evaluation

A Markov analytical model is developed in this section for the RLC/MAC-layer throughput performance evaluation in a cellular WCDMA network with dynamic

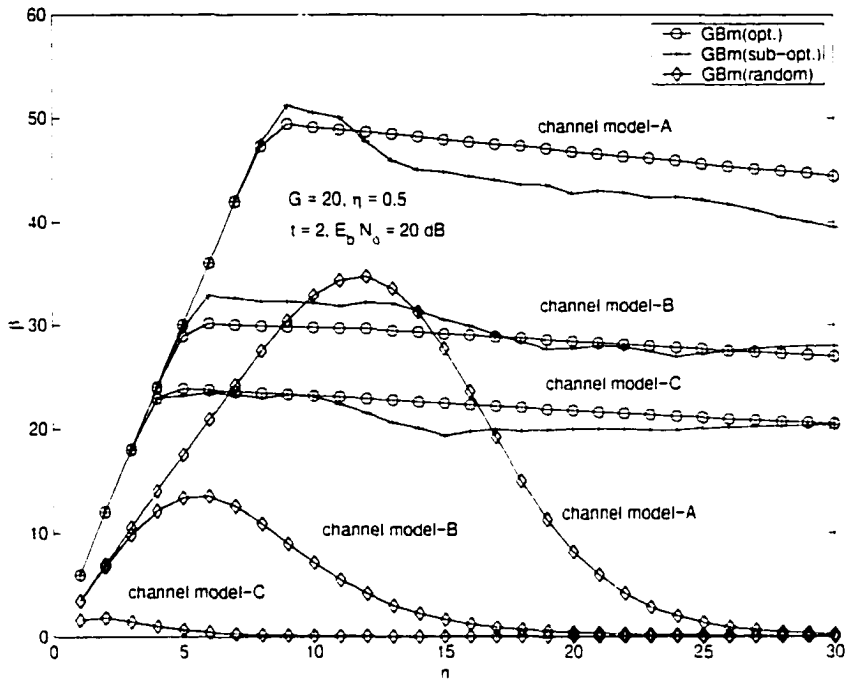

Figure 3.7. *Comparison between optimal and sub-optimal rate selection performance (with* GBm*-based error control).*

rate allocation. An ideal situation is considered where the BS has complete knowledge about the instantaneous number of transmission attempts. The issue of estimating the instantaneous traffic load at the BS is not considered here but an approximate load estimation procedure will be described later. It is to be noted that, even in the case of some approximate instantaneous traffic load estimation, the same analytical framework presented here can be used for performance evaluation.

In each slot, the mobile stations which are ready to transmit RLC/MAC-layer frames select the transmission rate based on the two-step rate selection procedure described above. The rate selection information, which is derived based on the instantaneous number of attempts, is assumed to be broadcasted to the mobiles by the BS.

Uplink transmissions from the finite number of distributed mobile stations in a WCDMA system will be controlled by some retransmission control (or access control) policy. Therefore, evaluation of the RLC/MAC-layer performance in a cellular WCDMA network with dynamic rate adaptation requires considering some retrans-

mission control (or access control) policy as well. The RLC/MAC-layer throughput performance is evaluated here assuming a constant probability retransmission control scheme.

When a transmission error occurs, mobile stations (those in the *blocked* state) attempt retransmission in the following timeslot(s) with probability ν. Mobile stations with new frames (i.e., mobiles in the *thinking* state) attempt transmission in a timeslot with probability μ. Implicit here is the assumption that the mobiles always have a sufficient number of frames to transmit under the dynamic rate adaptation scenario (for example, in cases where new frames are to be transmitted along with the old frames to match the desired transmission rate).

Under the above stated assumptions, a Markov model can be developed where the system states are specified by the number of users in *blocked* state, i.e., $\{0, 1, \ldots, K\}$ for a finite population of K users. Then, the state transition probabilities p_{ij} can be expressed by

$$p_{ij} = \Pr\left[N_{d+1} = j \mid N_d = i\right] \qquad i, j = 0, 1, \ldots, K \tag{3.16}$$

where N_d denotes the system state at the beginning of the dth timeslot. The probabilities, p_{ij}, can be derived as

$$\begin{aligned} p_{ij} = & \sum_{k=\max(0, j-i)}^{K-i} \sum_{g=\max(0, i-j)}^{i} B(K-i, k, \mu)\, B(i, g, \nu) \\ & \cdot \Pr\left[k + i - j \text{ successes} \mid k + g = n \text{ attempts}\right] \end{aligned} \tag{3.17}$$

where $B(k, l, p) \triangleq \binom{k}{l} p^l (1-p)^{k-l}$ and

$$\Pr\left[k+i-j \text{ successes} \mid k+g = n \text{ attempts}\right] = \sum_{\substack{s_1 + \ldots + s_r = k+i-j \\ 0 \le s_m \le n_m,\ \forall m}} \left[\prod_{m=1}^{r} B(n_m, s_m, P_c^m(x-m))\right] \tag{3.18}$$

for *SR*-based error control, while $P_c(x-m)$ would be substituted for $P_c^m(x-m)$ in the case of *GBm*-based error control. $P_c(x-m)$ would be replaced by $P_c(x \cdot L - m)$ and $P_c(x \cdot \rho_1 - m, x \cdot \rho_2 - m, \ldots, x \cdot \rho_L - m)$ for channel model-B and channel model-C, respectively.

For the numerical results presented here, the rate allocation for which $n_1 + \ldots + n_r = k + g = n$ and $x = \sum_{m=1}^{r} m n_m$, is assumed to be determined by the two-step rate selection procedure described previously.

The stationary probability distribution $\{\pi_i | i = 0, 1, \ldots, K\}$ can be computed by solving the equations $\pi_j = \sum_{i=0}^{K} \pi_i p_{ij}$ for $j = 0, 1, \ldots, K$ and $\sum_{i=0}^{K} \pi_i = 1$; therefore, the steady-state channel throughput is given by

$$\mathcal{S} = \sum_{i=0}^{K} \mathcal{S}(i)\,\pi_i \tag{3.19}$$

where $\mathcal{S}(i)$ is the system throughput when $N_d = i$ and can be expressed as follows (Appendix D):

$$\begin{aligned}\mathcal{S}(i) \cong & \sum_{k=0}^{K-i} \sum_{g=0}^{i} B(K-i, k, \mu)\, B(i, g, \nu) \\ & \cdot \sum_{s=0}^{k+g} \sum_{\substack{s_1 + \cdots + s_r = s \\ 0 \le s_m \le n_m,\ \forall m}} \left[\sum_{m=1}^{r} \left(m s_m + m P_c(x-m)\left[1 - P_c^{m-1}(x-m)\right](n_m - s_m) \right) \right. \\ & \left. \cdot \prod_{m=1}^{r} B(n_m, s_m, P_c^m(x-m)) \right].\end{aligned} \tag{3.20}$$

Note that $P_c^m(x-m)$ and $P_c^{m-1}(x-m)$ would be replaced by $P_c(x-m)$ and 1, respectively, in the case of *GBm*-based error control. Here, $P_c(x-m)$ depends on $P_{b,m}$ and hence on the channel model used.

Typical variations in throughput $\mathcal{S}$ with average system load G for the two-step rate selection procedure are shown in Figure 3.8. For a given number of users K, the values of μ and ν are chosen so that the average load experienced becomes G. Under the full load condition, this is achieved by choosing $\mu = \nu = \frac{G}{K}$ $(K \ge G)$ in this case. For the results presented here, K is assumed to be equal to 50. Simulation results for the *GBm* scheme are very close to the corresponding analytical results, while, in case of *SR* scheme, analytical results act as lower bounds, which is presumably due to the approximation used in (3.20). As the channel condition deteriorates, the throughput drops off rapidly with an increase in the value of average offered load G. When the average load is relatively small and/or channel condition improves, *GBm*-based error control may offer higher throughput than *SR*-based error control.

In an underload situation, the Markov model would be extended with the number of *blocked* states being equal to the number of available rates (as opposed to 1 in this case).

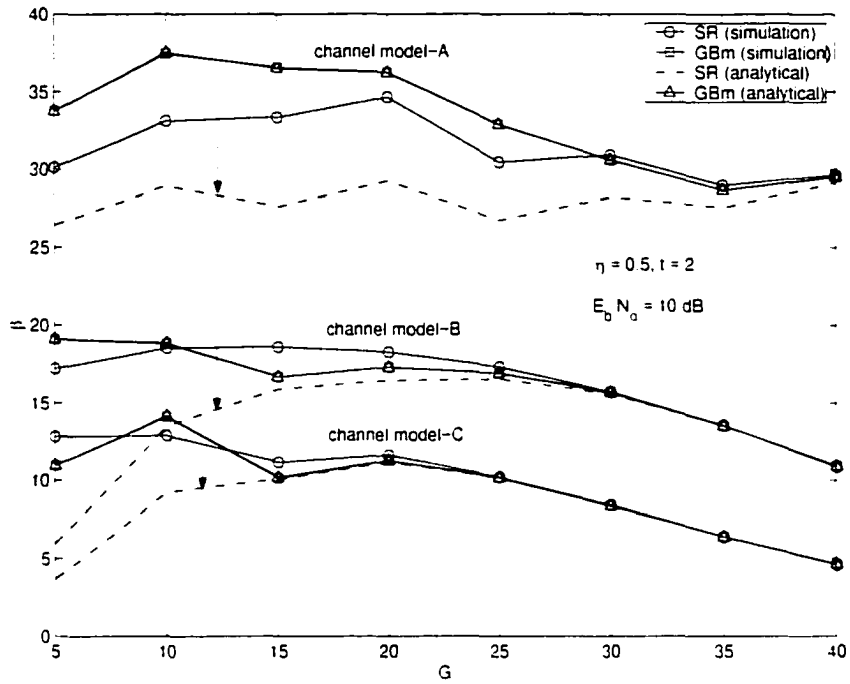

Figure 3.8. *Variation in S with G with two-step rate selection procedure.*

3.6 A BS-Assisted Distributed RLC/MAC Protocol with Integrated Rate and Error Control

In a cellular WCDMA environment, it may not be easy to control the variable rate transmission (in a timeslot) from the mobiles in a centralized fashion based on the instantaneous traffic load (i.e., the number of transmission attempts). This difficulty is due to the fact that information regarding the ready to transmit mobiles may not be known *a priori* unless a sub-rate channel is maintained for each active user to inform the BS of the next transmisison attempt which would cause extra overhead on channel resources. Again, in a fully distributed environment the mobiles may not be able to choose the suitable transmission rates without having any knowledge about the system load and channel condition. Therefore, a hybrid medium access control scheme employing adaptive rate selection procedure may be more desirable to achieve high radio spectrum efficiency.

In this section, a simple BS-assisted medium access control protocol based on

the dynamic rate adaptation is presented which can be implemented distributedly in the mobile stations. Based on an estimation of the uplink traffic load, the BS broadcasts the sub-optimal rate selection value in each timeslot, which is used by all the mobiles ready to transmit in the next timeslot. Note that all the mobiles use the same transmission rate in this case, which is derived through the first step of the sub-optimal rate selection procedure described above. The BS broadcasts the transmission status along with the rate selection value (i.e., m_o) in the next timeslot; that is, the feedback delay here is one timeslot.

The RLC/MAC procedure can be described as follows:

i. The BS estimates system load in a cell in terms of the number of mobiles attempting transmission in a slot. The load is estimated using a low-pass filter which computes average load $\bar{n}(d)$ based on the number of mobiles ($n_m(d)$) for which all the frames (at least one frame) have(has) been received successfully in case of *GBm(SR)*-based error control. That is,
$\bar{n}(d) = c \times n_m(d) + (1 - c) \times \bar{n}(d-1)$, where c is the low-pass filter gain.

ii. Based on the load estimate $\bar{n}(d)$, the BS computes the rate m_o using the first step of the two-step rate search procedure as previously described and broadcasts it to the mobiles in the $(d+1)$th slot. If $\bar{n}(d) = 0$, the highest rate is chosen.

iii. The mobiles which are ready to transmit, use rate m_o for frame transmissions in the $(d+2)$th slot.

iv. The retransmission control (or access control) probability is controlled locally at the mobiles by a truncated exponential backoff scheme [14]. In the case of successful transmission (of all the frames/at least one frame in the case of *GBm/SR*-based error control) the retransmission probability is increased multiplicatively while in the case of transmission failure (of all the frames), it is decreased multiplicatively. That is,

$$p(d+1) = \begin{cases} max\,(p_{min},\ p(d) \times q), & \text{if the transmissions were unsuccessful} \\ min\,(p_{max},\ p(d)/q), & \text{otherwise} \end{cases}$$

where $p(d)$ and $p(d+1)$ are the retransmission probabilities for the dth attempt and $(d+1)$h attempt, respectively. Here, $0 < p_{max} \leq 1$, $0 < q < 1$, $p_{min} = q^s$, $s \in I$. For the rest of the chapter $p(1) = q$ and $p_{max} = 1$.

It is to be noted that, although the channel load in the 'target cell' varies with time depending on the retransmission control policy, the average system load over all the cells is assumed to be constant so that the 'mean-sense' approach for inter-cell interference calculation can still be applied.

The performance of this medium access control scheme is evaluated using a *C*-coded time-driven simulator. A heavy traffic load condition is assumed at the mobiles so that each mobile can transmit as many packets as required in a timeslot to match the required transmission rate. The values of the *backoff rate* parameter (q) and *backoff limit* parameter (s) are assumed to be 0.5 and 8, respectively. The value of the low-pass filter gain (c) is assumed to be 0.5.

Simulation results, illustrating the performance of the above RLC/MAC scheme in terms of the RLC/MAC throughput (β) under a varying number of users (K), are presented in Figures 3.9 and 3.10. Note that K is the upper bound for the number of active users transmitting in a timeslot (n). To compare the performance with that of the sub-optimal case with a centralized two-step rate selection based on the instantaneous traffic load, the values of β^* corresponding to the varying n are also plotted.

Although the same rate (derived through the first step of the rate selection procedure) is used by all the mobiles during a timeslot, the observed throughput performance for the chosen parameters is fairly close to the sub-optimal performance in the case of a two-step rate selection procedure with exact instantaneous traffic load estimation. The increased throughput stability at higher values of n (for model-C in Figure 3.9 and both for model-A and model-C in Figure 3.10) is due to the retransmission backoff policy employed herein. Throughput degradation at higher values of n in the case of channel model-A (Figure 3.9) may be avoided by suitably adjusting of the retransmission control parameters (i.e., q and s). It is to be noted that the values for the parameters q and s assumed here have not been optimized.

Since the actual number of users transmitting in a slot is upper-bounded by K, for a certain value of K, the observed throughput with the distributed access control scheme would be upper-bounded by the maximum throughput obtained in the case of two-step rate selection over the range of n, extending from 0 up to K (i.e., $0 < n \leq K$). With enhanced channel reliability (e.g., for channel model-A) and/or when

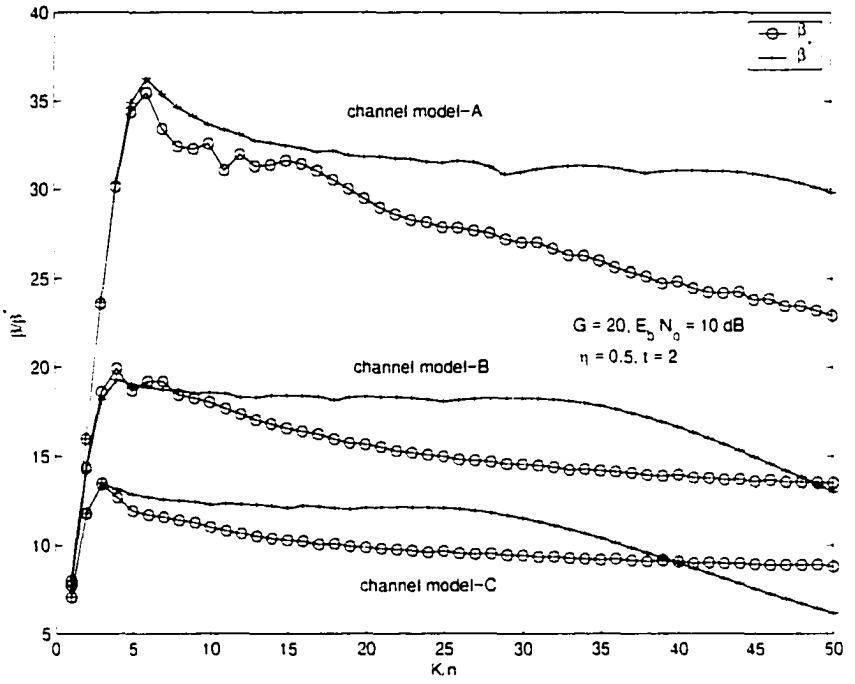

Figure 3.9. *Throughput performance of the proposed MAC scheme (with* SR*-based error control).*

the number of users (K) is relatively small. *GBm*-based error control outperforms *SR*-based error control. Simulation results make it evident that the proposed BS-assisted distributed packet access control mechanism with controlled rate adaptation can provide high throughput performance.

3.7 Chapter Summary

The problem of optimal dynamic rate allocation among mobile stations for variable rate packet data transmission in a cellular wireless network is NP-complete. Therefore, fast and simple sub-optimal schemes for rate adaptation are to be devised. Again, interference calculation under variable rate transmission is non-trivial in a packet-switched cellular CDMA network with heterogeneous traffic load in different cells.

A sub-optimal two-step dynamic rate selection procedure has been proposed in this chapter for uplink packet data transmission in cellular WCDMA networks, which has

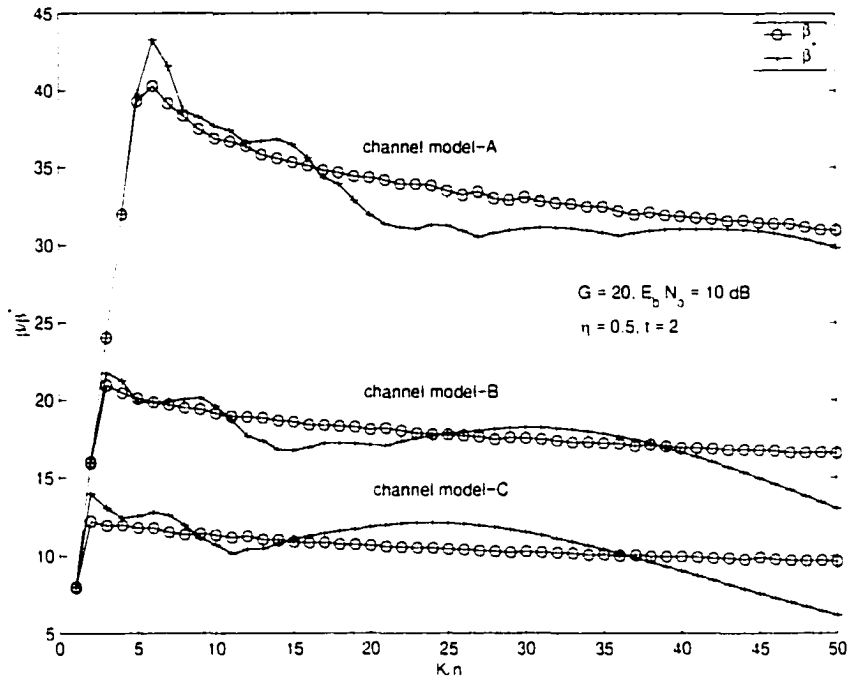

Figure 3.10. *Throughput performance of the proposed MAC scheme (with* GBm*-based error control).*

been based on a novel 'mean-sense' approach for inter-cell interference calculation under variable rate transmission. This 'mean-sense' approach is restricted by the requirement that, although the system load in different cells may vary, the mean load averaged over all cells is to be kept fixed. This requirement is not too restrictive for a cellular packet-switched WCDMA network where the RNC may be responsible for load/power management across different cells to ensure the stability of the network. Performance evaluation under a heterogeneous traffic load scenario, where the mean traffic load averaged over different cells is time-varying, needs further investigation.

The performance of the proposed fast and simple two-step rate selection procedure is fairly close to that of the optimal rate selection found through exhaustive search, which is not feasible from an implementation viewpoint. Two different error control alternatives in a variable rate packet data transmission scenario, namely, *SR* and *GBm*, have been presented and their performances have been analyzed for three different channel models. With an enhanced channel reliability and/or when the number of users is relatively small, the *GBm* scheme has been observed to outperform

the *SR* scheme. *SR*-based error control offers superior performance over a wide range of system load.

A general Markov model for evaluating the system throughput under dynamic rate allocation (in the ideal case of instantaneous exact load estimation) with two different error control alternatives has been presented, assuming a constant probability retransmision control. Lastly, a BS-assisted distributed medium access control scheme employing dynamic rate adaptation has been proposed. The throughput performance of this scheme, for chosen protocol parameters, has been observed to be fairly close to the performance of the ideal scheme with two-step sub-optimal rate adaptation.

Chapter 4

Higher Layer Protocol Performance in CDMA S-ALOHA Networks with Packet Combining in Rayleigh Fading Channels

4.1 Introduction

CDMA S-ALOHA is an efficient way of multiplexing packet data in the radio interface of wireless networks. As a multiple access protocol, CDMA possesses the advantages of zero channel access delay, potential for high throughput performance and low peak power in the transmitters. The random access scheme in the reverse link common-channel in WCDMA, which has been proposed as one of the radio transmission technologies for IMT-2000, is based on multi-code spread S-ALOHA ([35]-[36]). The performance of a CDMA S-ALOHA system is primarily limited by MAI (Multiple Access Interference) and channel fading; therefore, different techniques are needed at different protocol layers to combat the detrimental effects resulting from these phenomena.

Satisfactory throughput performance can be provided by limiting the number of simultaneous transmissions (and hence MAI) under dynamic system load through the adjustment of the RLC/MAC protocol parameters (e.g., *retransmission control*[1]

[1]The terms 'retransmission control' and 'access control' are used interchangeably here.

parameter). A CDMA S-ALOHA system that uses a constant retransmission probability is inherently unstable with a large population of terminals [37]. Scheme such as *MCLSP (Modified Channel Load Sensing Protocol)* was proposed to combat the effects of throughput degradation at large offered loads and was analyzed for infinite population systems [38]. As for the other random access protocols, retransmission probability (p_r) is the most appropriate RLC/MAC level parameter that can be varied here to achieve a good delay-throughput characteristic under dynamic system load.

Again, some diversity reception and packet combining schemes can be used in the physical layer to cope with the detrimental effects of MAI and channel fading ([39]-[43]). In particular, a retransmission diversity packet combining operation in CDMA networks can be well-suited for delay insensitive reliable data transmission services.

Either one or two levels of error recovery can be performed on top of the physical and MAC layers in a random access CDMA system. Two levels of error recovery are required for reliable data transmission when the underlying link layer is semi-reliable with some maximum allowed number of retransmissions in the case of unsuccessful transmission of a link-layer frame. Having two levels of error recovery implies the presence of a transport-level protocol on top of the physical and RLC/MAC layer (similar to the IS-99 protocol standardized for CDMA circuit-mode data). Two levels of error recovery may allow more flexibility in selecting 'packet' size at the higher level [44]. Performance comparisons of two retransmission protocols for CDMA circuit-mode data were carried out in [44].

Transport protocol performance is of paramount importance as far as user's application level QoS is concerned. The performances of two timer-based TCP implementations were investigated in [45] when the wireless channel is periodically taken away from the end to end connections. Retransmission control at the RLC/MAC layer can significantly affect the transport protocol performance. Slow frame loss recovery may trigger frequent transport protocol timer backoff resulting in very poor delay-throughput characteristics. Again, the RLC/MAC layer performance will be largely affected by the physical layer design.

In this chapter, the RLC/MAC-layer and transport-layer throughput performances of a CDMA S-ALOHA system have been analyzed with retransmission diversity

packet combining under frequency selective Rayleigh fading for three different heuristic-based RLC/MAC-layer retransmission control schemes, namely, *TEB (Truncated Exponential Backoff)*, *AIMD (Additive Increase and Multiplicative Decrease)* and *MB (μ-law Backoff)*. Two transport protocol timer control mechanisms, namely, *EB (Exponential Backoff)* and *EWMA (Exponential Weighted Moving Averaging)*, are considered. The identification of suitable RLC/MAC-layer access control technique and transport-layer timer control mechanism for a CDMA S-ALOHA network with packet combining and two levels of error recovery are of particular interest.

The organization of the rest of the chapter is as follows. Section 4.2 describes the background and motivation of the work. The system model is presented in Section 4.3 along with the different RLC/MAC-layer retransmission control schemes and transport protocol models for CDMA S-ALOHA networks that are being considered here. Subsequently, the performance results obtained through computer simulation are presented in Section 4.4. Finally, the key observations are summarized in Section 4.5.

4.2 Background and Motivation

Packet combining techniques are known to the wireless communications engineers and researchers as a means of improving channel utilization in packet radio networks ([39]-[42]). Recently, a bit level packet combining scheme employing selection diversity at the output of the correlator receiver has been proposed in [43] to improve system capacity and reliability in DS/CDMA networks. Lower bounds on system throughput were derived under Poisson distributed independent traffic arrival for different diversity order although no specific retransmission control strategy was assumed therein.

Since the time-varying nature of the number of interfering packets (with respect to a target packet) was deemed difficult to track analytically, the bounds derived in [43] are contingent upon the value of the maximum number of interfering packets (with respect to a target packet) during the course of retransmitting the target packet. Since the distribution of the number of interfering packets in such a system is determined by the retransmission control (or access control) policy ([13]-[14], [46]-[47]) employed at the RLC/MAC layer, evaluating the performance of such a system in

a practical network scenario requires considering the access control policy employed therein. Better throughput stability can be achieved by using an appropriate access control policy.

The work presented here complements the work in [43] in that the RLC/MAC performance of CDMA S-ALOHA networks with diversity packet combining is analyzed under three different retransmission control schemes through computer simulation. It focuses on the effects of macrodiversity packet combining on higher layer protocol performance to get insight into the identification of suitable protocol mechanisms at different layers which would be required for transmission protocol stack performance optimization. We demonstrate how the different transport protocol mechanisms (e.g., retransmission timer adjustment method) react differently to the different RLC/MAC layer access control techniques. The achieved throughput is sensitive to the reaction of different timer control mechanisms to the highly variable channel access and packet transmission time. Simulation results indicate how the different physical layer parameters (e.g., code rate, number of resolvable multipaths) affect the RLC/MAC-layer and the transport-layer throughput performances under different RLC/MAC-layer retransmission control schemes, which, in turn, reveals some alternative physical layer design choices.

4.3 System Model and Assumptions

4.3.1 Traffic Source Model

The data users are modeled as 'packet pipes', i.e., over the time considered, a user always has data to transmit in each timeslot. Although users having a very long stream of data traffic are less likely to use random access channels (except in cases where there are cost constraints), this traffic model allows us to evaluate the system performance under full load condition. It is to be noted that, under such conditions, the throughput of a transport protocol connection is maximized.

For a fixed number of mobile stations, traffic arrivals as intermittent packet bursts would result in a reduced system load (compared to that in case of full load situation). Therefore, results obtained with some bursty traffic arrival patterns (e.g., for Pareto distributed inter-arrival time and burst-length) do not give much more insight than

those obtained under full load condition by varying the number of active mobile stations.

Each of the mobile stations having a unique spreading code synchronizes its RLC/MAC-layer frame transmission so that it initiates transmission at the beginning of a timeslot. Each RLC/MAC-layer frame is channel coded so that it is correctly decodable even when a maximum of t bits are in error (bit errors are assumed to occur independently) although no specific coding and interleaving scheme is assumed herein. If the BS receiver can decode the frame successfully, the mobile station receives an ACK before the beginning of next timeslot. When the transmission is unsuccessful, the same frame is scheduled for retransmission using a retransmission control policy.

4.3.2 Channel and Receiver Model

A frequency-selective multipath fading channel is considered with a maximum of L resolvable multipaths. The maximum multipath delay spread is assumed to be less than the symbol duration to avoid intersymbol interference. It is assumed that the MAI follows a Gaussian process and that the channel has a constant multipath intensity profile (i.e., variance of energy is equal for all multipaths).

The BS uses a correlation (matched filter) receiver. Figure 4.1 depicts the system model, assuming K simultaneous transmissions and a non-diversity receiver. In the case of macrodiversity (i.e., retransmission packet diversity), instead of discarding the frames that are found to be erroneous, the receiver retains and processes all the copies of a frame using post-demodulation diversity combining. Selection diversity is employed at the bit level at the output of the correlator to form a more reliable packet by using all the stored copies. That is, the detection of a bit in a particular packet is based on the maximum measured SNIR (Signal-to-Noise plus Interference Ratio) corresponding to the bit among all the diversity copies. The decision criterion in the case of Mth order diversity is therefore $max[a_i]$, $i = 1, \cdots, M$, where a_i is the channel gain corresponding to the ith diversity branch. Microdiversity reception, which is not being considered here, will further enhance the system performance. The RLC/MAC-layer frame transmissions from the mobile stations can be depicted as shown in Figure 4.2.

The system model under the above stated assumptions allows for insight into the

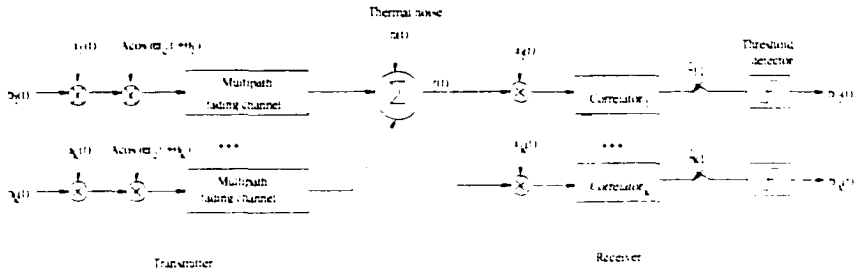

Figure 4.1. *Transmitter and receiver model.*

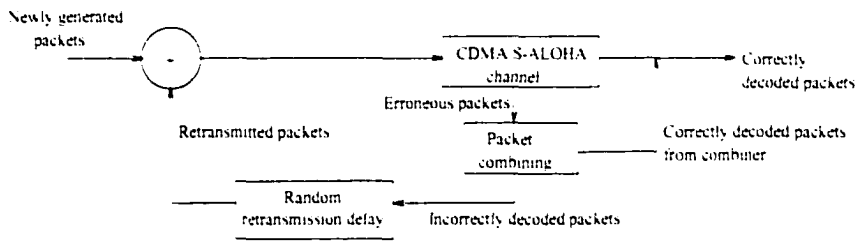

Figure 4.2. *Data flow between RLC/MAC layer and physical layer.*

system performance without being too unrealistic. Since the intent is to explore the inter-layer protocol interactions, more complex diversity combining mechanisms (e.g., maximal ratio diversity combining) are not considered here. The incorporation of a more complex diversity combining model may not provide much more insight as far as inter-layer protocol interactions are concerned.

If the combining operation is performed in an incremental fashion, i.e., the newly received retransmitted copy of a frame is combined with the stored combined frame from previously received copies (with errors), the additional buffer space required would be only one frame.

Under the above stated assumptions, BER without any diversity combining is [43]

$$P_b(K) = \frac{1}{2}\cdot\left(1 - \sqrt{\frac{\gamma_o}{1+\gamma_o}}\right) \tag{4.1}$$

with $\gamma_o = 0.5 \times \left(\frac{KL-1}{3N_c} + \frac{N_o}{2E_b[2\sigma_1^2]}\right)^{-1}$, where N_c = number of chips per data bit, E_b/N_o = ratio of bit energy to the AWGN noise spectral density, L = number of resolvable multipaths and σ_1^2 = variance of energy corresponding to the first multipath component.

BER with diversity combining is

$$P_b(K) = \frac{M}{2} \cdot \sum_{k=0}^{M-1} \binom{M-1}{k} \cdot \frac{(-1)^k}{k+1} \cdot \left(1 - \sqrt{\frac{\gamma_0}{1+k+\gamma_0}}\right) \tag{4.2}$$

where $\gamma_0 = 0.5 \times \left(\frac{KL-1}{3N_c} + \frac{N_o}{2E_b \cdot 2\sigma_1^2}\right)^{-1}$ and M is the diversity order.

After the BER is determined, for a frame size of N_p bits with t-bit forward error correction coding, the FER (Frame Error Rate) can be calculated as follows:

$$FER(K) = 1 - \sum_{r=0}^{t} \binom{N_p}{r} \cdot P_b^r(K) \cdot (1 - P_b(K))^{N_p - r}. \tag{4.3}$$

It is assumed that the ACK information from the BS are transmitted error-free on the downlink channel. This is a reasonable assumption since the small number of bits conveying the ACK information can be adequately protected through channel coding. Since a mobile station transmits a new frame only after the previous one has been successfully transmitted and since the despreading operation in the receiver is based on the unique spreading code for each mobile, the chance of wrong packet combining is minimized (or possibly eliminated).

4.3.3 Retransmission Control Schemes

The three following binary channel-status (*success/non-success*)-based retransmission control policies are considered:

- Truncated Exponential Backoff (TEB) Scheme
- Additive Increase and Multiplicative Decrease (AIMD) Scheme and
- μ-law Backoff (MB) Scheme.

TEB Scheme: The retransmission control probability in this case is updated according to the following:

$$p(r+1) = \begin{cases} max\,(p_{min},\ p(r) \times q_{LL}). & \text{if the transmission was unsuccessful} \\ min\,(p_{max},\ p(r)/q_{LL}). & \text{if the transmission was successful} \end{cases}$$

where $p(r)$ and $p(r+1)$ are the retransmission probabilities during the rth attempt and $(r+1)$th attempt, respectively. Here, $0 < p_{max} \leq 1$, $0 < q_{LL} < 1$, $p_{min} = q_{LL}^{s}$, $s_{LL} \in I$ $(s_{LL} > 0)$. In the rest of the chapter, we assume $p(1) = q_{LL}$ and $p_{max} = 1$.

AIMD Scheme: In this case, the binary feedback (*success/collision*) information is used to update the retransmission control probability according to the following:

$$p(r+1) = \begin{cases} max\,(p_{min},\ p(r) \times a)\,. & \text{if the transmission was unsuccessful} \\ min\,(p_{max},\ p(r) + (1 - p(r)) \times b)\,. & \text{if the transmission was successful} \end{cases}$$

where $0 < a < 1$ *and* $0 < b \leq 1$.

If the recent transmission by a mobile station is unsuccessful due to collision, then the mobile decreases the value of the retransmission probability multiplicatively. In the case of successful transmission, it increases the retransmission probability in an additive fashion. In this scheme, a decrease in the value of $p(r)$ in the case of collision is expected to be more abrupt compared to an increase in the value of $p(r)$ in the case of a successful transmission. This concept is similar to that of the TCP window adjustment in the case of network congestion.

MB Scheme: In this case, the binary feedback (*success/failure*) information is used to update the retransmission control probability according to μ-law and inverse μ-law [48] as follows:

$$p(r+1) = \begin{cases} min\left(\frac{ln(1-\mu p(r))}{ln(1-\mu)},\ p_{max}\right). & \text{if the transmission was successful} \\ max\left(\frac{(1-\mu)^{p(r)}-1}{\mu},\ p_{min}\right). & \text{if the transmission was unsuccessful.} \end{cases}$$

Here, μ is a non-zero positive integer and is called curvature control parameter. The rate of increase of $p(r)$ is high at low values of $p(r)$ while it decreases as $p(r)$ approaches 1. The rate of increase is determined by the parameter μ. Again, the rate of decrease is high for large values of $p(r)$ and it diminishes as $p(r)$ decreases.

All these schemes can be used either globally or locally. For global adaptation, the BS can update the value of the access control parameter, based on the transmission status during each timeslot, and can broadcast the value to the mobile stations. For local adaptation, each mobile can adjust the value of the access control parameter independently of the other mobiles based only on the feedback status corresponding to its own transmission, thus resulting in a more 'power-conservative' RLC/MAC protocol.

4.3.4 TP (Transport Protocol) Models for Two Levels of Error Recovery

A transport-layer protocol capable of end-to-end error recovery can be used for ensuring reliable transmission for non-real-time data traffic (e.g., FTP traffic). A TPDU (Transport Protocol Data Unit) consists of several RLC/MAC-layer frames and, in the case of RLC/MAC-layer failure, the transport protocol reacts by retransmitting the entire TPDU. In IS-99, the CDMA circuit mode data transmission standards, the link-layer frame size is 24 *bytes* (with 19 *bytes* of payload), which corresponds to a slot duration of 20 *ms* for a channel data rate of 9.6 *Kbps* and the TCP segment size is 536 *bytes*.

For each TPDU, the transport protocol at the transmitting mobile station maintains a *retransmission timer (RT)* expiration of which triggers retransmission of the corresponding TPDU. At the start of transmission of a TPDU, the value of retransmission timer is set to some *retransmission timeout (RTO)* value. For TCP, the three techniques that deal with the calculation of the RTO are [49]: *RTT (Round Trip Time) Variance Estimation (Jacobson's Algorithm), Exponential Backoff*, and *Karn's Algorithm*.

Here, two transport protocol timer mechanisms are considered: one is purely based on exponential backoff and the other is a hybrid scheme based on exponential backoff and on the weighted moving averaging of the transmission delays corresponding to the successfully transmitted TPDUs. It is to be noted that several other variations of these timer management schemes are possible depending on some different increase/decrease mechanisms for the RTO value. For brevity, discussions are limited to only these two fairly simple schemes that involve only a few parameters and enable us to clearly demonstrate the impact of the transport-layer timer control mechanism on system performance.

For S-ALOHA-based channel access, the retransmission strategy at the transport layer is 'stop and wait': in other words, the transport protocol transmission window size is always equal to one TPDU. Therefore, the transport protocol entity here does not employ any dynamic window management technique (e.g., conventional 'slow-start' and 'congestion avoidance' mechanisms used in TCP). The two variations of the transport protocol are described below.

4.3.4.1 TP with EB (Exponential Backoff)-Based *RTO* Control

In the *Exponential Backoff* scheme, after successful transmission of a TPDU, the RTO value is reduced multiplicatively (i.e., each source will now wait a shorter time before retransmitting the following TPDU, if required). In the case of transmission failure, the RTO value is increased multiplicatively so that each source will wait a longer time before timing out for a second retransmission, if it is required at all. Successive packet losses lead to RT backoffs and TPDU retransmissions and hence the transport protocol throughput decreases. The RLC/MAC layer should perform faster frame-loss recovery by using retransmission to reduce RT timeouts. Slow frame loss recovery may trigger frequent transport protocol timer backoff which causes very poor system performance.

The transport protocol in this case can be described as follows:

Procedure Transport_One() {

1. $RTO = RTO_{min}$. /* initialization done at the start of a session */
2. Send TPDU to RLC/MAC layer.
3. Start RT with timeout value = RTO.
4. Wait until (ACK for the TPDU is received from RLC/MAC layer)
   or (RT times out).

   If (ACK is received)

   $RTO = max(RTO \times q_{TL}, RTO_{min})$, $0 < q_{TL} < 1.0$

   Else if (RT times out)

   $RTO = min(RTO/q_{TL}, RTO_{max})$.
5. Go to step 2.

}

In conventional TCP implementations, after a timeout occurs, the timer value corresponding to a TCP connection is updated in a similar fashion as in *Transport_One*. But the RTO value is also updated regularly, depending on the RTT estimates that are calculated based on the measured RTT values corresponding to successfully received unretransmitted TCP packets.

4.3.4.2 TP with EWMA (Exponential Weighted Moving Averaging)-Based RTO Control

In this case, the estimated RTT delay for a TPDU is calculated using a low-pass filter, as in the *Jacobson's Algorithm*, except that the RTT calculation is not augmented with an estimate of the mean deviation in RTT. The RTO value is equal to the estimated RTT value.

The transport protocol can be described as follows:

Procedure Transport_Two() {

1. $RTO = RTO_{min}$ /* initialization done at the start of a session */
2. Send TPDU to RLC/MAC layer.
3. Start RT with timeout value = RTO
4. Wait until (ACK for the TPDU is received from RLC/MAC layer)
   or (RT times out)

   If ACK is received (say, after time t_{tp})
   $RTO = (1 - c) \times RTO + c \times t_{tp}$, $0 < c < 1.0$
   Else if (RT times out)
   $RTO = min(RTO/q_{TL}, RTO_{max})$, $0 < q_{TL} < 1.0$
5. Go to step 2.

}

Conventional TCP implementations update the RTO value as the sum of the smoothed average and four times the standard deviation of the RTT values (*Jacobson's Algorithm*).

During simulation, the values of RTT and RTO can be measured/controlled in a more 'fine-grained' manner than those for some of the current implementations. For example, in TCP Tahoe, the RTO values can be changed only with a minimum resolution of 500 *ms*.

4.3.5 Unreliable TP Throughput Performance

In the case of an unreliable transport protocol (e.g., UDP (User Datagram Protocol)) that may be appropriate for traffic which is moderately loss sensitive and delay sensitive (as opposed to delay insensitive but loss sensitive traffic), a TPDU is not re-

transmitted when one or more of the link-layer frames is (are) lost due to MAI and/or channel fading. Since it is a single-layer ACK-based error recovery mechanism, there is no need for any timer management.

When a frame transmission is unsuccessful, it is assumed that the RLC/MAC layer keeps retransmitting the corresponding frame until it is successfully transmitted or the delay reaches some predefined maximum limit D_l. In the latter case, the frame is discarded and the mobile station schedules the next candidate frame for transmission. In this case, the BS receiver should be able to flush the corresponding combining buffer after the delay limit.

4.3.6 Performance Metrics

- *RLC/MAC-Layer (or Link-Layer) Throughput (T_{LL})*: It is measured as the number of RLC/MAC layer frames successfully transmitted per frame-time (or timeslot). For single-level error recovery, application level throughput can be measured from T_{LL} as well.
- *Transport Protocol Throughput (T_{TL})*: It is measured as the minimum amount of transport-layer data transferred per unit time over the wireless channel. For the different values of the FEC (Forward Error Correction) parameter t, the Varsharmov-Gilbert bound [50] in (4.4) is used to determine the lower bound on the number of information bits (k) that can be allocated in an RLC/MAC-layer frame of size N_p bits so that up to t single bit errors can be corrected. Hence, based on the TPDU size, the lower bound on the number of information bits in a TPDU can be determined.

$$N_p - \log_2 \sum_{i=1}^{2t} \binom{N_p}{i} > k \geq N_p - \log_2 \sum_{i=1}^{2t-1} \binom{N_p}{i}. \qquad (4.4)$$

- *Frame Dropping Probability (P_{drop})*: It is measured as the fraction of the total RLC/MAC-layer frames dropped for which the waiting time exceeds the pre-specified delay limit D_l. It is one of the performance measures for the different access control schemes in the case of single-level error recovery for moderately loss sensitive and delay sensitive data traffic.

4.3.7 Simulation Architecture and Parameters

The system architecture considered here is similar to the one described in [51] where the BS is assumed to be connected to the PSTN through an interface called IWF (Inter-Working Function) and a reliable transport protocol provides a reliable connection between the mobile station and the IWF. Uplink transmissions, from mobile station to BS, are only being considered here.

A *C*-coded event-driven simulator is used for performance evaluation. In the simulator, the event set is comprised of the following two events: *Time_Slot* and *Update_Rexmit_Prob*. The *Time_Slot* event occurs periodically with the period being equal to the length of a slot in time. The event *Retransmission_Time_Out* occurs only when the timer corresponding to a TPDU in a mobile times out, in which case, the RTO value is updated accordingly. The set of tasks performed while executing the *Time_Slot* event can be described as follows:

i. *Simulate_Generate*: for each *free*[2] mobile station, a new TPDU is generated with probability p_g ($p_g = 1$ under full load assumption).

ii. *Start_Transmission*: each mobile station attempts to transmit an RLC/MAC-layer frame with probability p_r.

iii. *Update_Station_Status*: by using the number of simultaneous transmissions K in the current timeslot, the FER is calculated according to (4.3). For blocked stations, the total number of transmissions performed so far is used as the diversity order (M) for calculating the *FER*. The transmission status (*success/failure*) for a mobile station transmitting in the current slot is determined by a *Bernoulli trial* with probability *FER*.

iv. *Update_Rexmit_Prob*: with global adaptation, p_r is updated for *all* the mobile stations according to a retransmission control scheme while for local adaptation, p_r is updated only for those which have transmitted in the current timeslot.

v. *Update_TP_Statistics*: for two-level error recovery, mobile stations for which the number of successfully transmitted frames equals the TPDU size, the transport protocol statistics (e.g., number of successfully transmitted TPDUs) and the value of RTO are updated accordingly. For single-level error recovery, statistics

[2]It refers to a mobile station which is not blocked due to some previous transmission failure.

on the number of successfully transmitted frames are updated accordingly.

vi. *Check_Frame_Loss*: in the case of single-level error recovery, it is invoked to check whether the transmission delay time of a frame in a blocked station has reached the specified delay limit, in which case, the corresponding frame is dropped and the loss statistics are updated accordingly.

For each point in the simulation result set, the simulator is run until 5 *million Time_Slot* events have occurred and simulation statistics are flushed after 300000 *Time_Slot* events to avoid the transient effects.

The different simulation parameters are: the number of active mobile stations (N), the processing gain (N_c), the ratio of bit energy to AWGN noise spectral density (E_b/N_o), the number of resolvable multipaths (L), the variance of energy corresponding to the first multipath component (σ_1^2), the FEC parameter t, the RLC/MAC-layer frame length (N_p) and RLC/MAC-layer backoff parameters (i.e., s_{LL}, q_{LL}, a, b, μ, p_{min}, $p(1)$), the *MSS* (Maximum Segment Size), which refers to TPDU size in terms of number of RLC/MAC layer frames, and the transport protocol parameters related to the timer mechanism (i.e., q_{TL}, c, RTO_{min}, RTO_{max}). It is to be noted that, since service differentiation or priority among different mobile stations is not being considered here, E_b/N_o at the receiver is assumed to be the same for all users. Since all users are using the same transmission rate, this would induce approximately the same average power for all users at the receiver.

Values of some of the simulation parameters used for obtaining the simulation results presented in this chapter are shown in Table 4.1. The value of the parameter p_{min}, in the case of *AIMD* and *MB* schemes, is chosen to be 0.005. In a practical networking scenario, this value may be chosen based on the factors such as network capacity and user delay requirements. Simulation results are obtained while keeping the value of the transport protocol timer backoff parameter q_{TL} fixed to 0.5.

4.4 Simulation Results and Discussions

Some simulation experiments have been performed to observe the impact of the different access control protocol parameters on system performance without packet combining, assuming a threshold model of MAI (as in [52]) along with the presence of

Table 4.1. *Simulation parameters.*

Parameter	Value	Parameter	Value
Channel rate	9.6 *Kbps*	Slot size/N_p	0.02 s/192 *bits*
TPDU size	24 *LL* frames	RTO_{min}, slots	10 × TPDU size
RTO_{max}, slots	50 × TPDU size	N_c	63
E_b/N_o	20 dB	L, $2\sigma_1^2$	3, 7 dB
a, b (*AIMD*)	0.5, 0.25	s_{LL} (*TEB*)	8
q_{TL}, q_{LL} (*TEB*)	0.5	$p(1)$ (*AIMD* and *MB*)	0.5
μ (*MB*)	100	p_{min} (for *AIMD* and *MB*)	0.005

correlated multipath fading [20]. It has been observed that, in the case of *additive increase and multiplicative decrease (AIMD)*-based access control, the RLC/MAC-layer throughput is almost independent of the initial value $p(1)$ over a wide range of numbers of active mobile stations (N). Also, for a particular set of values of a, b and MAT[3] (Multiple Access Threshold), p_{min} can be properly chosen to have a high RLC/MAC-layer throughput.

With *μ-law backoff (MB)*-based access control, the variation in RLC/MAC-layer throughput with the number of active mobile stations depends primarily on p_{min} and the variation in the values of μ and $p(1)$ does not affect the throughput performance remarkably. For relatively large values of N, the RLC/MAC-layer throughput deteriorates as p_{min} increases. With *truncated exponential backoff (TEB)*-based access control, a reasonably high throughput performance is observed over a wide range of values of N as long as the values of the RLC/MAC-layer backoff parameters (i.e., q_{LL}, s_{LL}) are properly chosen. Details of these results will be reported somewhere else.

The values of the backoff parameters for the different retransmission control schemes assumed here are based on the above observations.

[3]It denotes the maximum number of simultaneous transmissions that allows the frame error probability to remain below some desired threshold.

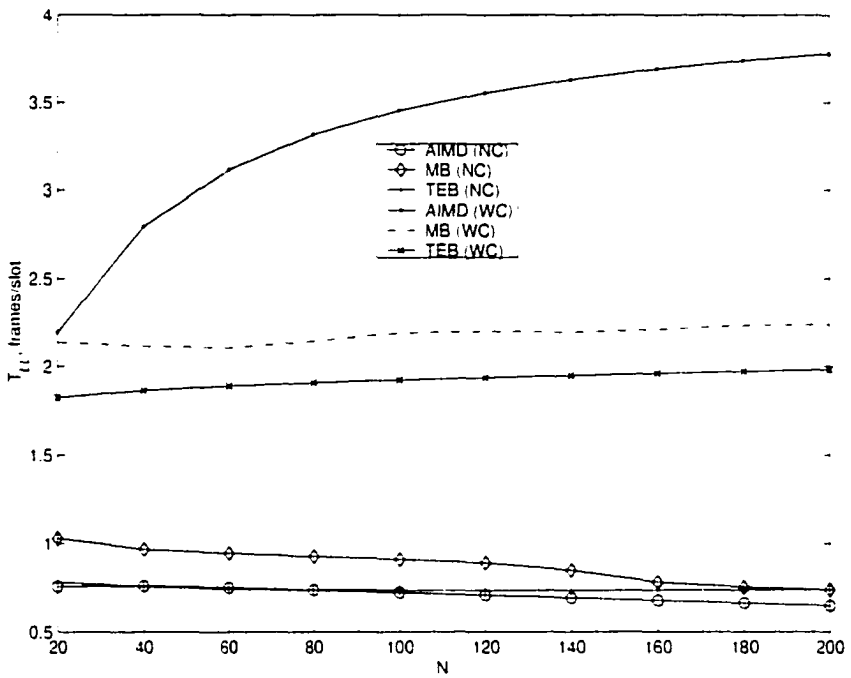

Figure 4.3. *T_{LL} for the different access control schemes (for t = 2; $NC \equiv$ no combining, $WC \equiv$ with combining).*

4.4.1 Effect of Diversity Combining on T_{LL}

As expected, diversity packet combining increases the RLC/MAC-layer throughput significantly and improves the throughput stability (Figure 4.3). Among the three access control schemes, the *additive increase and multiplicative decrease (AIMD)* scheme outperforms all of the other schemes except in those cases where the number of active mobile stations is relatively small. This is intuitive since, for the parameters chosen and for the case of low traffic load (i.e., small N), the additive increase of retransmission probability for the *AIMD* scheme may not result in a system load that is sufficiently high to achieve higher throughput. However, with a relatively large number of active mobiles, the additive increase (in case of success) and multiplicative decrease (in case of failure) of retransmission probability can effectively adjust the traffic load to offer a high throughput. The concept here is similar to that of the TCP window adjustment in the case of network congestion.

Since *μ-law backoff (MB)*-based retransmission control can adjust the value of re-

transmission probability more rapidly. for a relatively low traffic load. the RLC/MAC-layer throughput (T_{LL}) is higher compared to the throughput for the *AIMD* scheme. *MB*-based access control provides greater throughput than the *truncated exponential backoff (TEB)* scheme. It is observed that. without packet combining. the *MB* scheme provides the best RLC/MAC-layer throughput performance over a large range of system load. This is due to the interrelation between the increased channel reliability and the dynamics of the retransmission control policies. In this case. an increased channel reliability affects the performance of *AIMD*-scheme more positively than in the case of the other schemes. This illustrates the effects of physical layer techniques on higher layer protocol mechanisms.

The RLC/MAC-layer performance (and hence transport-layer protocol performance) is found to be poor even with large values of t (e.g.. for $t = 10$) when combining is not performed (Table 4.2). For example. with *AIMD*-based access control. the RLC/MAC layer throughput $T_{LL} \approx 3.9632$ without combining even when t is as large as 10. while with combining $T_{LL} \approx 3.1802$ (which is approximately 20% smaller than the former value) for $t = 2$ only. In other words. for the same value of t. the achieved throughput is much higher (depending on the access control policy) with combining compared to the case without combining. The implication is that. by performing diversity combining at the BS receiver. very low-rate channel coding computations at the mobile terminals may be avoided with a consequent decrease in battery power consumption to some extent. In this case. complex decoding operations at the BS receiver can be also avoided. This is. of course. at the expense of increased complexity (due to packet combining operation) at the BS physical layer. This observation leads to alternative physical layer design choices where network designers can opt for receiver structures with some compromise between FEC decoding hardware and packet combining hardware.

4.4.2 Effect of Diversity Combining on T_{TL}

Typical variations in transport protocol throughput (T_{TL}). with the number of active mobile stations (N). for *exponential backoff (EB)*-based timer management for the different RLC/MAC-layer retransmission control policies are shown in Figure 4.4. For the assumed values of the simulation parameters. with diversity packet combining.

Table 4.2. *T_{LL} with FEC with/without diversity packet combining (for N = 60: NC ≡ no combining, WC ≡ with combining).*

	AIMD		MB		TEB	
t	NC	WC	NC	WC	NC	WC
0	0.1820	1.7844	0.1417	0.8341	0.1124	0.5846
1	0.4633	2.5525	0.6641	1.4819	0.3908	1.2837
2	0.7568	3.1802	0.9621	2.1356	0.7579	1.9091
5	1.7850	4.8055	2.1955	4.0282	1.9437	3.7077
10	3.9632	7.3017	4.4078	7.2345	4.2520	6.8186

the *AIMD* scheme provides the highest throughput except when the number of active mobile stations is relatively small. Notable is the stable transport-layer throughput performance for the different retransmission control schemes over a wide range of numbers of active mobile stations.

When packet combining is not performed, *MB*-based retransmission control provides the highest throughput performance over a wide range of system load (i.e., N) while the *AIMD* scheme may even perform poorer than the *TEB* scheme. Similar results are obtained with some other values of E_b/N_o (e.g., for $E_b/N_o = 30\ dB$).

The efficacy of local adaptation of retransmission control probability over global adaptation for all the three schemes is evident from Figure 4.5. The reason behind this is that, in the case of transmission failure, for global adaptation, all the mobile stations, including those which have not transmitted in the recent timeslot delay their transmissions and this causes a decrease in channel load compared to the channel capacity. It is observed that, with global adaptation, for all three access control schemes, the transport protocol performance improves monotonically with the system load.

4.4.3 Effect of *RTO* Control Mechanism

The retransmission timer control mechanism in the transport layer can affect transport protocol throughput (T_{TL}) significantly. The impact of different timer manage-

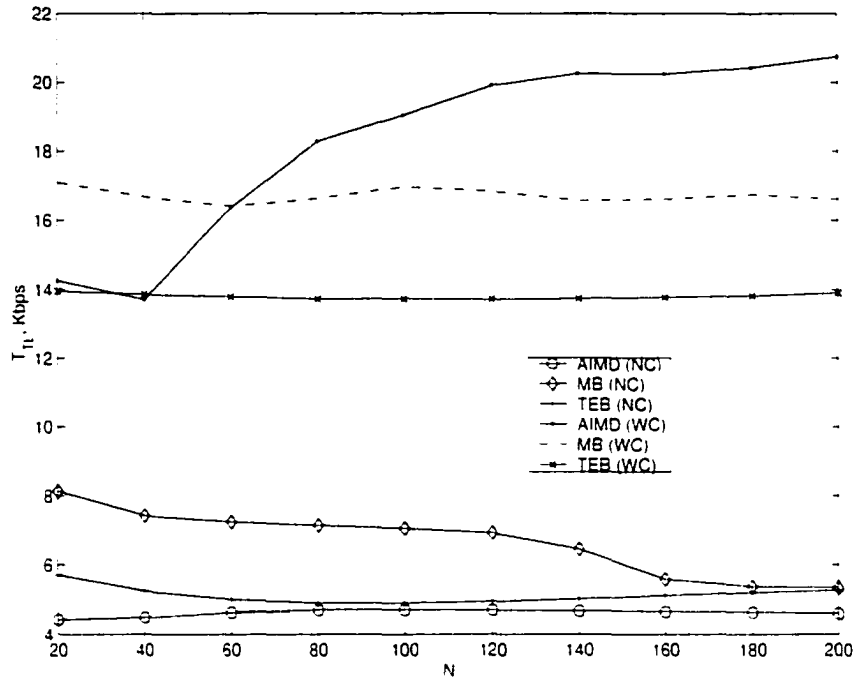

Figure 4.4. *T_{TL} for the different link-layer access control schemes (for t = 2, MSS = 24 and EB-based RTO control: $NC \equiv$ no combining, $WC \equiv$ with combining).*

ment mechanisms on transport protocol throughput varies depending on the underlying RLC/MAC-layer retransmission control scheme. Simulation results show that, for the chosen protocol parameters, with an *additive increase and multiplicative decrease*-based access control, an improved throughput performance is achieved for *exponential weighted moving averaging (EWMA)*-based *RTO* control mechanism and the performance improvement becomes more significant with increased channel reliability (i.e., with an increase in the value of FEC parameter t).

With *μ-law backoff (MB)*-based retransmission control, the *exponential backoff (EB)*-based RTO control offers a higher transport protocol throughput. It is observed that, with *truncated exponential backoff (TEB)*-based access control, the difference in transport protocol throughput for *EB* and *EWMA*-based *RTO* control mechanisms is smaller compared to the difference in the case of *AIMD* and *MB* schemes. Throughput is found to be higher with *EWMA*-based *RTO* control when the number of active mobile stations (N) is relatively small while throughput with *EB*-based *RTO* control is higher for relatively larger number of active mobile stations.

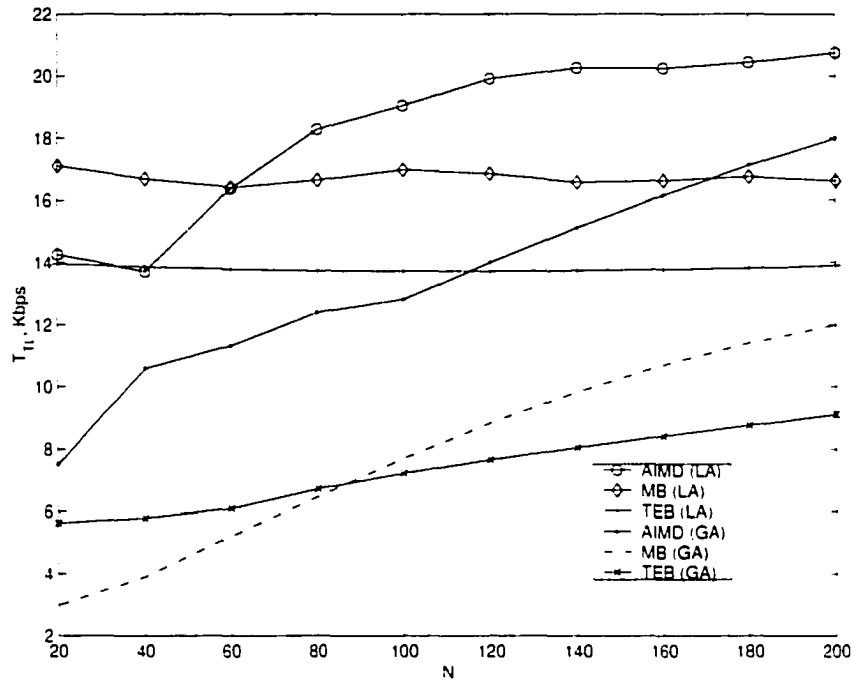

Figure 4.5. *T_{TL} for the different link-layer access control schemes (for t = 2, MSS = 24 and EB-based RTO control: GA ≡ global adapatation, LA ≡ local adaptation).*

The throughput performance for *EB*-based *RTO* control is sensitive to the value of RTO_{min}, particularly for *MB* and *TEB*-based access control schemes. The transport protocol throughput deteriorates as the value of RTO_{min} is decreased. Simulation results show that this deterioration is more evident for relatively lower values of t and/or N.

4.4.4 Sensitivity of T_{TL} to c

For all three access control schemes, simulation results show that, with *EWMA*-based *RTO* control, the transport protocol throughput (T_{TL}) improves as the value of the low-pass filter gain parameter c (which is used for RTT estimation) is reduced (Figure 4.7). The transport protocol throughput is observed to be more sensitive to variations in c for *AIMD*-based access control.

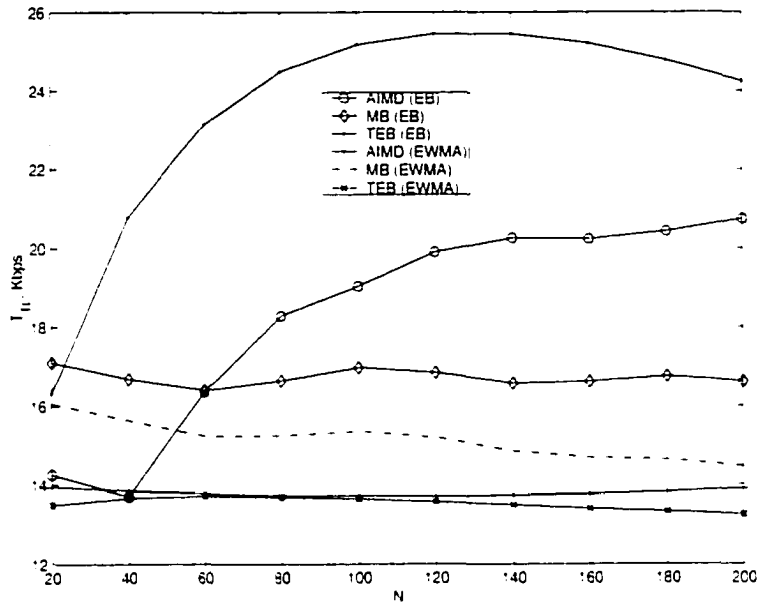

Figure 4.6. *T_{TL} with EB-based and EWMA-based RTO control (for t = 3, c = 0.3, MSS = 24).*

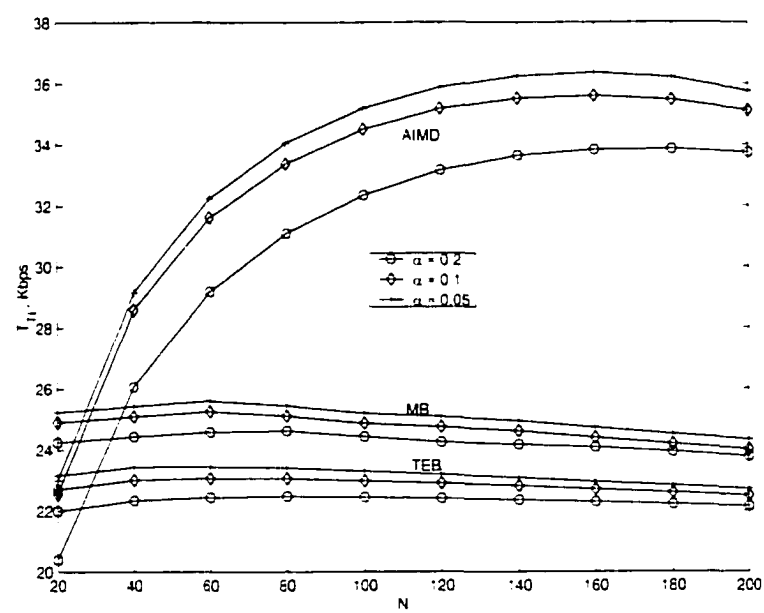

Figure 4.7. *Sensitivity of T_{TL} to c for EWMA-based RTO control (for t = 5, MSS = 24).*

4.4.5 Sensitivity of T_{TL} to Variation in MSS

To evaluate the effect of varying the TPDU size (MSS) on transport-layer protocol throughput (T_{TL}), simulation results are derived for different values of MSS for the different access control schemes. In the case of *additive increase and multiplicative decrease (AIMD)*-based access control, *exponential weighted moving averaging (EWMA)*-based *RTO* control is assumed in the transport layer while, for the two other access control schemes, *exponential backoff (EB)*-based *RTO* control mechanism is assumed. For different values of *MSS*s, the parameters RTO_{min} and RTO_{max} are set to $10 \times MSS$ and $50 \times MSS$ RLC/MAC-layer frame transmission time, respectively (Table 4.1). It is observed that the transport protocol throughput does not vary remarkably with varying TPDU size for the different underlying access control schemes (Figure 4.8).

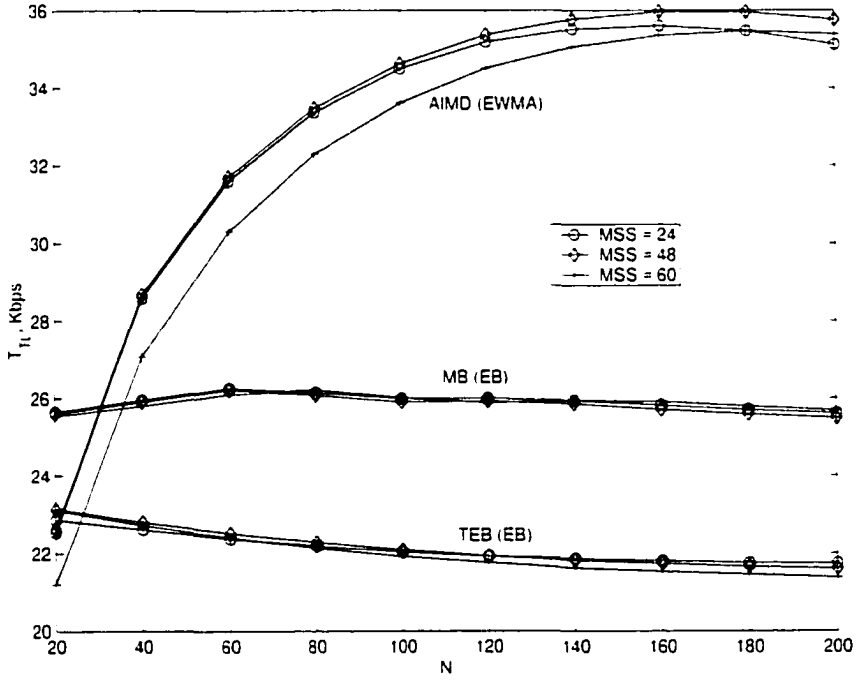

Figure 4.8. *Effect of variation of* MSS *on* T_{TL} *(for* $t = 5$, $c = 0.1$*).*

4.4.6 Effect of FEC Coding on T_{TL}

To analyze the relative performance of the different access control schemes under different forward error correction coding, simulation results are derived for different values of t while keeping the value of E_b/N_o constant. Some typical results are shown in Figure 4.9. For the chosen protocol parameters, *additive increase and multiplicative decrease (AIMD)*-based access control coupled with *exponential weighted moving averaging (EWMA)*-based *RTO* control provides the best throughput performance over a wide range of values of N except when N is relatively small. For relatively small values of N, the transport-layer protocol throughput (T_{TL}) is found to be higher for *μ-law backoff (MB)*-based access control scheme.

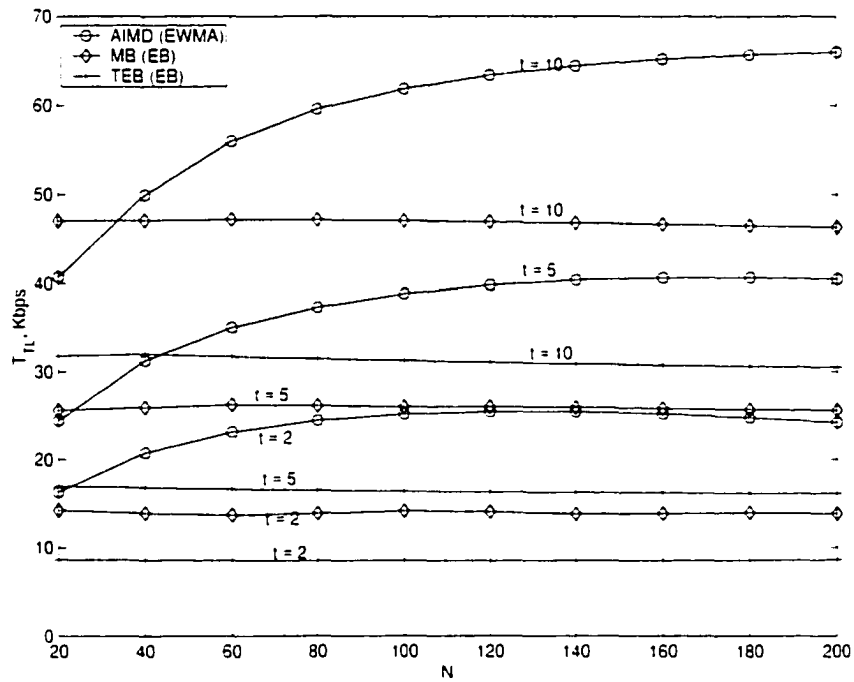

Figure 4.9. *Effect of FEC coding on T_{TL} (for ϵ = 0.2, MSS = 24).*

4.4.7 Effects of Channel Dispersiveness and the Number of Multipaths

For exponential multipath intensity profile (i.e., when interfering multipaths' signal power decay exponentially with corresponding time delay), throughput is higher in a DS/CDMA S-ALOHA network for higher decaying factors (i.e., when channels are less dispersive) [43]. Therefore, the performance of the different access control schemes under different dispersive channel scenarios will follow the same trend as in the case of FEC coding with different t.

Similar comments apply regarding the effect of L, the number of resolvable multipaths. The channel quality deteriorates with increasing L since the unresolved multipaths add to the multiple access interference. Simulation results for $L = 2, 3, 4$ are shown in Figure 4.10 for the three access control schemes under diversity packet combining. As is evident from the figure, the system performance under *additive increase and multiplicative decrease (AIMD)*-based access control improves more rapidly with an increase in channel reliability (e.g., a decrease in the value of L) compared to the performance of the other schemes.

4.4.8 Unreliable TP Performance

The RLC/MAC-layer throughput (T_{LL}) and frame dropping probability (P_{drop}) have been evaluated for different values of delay tolerance limit (D_l) under different access control schemes. Some typical simulation results are shown in Figure 4.11. It is observed that, although the differences among throughputs are not significant, the values of frame dropping probability differ remarkably for the different access control schemes. For the chosen protocol parameters, among the three access control schemes, with the *additive increase and multiplicative decrease (AIMD)* scheme, P_{drop} is the least for a particular value of D_l. The performance of *μ-law backoff (MB)*-based access control scheme is better than the performance of *truncated exponential backoff (TEB)*-based access control and the improvement is not significant when the value of D_l is relatively small (e.g., 1 s).

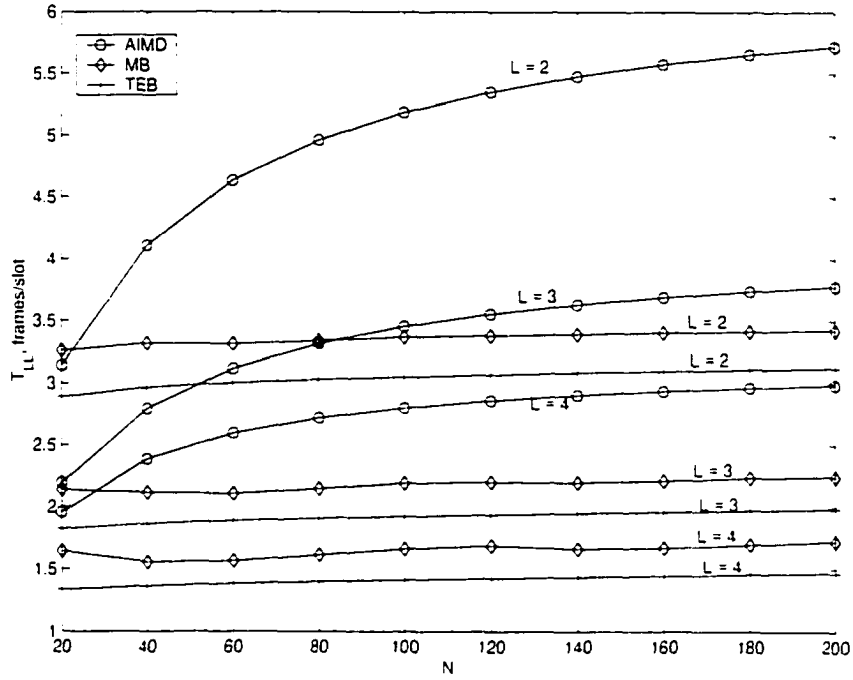

Figure 4.10. *Effect of L on T_{LL} (for t = 2, MSS = 24).*

4.5 Chapter Summary

In this chapter, the RLC/MAC-layer and the transport-layer protocol performances have been analyzed in a DS/CDMA S-ALOHA system with retransmission diversity packet combining under frequency selective Rayleigh fading channels. The performance results for three different RLC/MAC layer access control schemes have been presented. Two transport layer timer control mechanisms have been described and their effects on system performance have been analyzed. The system performance under single-level error recovery has also been evaluated for the three different RLC/MAC layer access control mechanisms. The results presented in this paper provide insight into the identification of appropriate higher layer protocol mechanisms, which would be required for transmission protocol stack performance optimization in DS/CDMA S-ALOHA systems.

The following provides a brief summary of the key results:

- Diversity combining significantly affects the relative performance of RLC/MAC-layer access control techniques in terms of RLC/MAC-layer and hence transport-

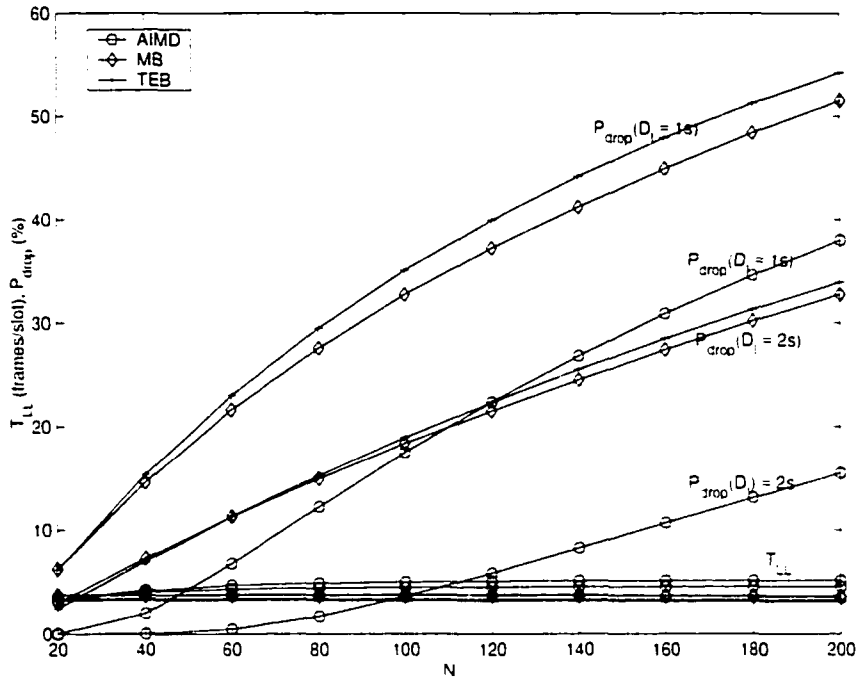

Figure 4.11. *Variation of T_{LL} and P_{drop} with N for the different access control schemes (for t = 5, MSS = 24).*

layer protocol throughput.

- The performance of the three access control schemes described above is always better with localized adaptation (i.e., distributed control) than with global adaptation (i.e., centralized control). *AIMD*-based access control provides the highest system throughput over a wide range of system load, except in cases where system load is low.
- In the case of two-level error recovery through a reliable transport protocol, the achieved throughput depends on the transport protocol timer control mechanism and a suitable mechanism can be identified for an underlying link-layer access control scheme. Different timer control mechanisms react differently to different access control schemes. With *AIMD*-based access control, the transport-layer throughput is higher with *EWMA*-based *RTO* control mechanism while, for *MB* and *TEB*-based access control, a higher throughput is achieved with *EB*-based *RTO* control.

- With diversity combining employed at the BS receiver, computations for very low-rate channel coding may be avoided with a consequent improvement in power conservation at the mobile terminals to some extent. This also leads to some alternative physical layer design choices.

Similar types of inter-layer protocol dependencies can be observed in systems employing narrowband modulation. Similar conclusions, therefore, may also be drawn for some non-CDMA systems (e.g., EGPRS (Enhanced General Packet Radio Service)) that employ physical layer performance enhancement techniques such as dynamic adaptive modulation and coding.

Chapter 5

Fair Bandwidth Allocation Through TDMA-Based Centralized MAC Protocols in Wireless IP Networks

5.1 Introduction

The introduction of DS (Differentiated Services) architecture [53] into the current Internet aiming at providing 'scalable service differentiation' is under way. The necessity to provide a seamless service to mobile users has been motivating the research on DS wireless networks ([54]-[55]). Bandwidth management for fair bandwidth allocation among traffic flows[1] will be a key requirement of the MAC protocols in the emerging DS wireless IP networks. Bandwidth allocation is done by a scheduling mechanism in the MAC protocol. The scheduler and the admission controller in the base station in combination with the packet classifiers in the mobile stations can effectively manage traffic in a wireless link to maintain the SLAs (Service Level Agreements) between a wireless domain and its wired counterpart and thus provide necessary QoS to the different traffic flows.

In this chapter, we propose a unified TDMA-based centralized MAC framework for *fair* bandwidth allocation among the competing uplink data flows in a wireless

[1] A *flow* is a transmission session of packets that are generated from the same source and destined to the same destination.

IP network in the presence of bursty channel errors. This unified scheme includes an admission control policy, a traffic conditioning scheme and a dynamic bandwith allocation mechanism. We consider both *packet-level* (or *slot-by-slot*) scheduling and *burst-level* scheduling for dynamic bandwidth allocation. Here, the notion of *fairness* means that the level of throughput achieved for the different flows over a relevant time interval would be proportional to their subscription levels (or profile rates), provided that the average traffic load generated by each flow is proportional to its profile rate[2]. This implies that almost the same average throughput would be achieved for flows with the same profile rate.

Extensive simulations with different flow *grain-size*[3], different subscription levels of the wireless link bandwidth, different composition of flows and different frame-lengths show that this fairness is robust against variations in the number of flows, flow grain-size, TDMA frame-length and channel-error correlation. The energy efficiency of the resulting MAC scheme is evaluated in terms of the mobile station radio transceiver usage while in *transmit* and in *receive* mode for both the packet-level and the burst-level bandwith allocation.

The organization of the rest of the chapter is as follows. Section 5.2 presents the background and the motivation of the work. The system model is described in Section 5.3. In Section 5.4, we present the proposed bandwidth allocation framework. The simulation model and the assumptions used for performance evaluation of the proposed bandwidth allocation framework are presented in Section 5.5. The simulation results are presented in Section 5.6. Finally, in Section 5.7, the main results are summarized.

5.2 Background and Motivation

Throughput fairness among flows is essential in a wireless data network since service providers are expected to prorate the available link bandwidth to the clients in proportion to their subscription level in the presence of bursty channel errors. Wireless fair queueing is one way to provide fair access to the shared wireless channel. The

[2]The maximum load generated by each flow is upper bounded by its profile rate.

[3]Here, the ratio of the contracted rate to link data rate is referred to as the *grain-size* of a flow.

goal of wireless fair queueing algorithms is to make the wireless channel errors transparent to users by allocating channel dynamically over small time scales. A unified wireless fair queueing architecture consists of an error-free service model, a lead and lag model, a compensation model and a method for channel monitoring and prediction [56]. Since the wireless fair queueing concepts are now fairly well understood (e.g., [56]-[62]), the issues concerning their efficient implementation in popular MAC frameworks need to be addressed.

In this chapter, we propose a 'unified' dynamic bandwidth allocation framework that provides throughput fairness among flows in a TDMA-based wireless access scenario. The proposed framework is 'unified' in the sense that it integrates dynamic bandwidth allocation (which most of the conventional MAC protocols (e.g., [63]-[65]) primarily deal with) with an admission control policy and a traffic conditioning scheme. The resulting MAC protocol provides throughput fairness which is robust against the number of admitted flows, their profile rates, data burst inter-arrival time, maximum tolerable delay of data bursts corresponding to the flows, TDMA frame-length and channel-error correlation. In addition, the proposed scheme can be used in an adaptive QoS framework for dynamically adjusting the QoS[4] for flows in order to accommodate dynamic variations in wireless link resources.

Since the mobile stations will always have limited power, energy conservation in the mobile stations is an important issue in the design of MAC protocols [66]. In a DS wireless IP network, the functionalities such as packet marking and policing are expected to be performed in the MSs with a consequent increase in the processing power requirement. Therefore, along with fairness, energy efficiency should be considered as another criterion for evaluating the performance of a MAC protocol for wireless IP networks. The effects of the channel-error correlation and the different protocol parameters on the energy efficiency of the resulting MAC protocol are evaluated for both packet-level and burst-level scheduling under the proposed bandwidth allocation framework. This enables us to better understand the different performance tradeoffs and to get insight into the identification of suitable scheduling mechanism for fair bandwidth allocation. The issue of energy efficiency has often been ignored in the wireless fair queueing algorithms [56] (e.g., IWFQ (Idealized Wireless Fair Queueing).

[4]Here, the MAC-level throughput is the primary QoS measure.

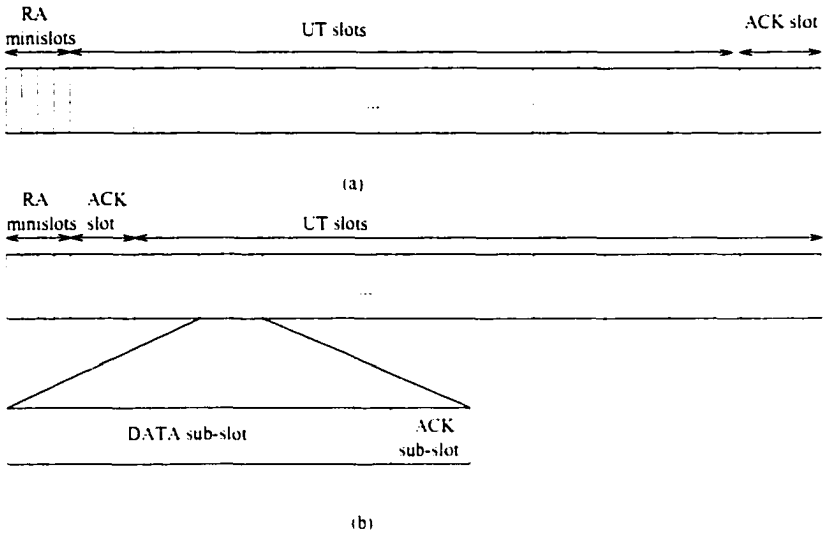

Figure 5.1. *TDMA frame structure for (a) burst-level scheduling, and for (b) packet-level scheduling.*

CIFQ (Channel Condition Independent Fair Queueing), SBFA (Server Based Fairness Approach) and WFS (Wireless Fair Service Algorithm)) proposed in the literature.

5.3 System Model and Description

We assume two different frame structures for burst-level and packet-level scheduling as shown in Figure 5.1. Each TDMA frame consists of a RA (Request Access) slot which is comprised of N_r minislots, N_s UT (Uplink Transmission) slots and an ACK (Acknowledgement) slot.

Each flow f_i has a contracted profile rate (or target rate) R_i. Applications may view the rate R_i as the one necessary to provide a specific level of service to a flow. In a DS network, the value of this parameter can be contracted as a part of the SLA. For each new data burst, a mobile station transmits the reservation request information, such as the queue-length status in one of the minislots during the RA slot. S-ALOHA-based contention resolution is assumed in the reservation minislots. A mobile terminal may have more than one flow active at a time.

In the case of burst-level scheduling, after transmitting the reservation request information, an MS can switch to a more power-conservative *standby* mode. After

the reservation request is successfully received, the BS registers the corresponding flow as 'active' for service and stores the reservation information. It then executes the scheduling algorithm (as described later) to compute the allocation of UT slots among the 'active' flows in the next frame.

For burst-level scheduling, the slot allocation information (e.g., total number of UT slots, UT slots allocated to each flow and their positions in the frame) is broadcasted by the BS in the ACK slot along with the acknowledgement information for the packets transmitted in the UT slots in the current frame after the UT slots in the current frame pass by (Figure 5.1(a)). The MSs in the *standby* mode switch to *receive* mode at the start of the ACK slot. This strategy will cause at most one transceiver turnaround between transmit and receive mode during a frame-time.

If a new data burst arrives while a mobile is active in transmitting the previous data burst, it can send the new buffer information to the BS without going through the reservation process. In this case, the packets corresponding to the previous data burst are assumed to be lost.

For packet-level (i.e., slot-by-slot) scheduling, the acknowledgements corresponding to the RA attempts are broadcasted during the ACK slot following the RA slot while the acknowledgement corresponding to a data packet transmission is broadcasted by the BS in the ACK sub-slot during each timeslot (Figure 5.1(b)) along with the scheduling information for the next timeslot. The length of each ACK sub-slot is assumed to be $\frac{1}{N_s}$ times the length of a timeslot. This implies that the information transmitted during a timeslot is reduced by $\frac{N_s-1}{N_s}$ in this case compared to that for the burst-level scheduling.

The bursty channel errors are accounted for by delaying transmission of data traffic corresponding to a flow for which the transmission channel is expected to be noisy and by scheduling transmissions from the flows for which the channel conditions are expected to be good. This is similar to the 'channel-state dependent scheduling' concept presented in [59]. The channel status information is exploited by the BS scheduler as follows: if the transmitted packet(s) is(are) not ACKed by the BS, the corresponding MS assumes that the packet(s) has(have) been lost and it starts transmitting 'probe' messages periodically during the RA period. Therefore, the 'probe' messages contend with the new reservation request messages in the RA minislots. Af-

ter the BS receives a 'probe' message correctly, it assumes that the channel condition has improved and it resumes scheduling the traffic corresponding to that flow again for transmission.

5.4 Traffic Conditioning and Scheduling for Fair Bandwidth Allocation

5.4.1 Admission Control Algorithm

A differentiated services enhanced wireless network would require an admission control mechanism based on active flow count to combat large throughput fluctuation due to large numbers of active flows and hence to provide a certain level of QoS to the admitted flows. The admission control algorithm that we consider here ensures that at any instant of time the sum of the contracted rates (or profile rates) of all the flows (i.e., $\sum_i R_i$) is always less than or equal to the targetted data rate in the channel. This simple admission control policy can be described as follows:

Procedure Admission_Control() {
/* R_i: Subscription level for flow i which is included as a part of SLA.
C: Total uplink channel capacity.
R: Sum of the contracted rates of all the currently admitted flows.
μ: Targetted link utilization level. $0 < \mu \leq 1$. */
If $(R + R_i) \leq \mu C$ {
$R = R + R_i$
admit flow i
}
Else
reject flow i
}

5.4.2 Traffic Conditioning at Mobile Terminals

The DS network architecture is based on traffic classification, conditioning and scheduling mechanism. Traffic classification marks the packets with appropriate DS code points. Traffic conditioners are used to police the traffic entering a DS domain to enforce conformance to SLA with downstream DS domain. Traffic conditioning is required to check against the greedy flows which may try to receive more bandwidth than the contracted bandwidth by sending more traffic. In a differentiated services enhanced wireless network, wastage of valuable link bandwidth available between the mobiles and the BS can be avoided by using traffic conditioners and shapers in the mobiles and by allowing the BS to adjust the traffic conditioning parameters at the mobiles. Marking the traffic originating from the MSs before they enter the wired DS domain is also advantageous in the sense that the MSs can more easily take the application's preferences into account during marking.

Upon the arrival of a traffic burst (of length b_i packets), the average rate estimator in the traffic conditioner computes the estimated number of arrivals (e_i packets) using the contracted profile rate for the corresponding flow (i.e., R_i) and the measured burst inter-arrival time (t_i). If $b_i < e_i$, then the corresponding flow accumulates $(e_i - b_i)$ as *token* for retaining packets in excess of the allowable burst-length in future. When the data burst-length exceeds the allowable length (i.e., estimated length plus the number of currently available *token*), the traffic conditioner can either drop the 'excess' packets or can mark them as *OUT profile* packets. Here, we assume that when the packet arrival rate exceeds the contracted rate, packets in excess of the number of *IN profile* packets are dropped.

The traffic conditioning algorithm can be described as follows:

Procedure Traffic_Conditioning() {

/* f_i: Flow i, $f_i(token)$: Token counter for f_i.

$f_i(q_i)$: Total *IN profile* packets in current data burst corresponding to f_i.

t_i: Measured burst inter-arrival time, L: Packet length. */

/* Initialization done after admission of f_i */

$f_i(token) = 0$

/* Traffic conditioning/shaping actions after arrival of a data burst */

$e_i = \lfloor (R_i \times t_i)/L \rfloor$

```
If (b_i > e_i) {
   If  (f_i(token) ≥ 0) {
      If ((b_i − e_i) ≤ f_i(token)) {
         f_i(q_i) = b_i
         f_i(token) = f_i(token) - (b_i − e_i)
      } /* If */
     Else {
      f_i(q_i) = e_i + f_i(token)
      f_i(token) = 0
     } /* Else */
    } /* If */
   } /* If */
  Else {
        f_i(q_i) = b_i
        f_i(token) = f_i(token) + (e_i − b_i)
  } /* Else */
}
```

5.4.3 The Bandwidth Allocation Algorithm

A certain number of slots a_i is allocated to flow f_i in each frame-time depending on the contracted profile rate R_i. Since data traffic is expected to be bursty, a mobile terminal may not have packets to transmit during each frame-time. The flow f_i accumulates a *credit* of a_i units during each frame-time and loses one unit of *credit* after a packet corresponding to f_i has been successfully transmitted. Scheduling is based on the *credit* values of the competing flows, where flows with higher *credit* values are prioritized over the others.

We consider both *burst-level* and *packet-level* scheduling for dynamic bandwidth allocation. For burst-level scheduling, the BS performs the scheduling computations only once during each frame-time and allocates timeslots to a flow within a frame-time in a contiguous fashion. The mobile terminals are informed of the slot allocation corresponding to frame t during the ACK period in frame $(t-1)$. For packet-level scheduling, the BS performs the scheduling computations during each timeslot and

allocates one timeslot to a flow at a time. This 'fine-grained' scheduling would presumably enable to utilize the link bandwidth more efficiently in the presence of bursty channel errors. On the otherhand, since the bandwidth is allocated in a slot-by-slot basis, the active MSs have to 'snoop' the channel during each timeslot and consequently the average number of transceiver turnaround per frame-time increases.

Let us denote by $f_i(credit)$ the credit parameter corresponding to f_i, and by $P(t)$ the set of flow-ids (corresponding to the flows for which the channel state is predicted to be good during frame t and $f_i(q_i) > 0$) arranged in the descending order of the $f_i(credit)$ values. In $P(t)$, among f_i with the same $f_i(credit)$ value the one with the highest $f_i(q_i)$ value precedes the others. Let N_s denote the number of UT slots in frame t and let n_i denote the number of admitted flows with profile rate (or subscription level) of R_i.

The burst-level bandwidth allocation policy can then be described as follows:

Procedure Burst_Level_BW_Allocation() {

/* Initialization done after admission of f_i */

$a_i = max(1, \lfloor N_s \times \frac{R_i * n_i}{\sum_i R_i * n_i} \rfloor)$

$f_i(credit) = a_i$

/* Distribute N_s slots among f_i in $P(t)$ */

$Z = N_s$, $i = 0$

While ($Z > 0$ and $i < |P(t)|$) {

$z = min(f_i(q_i), a_i)$

allocate z slots to f_i

$f_i(q_i) = f_i(q_i) - z$

$Z = Z - z$

$i++$

} /* While */

/* Distribute the remaining slots */

$i = 0$

While ($i < |P(t)|$ and $Z > 0$) {

$z = min(f_i(q_i), Z)$

allocate z slots to f_i in $P(t)$

$f_i(q_i) = f_i(q_i) - z$

```
        Z = Z - z
        i++
      } /* While */
  }
```

The packet-level scheduling policy can be described as follows:

```
  Procedure Packet_Level_BW_Allocation() {
    For each timeslot during a frame-time {
      Compute P(t)
      Allocate the next timeslot to the first-flow in P(t)
    }
  }
```

The BS scheduler updates the value of $f_i(credit)$ for each f_i during each frame-time as follows:

```
  Procedure Update_Credit() {
    f_i(credit) = f_i(credit) + a_i, ∀ i /* during frame t */
    For (j = 1 to N_s) { /* during all slots in frame t */
      If (a packet is successfully transmitted during slot j)
        decrement the credit value of the corresponding flow
    } /* For */
  }
```

For FFQ (Fluid Fair Queueing), when a flow has nothing to transmit during a time window $[t, t + \Delta]$, it cannot reclaim the channel capacity that would have been allocated to it during $[t, t + \Delta]$ if it were backlogged at t. But in the scheme described above, every flow accumulates credit equal to its share in each frame-time, irrespective of whether or not it has packets to transmit. For a flow f_i which has been allocated slots in a certain frame-time, the minimum credit withdrawn during that frame time is $min(q_i, a_i)$. Since there is no bound on credit, a flow, which has accumulated large number of credits compared to the other flows, can transmit during several consecutive frame-time. It is to be noted that, even if a flow has accumulated large credit compared to one with similar subscription level, it will be allocated the maximum of a_i slots per frame-time. Therefore, it will not necessarily 'hog' the channel. Since traffic conditioning is used at the MSs, no flow should be considered

as 'misbehaving' to any other flow. Therefore, isolation among the flows is achieved in the sense that there is no misbehaving flow.

The long-term fairness in bandwidth allocation provided by the above credit-based scheduling can be appreciated by the following simple intuitive analysis. Let us denote by $w_i(t_2, t_1)$ and $w_j(t_2, t_1)$ the service received by f_i and f_j, respectively, between frame-time t_1 and t_2. Let the normalized bandwidth share per frame-time are a_i and a_j for f_i and f_j, respectively. Then, $f_i(credit) = a_i \times (t_2 - t_1)$ - $w_i(t_2, t_1)$ and $f_j(credit)$ $= a_j \times (t_2 - t_1)$ - $w_j(t_2, t_1)$. Since for the above credit-based scheme, the flow with the highest *credit* value is prioritized over the others (for both packet-level and burst-level scheduling), the algorithm tries to minimize $\mid f_i(credit) - f_j(credit) \mid, \forall i, j$ over an arbitrary time-window $[t_1, t_2]$. If we assume that, $\mid f_i(credit) - f_j(credit) \mid \leq \Gamma$, for some finite Γ, it follows after some algebraic manipulation that, $\mid \frac{w_i(t_2,t_1)}{a_i \times (t_2-t_1)} - \frac{w_j(t_2,t_1)}{a_j \times (t_2-t_1)} \mid$ is bounded. This bounded unfairness is observed (from the simulation results that will be described later) to be very small for flows with similar subscription level, while the relative unfairness for flows with different levels of subscription can also be achieved by controlling the system load through admission control.

The computational complexity of the scheduling algorithm is of $O(n)$ where n is the total number of admitted flows. Therefore, in a typical wireless setup, the processing overhead will not be significant at all. Obviously, packet-level scheduling would incur more processing and transmission overhead (and hence more power) in the BS compared to burst-level scheduling.

5.5 Simulation Model and Assumptions

5.5.1 Channel Model

Here, we assume that the slowly varying shadow and path losses in wireless links are perfectly compensated and that the channel quality variations due to multipath fading remain uncompensated. The multipath fading model is assumed to be same as that in Chapter 2. We assume that the fading envelope does not change significantly during the data block duration T. The value of T determines the minimum fade duration.

At a carrier frequency of 900 MHz, user speed of 5 km/hr, 20 km/hr and 40

km/hr correspond to f_dT values of 0.00833, 0.03333 and 0.06667, respectively, when the minimum fade duration $T = 2\ ms$. When the average block error rate $P_E = 0.1$ (i.e., channel fading margin, $F = 9.773\ dB$), the corresponding error burst-lengths (in terms of number of blocks) are 14.22390, 4.14930 and 2.16696, respectively. The error burst-lengths decrease with decreasing P_E. For a particular P_E and v, as the minimum fade duraion T increases, the error burst-length decreases.

5.5.2 Flow Models

5.5.2.1 Poisson Flow Model

We assume that a Poisson traffic source follows an ON/OFF model where the ON time and the OFF time are exponentially distributed. During each ON period, data is generated at a constant rate (i.e., at source coding rate). This model is applicable for packetized voice sources.

5.5.2.2 GBAR Flow Model

If X_{n-1} is a gamma distributed random variable with shape parameter β and scale parameter λ (i.e., $Ga(\beta, \lambda)$), A_n is a beta distributed random variable with parameters α, $\beta-\alpha$ (i.e., $Be(\alpha, \beta-\alpha)$), and B_n is $Ga(\beta-\alpha, \lambda)$, and they are mutually independent, then

$$X_n = A_n \circ X_{n-1} + B_n \tag{5.1}$$

defines a stationary stochastic process X_n with a marginal $Ga(\beta, \lambda)$ distribution which is called a GBAR(1) process whose autocorrelation function is given by $r(k) = (\frac{\alpha}{\beta})^k$, $(k = 0, 1, 2, \cdots)$. Since the mean and variance of a $Ga(\beta, \lambda)$ distribution are $\frac{\beta}{\lambda}$ and $\frac{\beta}{\lambda^2}$, respectively, if m and v denote the mean and variance of the data, then the estimated values of the parameters λ and β are: $\hat{\lambda} = \frac{m}{v}$ and $\hat{\beta} = \frac{m^2}{v}$. If it is assumed that data have the property: $r(k) \approx \rho^k$, $k = 1, 2, \cdots, K$ for some sufficiently large K, we have $\hat{\alpha} = \hat{\rho}.\hat{\beta}$.

It was shown that video traffic (e.g., real-time video-conferencing traffic) in isolation (i.e., traffic from a single source/flow) can be modeled accurately by an *Autoregressive Model* such as the GBAR model described above [67]. Here, we assume that a GBAR traffic source follows an ON/OFF model where the number of packets

generated during an ON period follows a GBAR(1) process and the OFF time is exponentially distributed. The number of packets generated during an ON period is found out by generating a non-integer value according to (5.1) and then rounding it to the nearest integer. For data generated from a H.261 compatible video coder that uses both DCT and motion compensation, some typical values for the case where head and shoulder scenes are shown with moderate motion and scene changes and with little camera zoom or pan is [67]: $m = 104.9$ ATM cell payload, $v = 882.09$ and $\rho = 0.984$.

5.5.2.3 LRD Flow Model

It has been observed that the traffic patterns generated by the web browsers are self-similar. For interactive TELNET traffic and bulk transfer traffic due to FTP, although the burst arrivals are well modeled as Poisson processes, the packet arrivals well fit to a distribution with a heavy upper tail (e.g., Pareto distribution) [68]. The degree to which the burst sizes are heavy tailed at the source level directly determines the degree of traffic self-similarity in the aggregate network traffic and is independent of the distribution of the burst inter-arrival times, of topology and of changes in network resources. In the LRD flow model that we consider here, the sources are of ON/OFF type, the number of packets generated during the ON period (i.e., burst-length) is Pareto distributed and the OFF time is assumed to be exponentially distributed.

The probability density function and the expected value of a Pareto-distributed random variable are given by

$$p(x) = \alpha . k^{\alpha} . x^{-\alpha-1}, \alpha, k > 0, x \geq k \quad and \quad E[X] = \frac{\alpha . k}{\alpha - 1}, \alpha > 1 \qquad (5.2)$$

where the parameter k represents the smallest possible value of the random variable. If $\alpha \leq 2$, the distribution has infinite variance and if $\alpha \leq 1$, the distribution has also infinite mean.

5.5.3 Measure of Energy Efficiency

Most of the battery power in a mobile station is consumed during *transmit* and *receive* mode while the least power is used when a mobile is in *standby* mode. Power consumption in each mode varies depending on the transceiver electronics [96].

Here, the energy efficiency of the MAC protocol is measured by the average *transmitter usage time* (e_t) and the average *receiver usage time* (e_r) required for successful transmission of a data packet. The energy spent in *standby* mode and in transceiver mode switching are not being considered here. The value of e_t for transmission of an RA packet/'probe' message is assumed to be $1/N_r$ slot[5]. The value of e_t for transmission of a data packet is 1 slot and $\frac{N_s-1}{N_s}$ slot for frame format (a) and (b) (Figure 5.1), respectively. For frame format (a), the value of e_r is assumed to be 1 slot while for frame format (b) the value of e_r is 1 slot and $1/N_s$ slot for ACK slot and ACK sub-slot, respectively.

The energy efficiency is largely affected by the TDMA frame-length, the channel quality, the channel-error correlation pattern and the traffic arrival pattern. For a particular channel quality and channel-error correlation pattern, since a finer granularity in channel state prediction can be achieved with packet-level scheduling, the average transmission power required for a successfully transmitted packet would be presumably lesser than that for burst-level scheduling, but the receiver usage for an active mobile may increase (depending on the frame-length and the channel-error burstiness) for packet-level scheduling due to continuous channel snooping by each active mobile.

5.5.4 Simulation Environment

A *C*-coded event-driven simulator is used for performance evaluation of the proposed bandwidth allocation framework. We obtain simulation results for both homogeneous and heterogeneous mixture of flows. In both the cases, the performance results are obtained for flows with different grain-size, for different values of TDMA frame-length, burst inter-arrival time (I), maximum tolerable delay limit (D_{max}), Pareto shape parameter α (for LRD flows) and for different channel quality (i.e., for different P_E) with different channel-error correlation pattern (i.e., for different values of f_dT).

For a fixed set of admitted flows, the system traffic load is modified by changing the length of the burst inter-arrival time (I) and/or the source coding rate and/or the the profile rate (R_i). We assume that, during the course of data transfer the wireless

[5]The actual amount of energy consumed (in joule) = $\frac{1}{N_r}$ × slot duration (in sec) × transmit-mode wattage.

Table 5.1. *Selected simulation parameters.*

Parameter	Value	Parameter	Value
Channel rate, R	2.0 $Mbps$	Slot size	0.0004 s
No. of RA minislot, N_r	4	Carrier frequency, f_c	900 MHz
Mean burst duration (for Poisson flows)	1 s	Mean burst length (for LRD flows)	20 packets
Mean burst length (for GBAR flows)	52.5 packets	Variance of burst length (for GBAR flows)	882.09
Correlation coefficient, ρ (for GBAR flows)	0.984	Source coding rate (for Poisson flows)	same as R_t

channels between the BS and the MSs suffer burst errors independently. The values of some of the simulation parameters are shown in Table 5.1.

The sequence of the events in our simulation environment can be described as follows:

i. When a data burst arrives, the burst-length is checked against the estimated number of packet arrivals using the corresponding R_t value and the measured burst inter-arrival time. When the number of packets in the data burst is greater than the sum of the corresponding estimated value and the *token* value, the excess packets are dropped; otherwise all the packets are considered as *IN profile* packets candidate for transmission. In both the cases, the value of the corresponding *token* value is adjusted accordingly.

ii. Timeslots within the TDMA frames are generated in a periodic fashion and, at the first timeslot, the contention resolution among the RA packets and the 'probe' messages in the RA minislots is simulated. In addition, slot allocations for the next frame-time are calculated and the *credit* value for each flow is increased based on the corresponding value of a_t. The link status corresponding to each flow is updated periodically after each T–the minimum fade duration.

iii. During each timeslot, a packet (from the flow for which that timeslot has been allocated) is transmitted only if the channel status is expected to be good. For successful transmission, the corresponding *credit* value is decremented. A local

variable toe_i is used to store the current TOE (Time of Expiry) value of a data burst corresponding to the flow f_i. For all the flows with non-zero values of q_i, the corresponding toe_i values are decremented. When any toe_i value reaches zero, the remaining packets of the corresponding data burst which have not yet been transmitted are discarded.

5.6 Simulation Results

The throughput achieved by flow i (T_i) is a function of R_i and the relative differences among the values of R_i (and hence flow grain-size), number of UT slots per TDMA frame N_s (and hence TDMA frame-length), traffic *burstiness*[6], total number of admitted flows, delay tolerance of data packets in a burst corresponding f_i and channel-error burstiness. The end-to-end throughput corresponding to a flow is affected largely by the packet marking strategies and the packet dropping policies adopted in the routers in the wired part of the DS Internet.

5.6.1 Homogeneous Poisson Flows

Typical variations in T_i for different values of frame-length, maximum tolerable delay limit (D_{max}) and burst inter-arrival time (I) are shown in Figures 5.2 and 5.3 for burst-level and packet-level scheduling, respectively. In this case, there are 30 admitted Poisson flows with different R_i values-R_i = 32 $Kbps$, 64 $Kbps$ and 100 $Kbps$ for $i = 0-14$, $i = 15-24$ and $i = 25-29$, respectively. It is observed that, the throughput variability among flows with the same value of R_i is very small irrespective of N_s, D_{max} and I. For a particular channel condition, although the achieved throughput for a flow depends on the values of R_i, N_s, I, D_{max} and the number of admitted flows, fair throughput share among flows with the same value of R_i is maintained even if the values of these parameters are changed.

The relative throughput fairness among flows with different profile rates is also achieved. For instance, simulation results show that, with $N_s = 10$, $I = 0.25\ s$ and $D_{max} = 1\ s$, $\frac{T_i}{T_j} \approx 0.5 = \frac{R_i}{R_j}$, for $i = 0 - 14$, $j = 15 - 24$; $\frac{T_i}{T_j} \approx 0.64 = \frac{R_i}{R_j}$.

[6]The arrival pattern of traffic within a flow determines its burstiness.

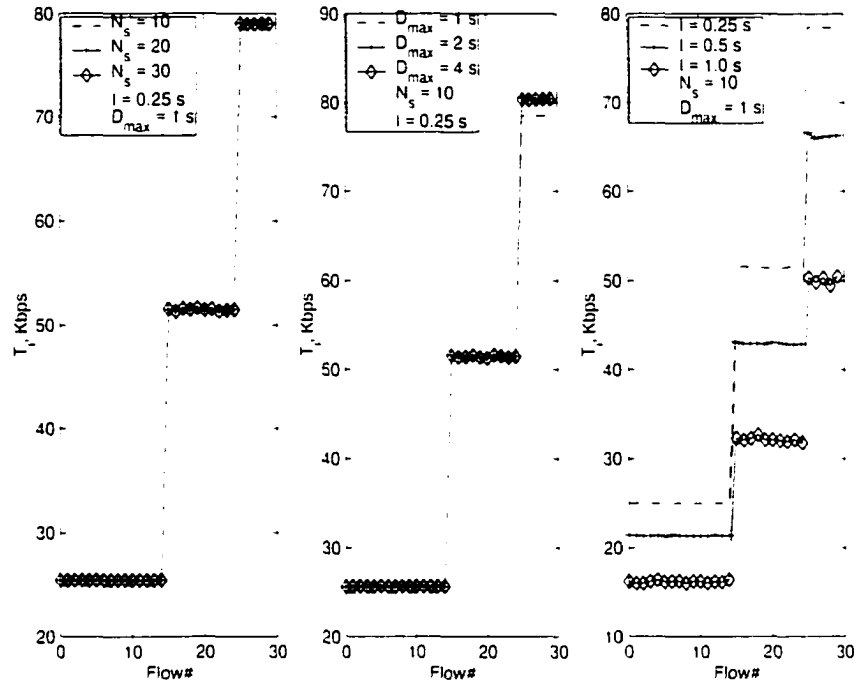

Figure 5.2. *Long-term throughput for Poisson flows under* burst-level *scheduling (with* $R_i = 32,\ 64,\ 100$ *Kbps for* $i = 0 - 14,\ 15 - 24$ *and* $25 - 29$, *respectively, and* $P_E = 0.1,\ v = 5\ km/hr,\ f_dT = 0.00833$*).*

for $i = 15 - 24,\ j = 25 - 29$: and $\frac{\Gamma_i}{\Gamma_j} \approx 0.32 = \frac{R_i}{R_j}$, for $i = 0 - 14,\ j = 25 - 29$. Under such conditions, for both burst-level and packet-level scheduling, the observed *channel utilization*[7] is 77%, 78% and 78% for $D_{max} = 1\ s,\ 2\ s$ and $4\ s$, respectively. When D_{max} is relatively lower and/or system load is higher, the relative throughput fairness among flows deteriorates.

It is observed that the *post facto* average packet delay (D_{avg}) and the average packet loss ratio (P_{drop}) *increase* as the frame-length and/or the burst inter-arrival time is(are) reduced. With increasing D_{max}, the value of D_{avg} increases, but P_{drop} decreases. The achieved throughput corresponding to a flow is higher for burst-level scheduling. The relative loss and delay performances for a flow under burst-level scheduling and packet-level scheduling depend on the corresponding profile rate

[7]Here, *channel utilization* is measured as the average number of packets successfully transmitted per timeslot.

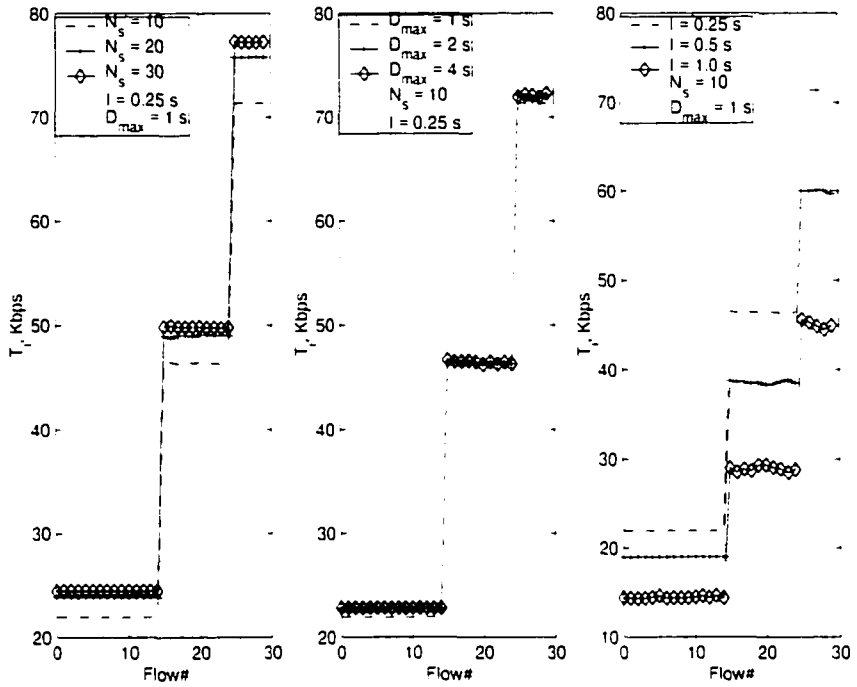

Figure 5.3. *Long-term throughput for Poisson flows under* packet-level *scheduling (with R_i = 32, 64, 100 Kbps for i = 0 − 14, 15 − 24 and 25 − 29, respectively, and P_E = 0.1, v = 5 km/hr, f_dT = 0.00833).*

(Figure 5.4). For both the cases, the channel-error correlation does not affect the throughput fairness but the average delay and loss performances *improve* as the channel-error correlation increases (i.e., when the terminal speed decreases).

For burst-level scheduling, the average *transmitter usage time* (e_t) and the average *receiver usage time* (e_r) increase (i.e., energy efficiency decreases) as the frame-length increases (Figure 5.5). On the other hand, for packet-level scheduling, the average transmitter and receiver usage time decrease with increasing frame-length. For both the cases, e_t and e_r increase as the data burst inter-arrival time decreases. The transmitter and receiver usage times *decrease* as channel errors become more correlated (Figure 5.6) which is due to the increased effectiveness of the channel-state aware scheduling for dynamic bandwidth allocation. The maximum tolerable delay limit D_{max} may not affect the energy efficiency significantly.

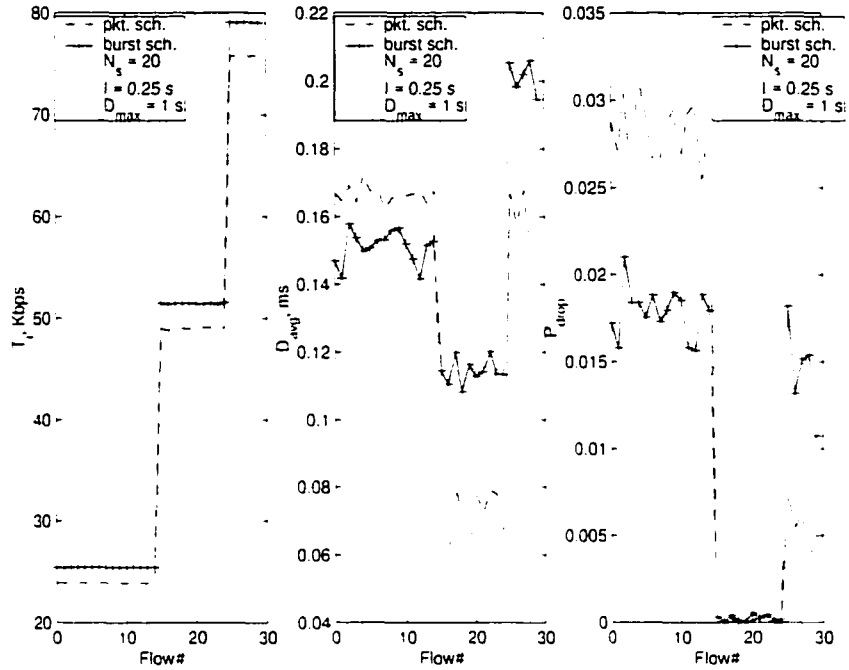

Figure 5.4. *Comparison among the throughput, delay and loss performances under burst-level and packet-level scheduling (for Poisson flows with R_i = 32, 64, 100 Kbps for $i = 0 - 14$, $15 - 24$ and $25 - 29$, respectively, and $P_E = 0.1$, $v = 5$ km/hr, $f_dT = 0.00833$).*

5.6.2 Homogeneous GBAR Flows

As in the case for Poisson flows, the throughput fairness among flows is maintained for different values of N_s, I, D_{max} and f_dT for channel utilization as high as 75% even when P_E is as large as 0.1. The effects of variations in N_s, D_{max}, I and f_dT on the average loss and delay performances and energy efficiency are similar as in the case of Poisson flows. The loss and delay performances improve as the system load is lowered either by increasing the burst inter-arrival time or by tightening the flow profile rate. For both burst-level and packet-level scheduling, the effects of N_s, D_{max}, I and f_dT on e_t and e_r are similar to those for Poisson flows.

The variations in T_i, e_t and e_r with f_dT are shown in Figure 5.7 for 30 flows with different R_i values-R_i = 200 *Kbps*, 250 *Kbps* and 300 *Kbps* for $i = 0 - 9$, $i = 10 - 19$ and $i = 20 - 29$, respectively. In this case, the channel data rate (R) and the

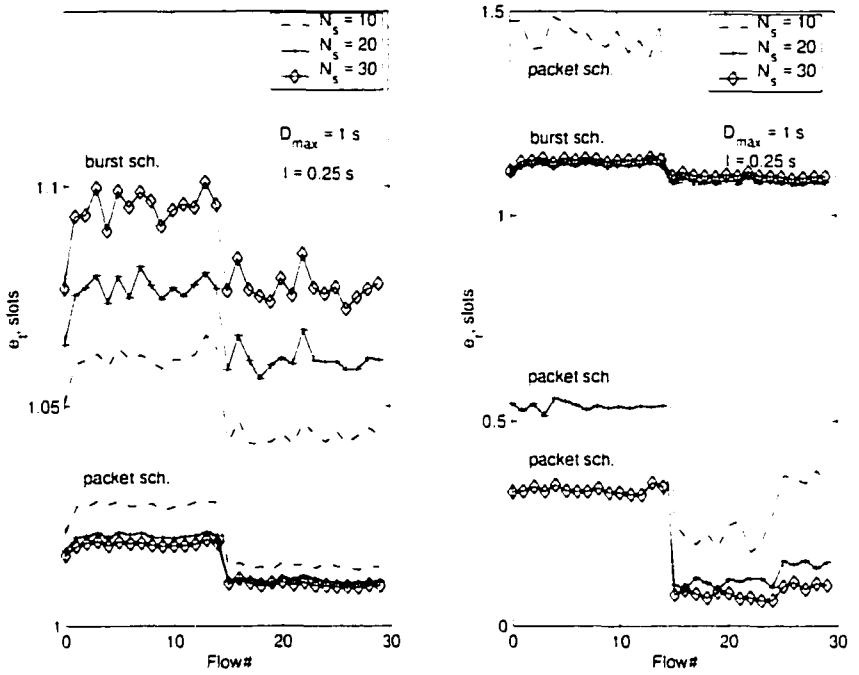

Figure 5.5. *The average* transmitter usage time *and the average* receiver usage time *under packet-level and burst-level scheduling (for Poisson flows with* $R_i = 32, 64, 100$ *Kbps for* $i = 0-14, 15-24$ *and* $25-29$, *respectively, and* $P_E = 0.1$, $v = 5\ km/hr$, $f_dT = 0.00833$*).*

average block error rate (P_E) are assumed to be 10 $Mbps$ and 0.05, respectively. The channel-error correlation does not affect throughput fairness, but the enegy efficiency improves with increasing channel-error correlation. The relative throughput for flows are observed to be proportional to their profile rates-$\frac{T_i}{T_j} \approx 1.18$, for $i = 20-29$, $j = 10-19$ (compared to $\frac{R_i}{R_j} = 1.20$): $\frac{T_i}{T_j} \approx 1.23$, for $i = 10-19$, $j = 0-9$ (compared to $\frac{R_i}{R_j} = 1.25$): and $\frac{T_i}{T_j} \approx 0.69$, for $i = 0-9$, $j = 20-29$ (compared to $\frac{R_i}{R_j} = 0.67$). The achieved channel utilization is approximately 82% in this case.

5.6.3 Homogeneous LRD Flows

In the case of LRD flows, the throughput fairness among flows with similar profile rate is largely affected by the value of the Pareto shape parameter α. It is observed

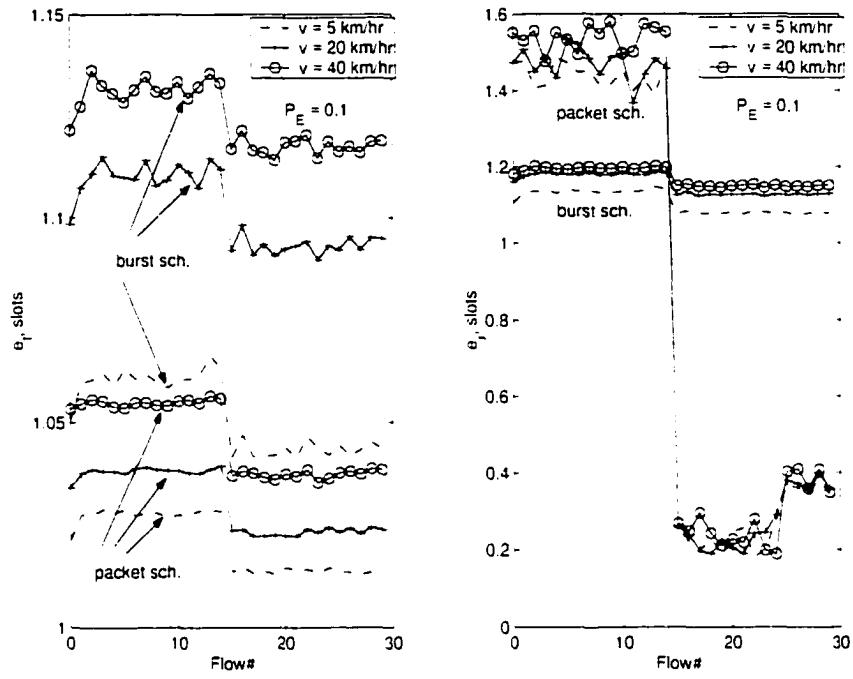

Figure 5.6. *The effects of channel-error correlation on the average transmitter and receiver usage times under packet-level and burst-level scheduling (for Poisson flows with* R_i = 32, 64, 100 *Kbps for* $i = 0 - 14$, $15 - 24$ *and* $25 - 29$, *respectively, and* $N_s = 10$, $I = 0.25$ *s*, $D_{max} = 1$ *s).*

that as α decreases, which results in an increased traffic burstiness, the packet loss ratio increases and hence the throughput corresponding to a flow decreases. Also, the variability in throughput increases among flows with the same value of R_i (Figure 5.8). The average delay and loss performances deteriorate, especially for burst-level scheduling, as the frame-length increases. For larger frame-length, the throughput, delay and loss performances are better for packet-level scheduling compared to those for burst-level scheduling. As D_{max} is increased, the average packet delay (D_{avg}) increases while the average packet loss ratio (P_{drop}) decreases. It is observed that, in the case of packet-level scheduling, the *post facto* loss and delay performances of a flow may deteriorate significantly as channel errors become more correlated, especially when the frame-length is relatively small. It may not be possible to achieve high channel utilization while providing throughput fairness in the case of bounded-delay LRD traffic.

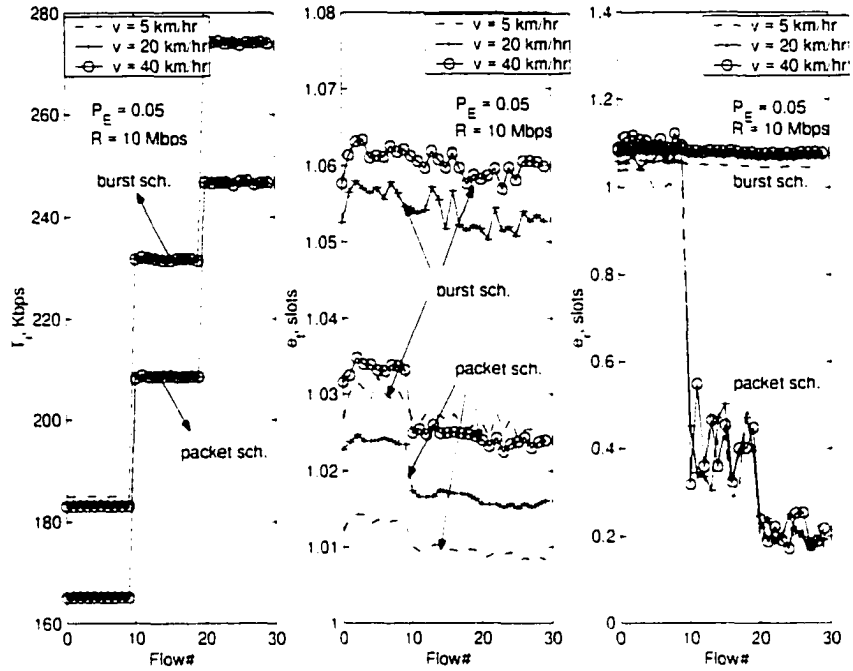

Figure 5.7. *The effects of channel-error correlation on the throughput, average transmitter and receiver usage times under packet-level and burst-level scheduling (for GBAR flows with R_i = 200, 250, 300 Kbps for $i = 0-9$, $10-19$ and $20-29$, respectively, and N_s = 10, I = 0.1 s, D_{max} = 1 s).*

As in the cases for Poisson and GBAR flows, for packet-level scheduling, the average transmitter usage time (e_t) and the average receiver usage time (e_r) increase as the frame-length decreases, while for burst-level scheduling, the average transmitter and receiver usage times increase as the frame-length increases. For both the cases, e_t and e_r increase as the traffic burstiness increases (i.e., α decreases).

5.6.4 Heterogeneous Flows

Typical simulation results demonstrating the throughput performance and energy efficiency in a heterogeneous scenario with Poisson, GBAR and LRD flows are shown in Figure 5.9. With 10 admitted flows of each type, we consider a total of 30 flows where $R_{0-9} = 64$ *Kbps*, $R_{10-19} = 100$ *Kbps* and $R_{20-29} = 20$ *Kbps* for Poisson, GBAR

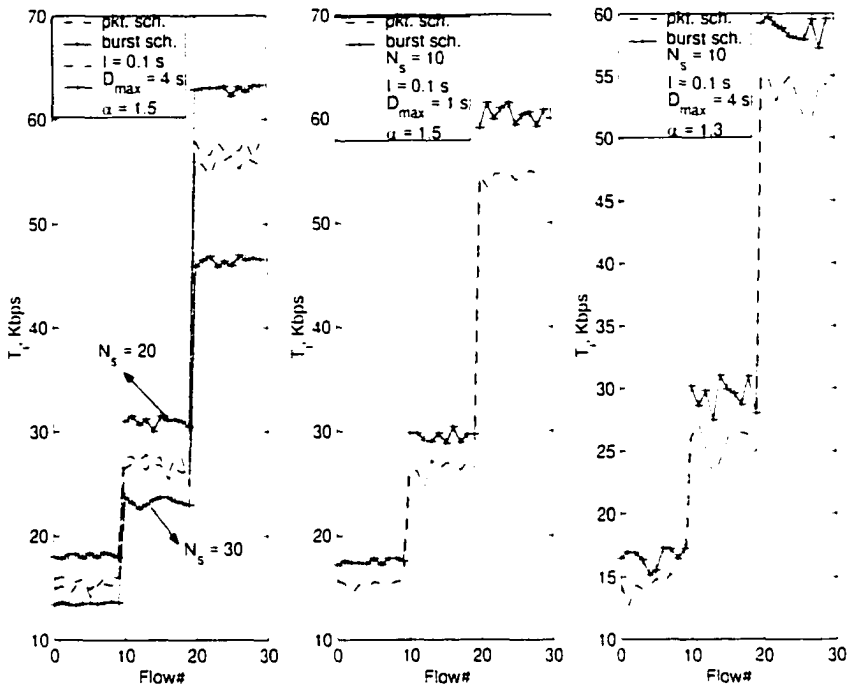

Figure 5.8. *Long-term throughput for LRD flows under packet-level and burst-level scheduling (for R_i = 20, 40, 100 Kbps for $i = 0-9$, $10-19$ and $20-29$, respectively, and $P_E = 0.01$, $v = 5$ km/hr, $f_dT = 0.00833$).*

and LRD flows, respectively. We assume $\alpha = 1.5$ and $I = 0.25$ s, 0.25 s, 0.1 s for Poisson, GBAR and LRD flows, respectively. The variability in throughput among flows with similar profile rates is considerably small. The relative throughput fairness is also considerably high in this case. For the assumed values of I, D_{max}, P_E and f_dT, it is observed that $\frac{T_i}{T_j} \approx 1.42$, for $i = 0-9$, $j = 10-19$ (compared to $\frac{R_i}{R_j} = 1.56$); $\frac{T_i}{T_j} \approx 4.73$, for $i = 10-19$, $j = 20-29$ (compared to $\frac{R_i}{R_j} = 5$); and $\frac{T_i}{T_j} \approx 3.36$, for $i = 0-9$, $j = 10-19$ (compared to $\frac{R_i}{R_j} = 3.20$). Under such condition, the observed channel utilization is approximately 78% and 77% for burst-level and packet-level scheduling, respectively.

Other parameters remaining the same, the channel utilization decreases with increasing frame-length. As α decreases, the throughput variability among the LRD flows and the *post facto* average packet loss rate for an LRD flow increase considerably. For Poisson and GBAR flows, the variability in loss and delay performances are fairly small among flows with similar R_i. But for LRD flows, comparatively higher vari-

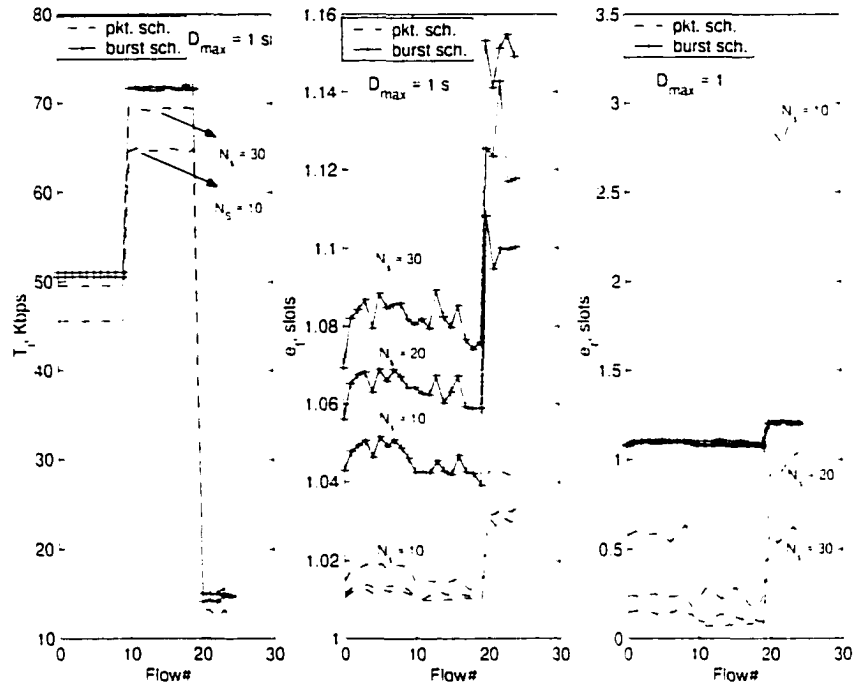

Figure 5.9. *Variation in long-term throughput and average transmitter and receiver usage times under packet-level and burst-level scheduling in a heterogeneous traffic scenario (for R_{0-9} = 64 Kbps. Poisson: R_{10-19} = 100 Kbps. GBAR: R_{20-24} = 20 Kbps. LRD: P_E = 0.1. v = 5 km/hr. f_dT = 0.00833).*

ability in loss performance is observed. The average packet loss rate is small for flows with higher D_{max} and a decrease in the value of μ improves the delay performance.

For a particular channel quality, increased channel-error correlation results in improved throughput, loss and delay performances. For burst-level scheduling, the average transmitter usage time (e_t) and the receiver usage time (e_r) increase with increasing frame-length, while for packet-level scheduling, the average transmitter and receiver usage times decrease (and hence the energy efficiency improves) with increasing frame-length (Figure 5.9). The energy efficiency increases as the channel-error correlation increases. In a heterogeneous traffic scenario, the variation in the burstiness of LRD traffic (i.e., variation in α) may not significantly affect the average transmitter and receiver usage times corresponding to the non-LRD flows (Figure 5.10).

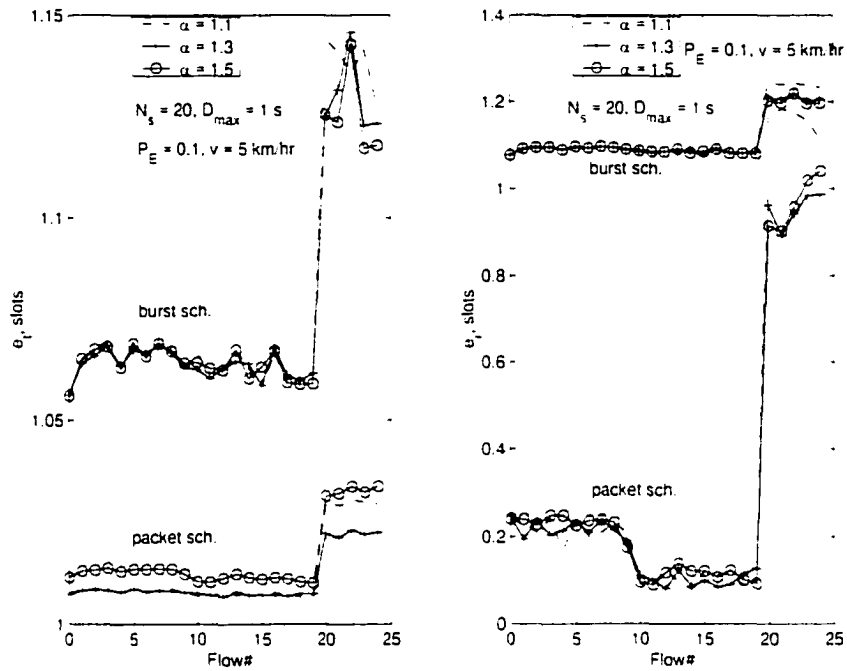

Figure 5.10. *Effect of variation in α on the average transmitter and receiver usage times under packet-level and burst-level scheduling in a heterogeneous traffic scenario (for R_{0-9} = 64 Kbps. Poisson: R_{10-19} = 100 Kbps. GBAR: R_{20-24} = 20 Kbps. LRD: I = 0.25 s. 0.25 s and 0.1 s for Poisson. GBAR and LRD flows. respectively).*

5.7 Chapter Summary

In this chapter. we have proposed a centralized TDMA-based wireless access scheme to provide throughput fairness among bursty data flows in a Rayleigh fading channel. The scheme consists of an admission control policy. a traffic conditioning mechanism and a bandwidth allocation policy. Both the packet-level and the burst-level scheduling policies for dynamic bandwidth allocation have been considered. The energy efficiency of the resulting MAC protocol in terms of the average *transmitter usage time* and the average *receiver usage time* for successful transmission of a packet have been evaluated for both these scheduling policies. Detailed performance evaluation for the proposed scheme has been carried out by computer simulations for homogeneous Poisson. GBAR and LRD flows and for heterogeneous mixture of these flows.

The following provides a brief summary of the simulation results:

i. Throughput variability among flows with the same level of subscription (as specified by R_i) is considerably small for Poisson and GBAR flows. For LRD flows, the distribution of bandwidth among flows becomes unfairly skewed as the traffic burstiness increases (i.e., the value of the Pareto shape parameter decreases). This is mainly due to the very high variability in the packet generation process which causes the traffic load generated by the different LRD flows over a time interval to vary significantly. The relative throughput fairness among flows with different profile rates can be achieved as long as the channel utilization is below some threshold and it improves for higher values of D_{max} and/or lower system load.

ii. The observed throughput fairness is robust against the number of admitted flows, their profile rates (i.e., R_i), data burst inter-arrival time, maximum tolerable delay of data bursts, TDMA frame-length and channel-error correlation. In a heterogeneous traffic scenario, channel utilization as high as 75% may be achieved while maintaining the relative throughput fairness among flows with different profile rates for error rate as large as 10%. Higher level of channel utilization can be achieved as the channel quality improves.

iii. For a certain TDMA frame-length and for fixed D_{max} and R_i, under certain channel condition, the actual throughput performances of flows depend mainly on the traffic generation pattern (as manifested in the burstiness) of the flows.

iv. Under certain channel condition, for a fixed frame-length, the *post facto* loss performance of flows depends on the corresponding value of D_{max} and the system load. An increase in the value of D_{max} results in an improved average packet loss rate (and hence higher throughput) but it increases the average packet delay. The average packet delay and the average packet loss rate can be controlled by varying the value of targetted link utilization level (μ) during admission control.

v. For a fixed frame-length, the average packet loss rate is found to be higher for bounded delay LRD flows. As the traffic burstiness of LRD flows increases, they incur more packet loss and consequently the throughputs fall. A reduction in the system traffic load (e.g., decreasing μ and/or increasing the data burst inter-arrival time) can significantly improve the throughput, loss and delay per-

formances of Poisson and GBAR flows.

vi. The achieved throughput for packet-level scheduling can be considerably lower than that for burst-level scheduling, especially when the frame-length is relatively small. The average transmitter usage time is lower for packet-level scheduling than for burst-level scheduling. The average receiver usage time is always higher with burst-level scheduling except when the frame-length is relatively small. The selection between the packet-level and the burst-level scheduling would be based on this tradeoff between channel utilization and energy efficiency.

vii. For a particular channel quality, throughput, loss and delay performances of flows and the energy efficiency of the MAC protocol improve as the channel-error correlation increases.

The results presented here apply to multi-cell systems as well. The centralized control functionalities can be put either in the IP-aware BS or further deep into the network. It's an implementation issue. Link-level rate adaptation can be combined with scheduling in the proposed framework for dynamic bandwidth allocation. For instance, by adjusting the parameter R_i, the throughput for each traffic flow can be controlled and this sort of adaptation would be required to respond to network dynamics. In a multi-cell system, in order to accommodate handoff traffic, the traffic profile of individual flow in a cell can be tightened in a fair manner so that each traffic flow experiences a fair service degradation.

Chapter 6

Channel Utilization and TCP Throughput Fairness in Wireless IP Networks

6.1 Introduction

Since TCP/IP is the standard network protocol stack on the Internet, its use over the next-generation wireless mobile networks is a certainty. Efforts in developing IP-centric wireless networks are ongoing and architectures and protocols for supporting multimedia traffic are evolving ([6], [8], [69]-[70]). The transport-layer protocol performance is one of the most critical issues in data networking over noisy wireless links. The performance of the transport protocol depends on the service provided by the underlying network and RLC/MAC layer. High utilization of wireless link bandwidth along with fairness towards competing TCP traffic are required in a wireless packet data network. Although maximizing wireless channel utilization while maximizing fairness are conflicting goals, efficient and robust transport protocol along with efficient packet scheduling mechanism over the air interface can provide throughput fairness with reasonably high channel utilization.

The work presented here complements that in [59], where the authors introduced the concept of CSDPS (Channel State Dependent Packet Scheduling) as a link-layer solution to increase channel utilization and reduce the unfairness problem. A simulation-based study on TCP performance in wide-area wireless networks is presented here. More specifically, the performances of a set of transport-

level packet scheduling policies are investigated in terms of channel utilization and throughput fairness among competing TCP connections in the wireless last hop. All these schemes exploit the CSDPS concept at the TCP level. Also, a time-frame-based channel-state aware transport-level packet scheduling scheme, namely, *Per-Destination Queueing* with *Time Frame-Based Scheduling (PD-TFS)*, is presented that imitates fair-queueing in large time scale.

Neither the production TCP code is used nor the exact behavior of specific implementations of TCP is imitated here. But the simulator used here supports the underlying TCP congestion and error control algorithms including slow-start, congestion avoidance, fast retransmit and fast recovery. A semi-reliable RLC/MAC layer protocol is assumed underneath the transport protocol. TCP performance under the *PD-TFS* scheme is compared to that under *Single-Queue FIFO (SQ-FF)*, *Per-Destination Queueing with FIFO (PD-FF)*, *Per-Destination Queueing with Round-Robin (PD-RR)* and *Per-Destination Queueing with Longest Queue First (PD-LQF)*-based scheduling policy. *PD-FF*, *PD-RR*, *PD-LQF* and *PD-TFS* schemes use destination-based queueing, that is, transmission queues at the BS for the different mobile stations are different and all the packets corresponding to a certain mobile station are transmitted from the same queue.

Since both high channel utilization and improved throughput fairness are desirable networking goals, a new performance metric, namely *Throughput-Fairness Product (TFP)* is defined for the performance assessment of different scheduling schemes.

The *PD-TFS* scheme provides bandwidth allocation that is more fair among similar connections (i.e., connections with the same round trip time and same maximum transmission window size) than *SQ-FF*, *PD-FF*, *PD-RR* and *PD-LQF* based schemes though the overall link utilization may not be optimized. It integrates the concept of channel-state dependent scheduling with a credit-based scheduling policy. This type of scheduling scheme can be useful to provide uniform service to all mobile stations in a service area in the presence of correlated channel errors and to check against some malicious receivers trying to maximize their own throughputs exploiting TCP vulnerabilities [71].

The rest of the chapter is organized as follows. In Section 6.2, the background and motivation for this work are described. In Section 6.3, some related works on wireless

TCP performance are described briefly. Section 6.4 describes the system model that is simulated. Simulation results are presented in section 6.5. The integration of TCP-level packet scheduling in the GPRS (General Packet Radio Service) transmission protocol stack is described in Section 6.6. In Section 6.7, the findings are summarized.

6.2 Background and Motivation

TCP suffers serious performance degradation due to high channel error rate and the consequent invocation of its sliding window based congestion control mechanism [72]. As a result of TCP's congestion control algorithm, a connection's throughput varies as the inverse of the connnections round trip time. Given a packet drop rate of p, the maximum sending rate for a TCP connection is T *bytes/sec*, for $T \leq \frac{1.5 * \sqrt{2/3} * B}{R * \sqrt{p}}$ where B is the packet size and R is the round trip time [73]. Wireless links are generally more error-prone than wired links and, in a typical wireless scenario, packet losses can be as high as 10%, depending on the user mobility, location and co-channel interference. Again, some of the mechanisms like *fast retransmit*, proposed for enahancing the basic TCP algorithm, may even fail in a wireless network due to the low bandwidth-delay product (and hence small TCP window size) and correlated fading (and hence loss of back-to-back packets and ACKs). Similar to link errors, when a handoff from one cell to the other occurs, there may be extra delay and packet loss. TCP recognizes loss due to link-error and/or handoff as signal of congestion and invokes the coresponding congestion avoidance algorithm. All these contribute to significant throughput degradation of TCP connections in wireless networks.

There has been a flurry of recent works on improving the TCP performance over wireless networks. While most of the works ([74]-[84]) try to shield the packet loss due to channel errors to the sender, some of them ([59]) try to exploit the channel error-correlations in the wireless link to improve TCP performance. In fact, these approaches supplement each other with a common goal: to improve reliable transport protocol performance in wireless networks.

The realized throughput for a flow in a wide area wireless network is a result of the combination of packet marking (as in Differentiated Services IP networks [85]) and dropping policy of the network routers and the policy of the transport protocol

in how it reacts to packet-dropping. Here, the focus is on the effects of different packet scheduling policies at the last hop router (e.g., the IP-aware BS) on channel utilization and realized throughput fairness at the TCP level. The unfairness is induced by link-level dynamics at the wireless last hop. The unfairness induced by the TCP mechanism (e.g., TCP bias against flows with long round trip time) is not being considered here. In the rest of the chapter, the terms 'BS' and 'IP-aware last hop router' are used interchangeably.

While the fairness of TCP congestion avoidance among competing flows can be maintained through improved traffic management techniques (e.g., [86]) and/or modified congestion avoidance algorithms (e.g., [87]) in the wired Internet, it may deteriorate in the wireless last hop due to time and location dependent channel errors. This problem can be alleviated through using a wireless fair-queueing mechanism [56]. In a wireless fair-queueing model, lagging flows are allowed to make up their lag by causing leading flows to give up their lead. Multiple TCP connections going over the same network path do not need to achieve the same throughput on a time scale comparable to the round trip time. Thus fairness needs to be measured over time intervals longer than a few round trip times.

Throughput fairness among different flows in a wireless network is dependent on channel quality and transport level packet scheduling policy at the BS. Due to lower values of bandwidth-delay product, congestion and consequent packet drop at the BS buffer is less likely to be a problem. The problem lies with the wireless channel unreliability. Channel-state dependent packet scheduling at the transport level can improve TCP throughput and hence fairness (for example, throughput fairness among TCP and UDP flows). At the same time, some type of fair-queueing mechanism can improve throughput fairness among similar connections.

The motivation of this work has been to better understand the issues involved in the design of transport level protocols for multipoint to multipoint packet data transmission over fading wireless channels, the implications of the channel-error correlations and the lower layer (e.g., RLC/MAC layer) protocol parameters on TCP performance and to gain insight into the relationships among the various system parameters which would be required for designing packet-switched services in the emerging IP-centric wireless mobile networks. The effects of the TCP window size

(W). TD (Transmission Delay) in the wired-network, number of concurrent TCP connections (N), channel error-correlation, fast retransmit threshold (K) on total TCP throughput (and hence channel utilization) and throughput fairness are investigated for different channel state dependent packet scheduling schemes at the transport layer. Rather than counting on a specific implementation, the TCP congestion and flow control mechanisms, including slow-start, congestion avoidance, fast retransmit and fast recovery, are simulated. The *TFP* metric is evaluated for the proposed *PD-TFS* scheme and is compared to those of the *SQ-FF*, *PD-FF*, *PD-RR* and *PD-LQF* schemes. Relevant works exploring the effects of TCP parameters and channel impairments on TCP throughput fairness in wireless networks are yet to appear in literature.

Simulation results show that with properly chosen transport-layer scheduling policy and transport protocol paremeters (e.g., TCP window size, fast retransmit threshold), a reasonably high channel utilization and fair throughput performance can be achieved even when the TCP packet loss rate is as high as 10%. The presence of some intermediate 'agent' (e.g., 'snoop agent' in [77]) in the BS or use of end-to-end mechanism like ELN (Explicit Loss Notification) [78] may enhance the TCP performance further.

6.3 Related Works

'Indirect-TCP' is a 'split-connection' solution that shields the original TCP sender from the wireless link [75]. Here, each TCP packet has to go through the TCP protocol processing twice at the BS as compared to zero time in a non split-connection approach. In addition, the end-to-end semantics of TCP ACKs is violated.

The 'snoop protocol' introduces a 'TCP snoop agent' at the BS that maintains a cache of TCP segments sent across the wireless link that have not yet been acknowledged [77]. When a TCP packet loss on wireless link is detected by the 'snoop agent', it retransmits the lost TCP packet and suppresses the duplicate ACKs, thereby avoiding unnecessary retransmissions and congestion control invocations by the sender. In this case, the states of all TCP connections are to be maintained at the BS.

With SACK TCP, each ACK contains information of about up to three non-

contiguous blocks of data that have been received successfully by the receiver [76]. The ACKs contain complete information of which packets have been received and which have not. This complete information allows the TCP sender to retransmit packets selectively and avoid unnecessary retransmissions, but the protocol overhead is increased in proportion to the number of missing segments detected at the receiver.

With ELN (Explicit Loss Notification) [78], when a packet is lost due to channel errors, future cumulative ACKs corresponding to the lost packet are marked to notify the sender of the non-congestion related loss so that, upon receiving this information, the sender can perform retransmission without invoking the normal congestion control function.

The impact of different scheduling mechanisms on the realized throughputs of different TCP connections under *Channel-State Dependent Packet Scheduling* was investigated in [59]. The authors proposed a class of link-layer packet scheduling policies, namely, CSDP-FIFO, CSDP-LQF, CSDP-ETF and CSDP-RR, which can be deployed at the BS to increase the utilization of the wireless channel. All of these scheduling policies are based on a per-destination queueing of the TCP packets at the BS. For FTP sessions running on top of production TCP implementation, CSDP-RR scheduler was found to outperform all other policies both in terms of performance and implementation complexity. Under the Poisson traffic and random error model, CSDP-LQF scheme was shown to provide maximum channel utilization. FIFO scheduling with a common queue for the TCP packets at the BS was found to provide poor wireless channel utilization and unfair bandwidth allocation among the competing connections. The relative difference in transmission completion time of different FTP sessions was taken as the measure of fairness. However, the paper did not deal with the impacts of different TCP and link-layer parameters and network characteristics (e.g., delay-bandwidth product, channel error-correlation pattern) on channel utilization and TCP throughput fairness.

A link-level approach, namely TULIP (Transport Unaware Link Improvement Protocol), was proposed to alleviate TCP's performance degradation over wireless links in [84] where the underlying MAC protocol is contention-based. TULIP approach was shown to provide improved throughput, delay and delay variance performance over 'snoop protocol', due to its 'MAC-level acceleration' feature but at the expense

of significant modification in the protocol stack.

6.4 System Model and Assumptions

6.4.1 TCP Model

During connection set-up, the receiver advertises a maximum window size W_{max} so that the transmitter is not allowed to have more than W_{max} unacknowledged data packets outstanding at any given time. Senders always have data to send and will send as many packets as their windows allow. Here, it is assumed that the bulk data connections send 576-*byte* TCP packets and the start times for the bulk data connections are staggered over a 10 *ms* interval. This type of connections allow us to study the steady-state behavior of the protocol.

The basic window adaptation procedure is similar to that of all currently available TCP implementations. Let $W(t)$ be the transmitters congestion window width and $W_{th}(t)$ be the slow-start threshold at time t. For a new connection, $W(t)$ is initialized to 1. The evolution of $W(t)$ and $W_{th}(t)$ are triggered by ACKs and timeouts as follows:

i. if $W(t) < W_{th}(t)$, each ACK causes $W(t)$ to be incremented by 1. This is the *slow-start* phase.

ii. if $W(t) \geq W_{th}(t)$, each ACK causes $W(t)$ to be incremented by $\frac{1}{W(t)}$. This is the *congestion avoidance* phase.

In the case of a timeout, the TCP source updates $W(t)$ and $W_{th}(t)$ as follows: $W_{th}(t+) = \lceil \frac{W(t)}{2} \rceil$ and $W(t+) = W_{th}(t+)$ and then starts retransmitting from the first lost packet. After a timeout the sender also updates the RTO (Retransmission Time Out) value using exponential backoff. The backoff continues until an ACK is received for a packet transmitted exactly once. The TCP sender does not decrease the slow start threshold (W_{th}) if it has been decreased in the last round trip time.

The TCP sinks can accept packets out of order but deliver them only in sequence to the user. The TCP sinks generate immediate ACKs. If a sink receives a packet out of order, it issues an ACK immediately for the last in order segment received.

If a packet is lost (after a stream of correctly received packets), the transmitter

keeps receiving ACKs (called duplicate ACKs) with the sequence number of the first lost packet, even if packets transmitted after the lost packet succeed in being correctly received at the receiver.

A fast retransmit procedure is implemented after a packet loss (as in TCP-Tahoe and TCP-Reno), i.e., if subsequent to a packet loss, the transmitter receives the K^{th} duplicate ACK before the timer expires, then the transmitter updates $W(t)$ and $W_{th}(t)$ as follows: $W_{th}(t+) = \lceil \frac{W(t)}{2} \rceil$ and $W(t+) = W_{th}(t+) + K$ and then retransmits the packet after the last ACKed packet. It can transmit more packets if the current value of $W(t)$ allows it to do so. For each additional ACK received, the congestion window $W(t+)$ is updated according to the basic algorithm. Each lost packet is retransmitted at most once during a single round trip time.

In the case of multiple packet loss in the loss window, fast retransmission may not be triggered due to a lack of duplicate ACKs, in which case, the sender will have to wait for a timeout to occur for loss recovery.

The round trip time (RTT) is estimated using exponential weighted averaging. The RTT is measured for each successfully received unretransmitted packet. Based on these measurements, a moving estimate of the round trip time (RTT_{mean}) is made according to the following: $RTT^{t-1} = c \times RTT_{measured} + (1 - c) \times RTT^{t}$. The value of c is assumed to be 0.1 here.

Simulation results in a sender-based one way traffic scenario are presented. It is assumed that ACKs are not lost in the wired part of the network although they may be lost in the wireless link. ACK packets undergo no extra queueing delays in addition to the propagation delay. All of the TCP connections in a simulation scenario are assumed to have the same TD value.

6.4.2 Link-Layer Model

A TCP data unit is typically equivalent to several link-layer frames. For example, in IS-99, the CDMA circuit mode data transmission standards, the link-layer frame size is 24 *bytes* (with 19 *bytes* of payload) and the TCP segment size is 536 *bytes*.

The link-layer model considered here is as follows. It is assumed that immediately after a successful frame transmission, the mobile station generates a link-level ACK which arrives at the BS immediately before the next timeslot. If the BS can decode

it successfully, it transmits the next frame; otherwise it retransmits the lost frame immediately in the next timeslot. In the event of link-layer failure, after a maximum number of n_{max} transmissions, the TCP packet is dropped and control is passed to the TCP for end-to-end error recovery. The BS scheduler then starts transmitting the HOL packet from the next candidate queue. The maximum number of link-layer retransmissions allowed for each frame is to be chosen properly so that the link-layer retransmissions do not interfere with the end-to-end retransmissions.

6.4.3 Channel Model

A system is assumed in which transmissions from the different mobile stations are power-controlled in such a way that their slowly varying path and shadow losses are perfectly compensated whereas the rapidly varying multipath fading remains uncompensated. At a high data rate, for the typical values of probability of channel outage, packet errors are correlated. The correlation in the multipath Rayleigh fading process can be modeled using a first-order Markov model [19]. The same correlated error model for Rayleigh fading channels is assumed as in [19]. It is assumed that during the course of data transfer, channels between the BS and the mobile stations suffer burst errors independently.

Under this error model, for example, packet error probability $P_E = 0.01$, 0.05 and 0.1 correspond to channel fading margin $F = 19.978\ dB$, $12.899\ dB$ and $9.773\ dB$, respectively. With carrier frequency of 900 MHz and packet duration of 2.304 ms (which corresponds to the transmission time of a packet of size 576 $bytes$ in a 2 $Mbps$ channel), normalized Doppler bandwidth $f_dT = 0.0096$, 0.0192, 0.0384 and 0.11520, when mobile terminal speed is 5 km/hr, 10 km/hr, 20 km/hr and 60 km/hr, respectively. For $P_E = 0.05$, with these values of f_dT, the average length of error bursts (in TCP packets) are 9.68686, 4.87280, 2.50669 and 1.22081, respectively. For $P_E = 0.1$, the corresponding error burst lengths are 13.84552, 7.14441, 3.61690 and 1.46006 packets, respectively.

6.4.4 Channel Probing Model

It is assumed that the channel-state is monitored by small 'probe' messages, sent from the mobile stations, which are sufficiently protected through error-correcting codes so that they can always be decoded correctly at the BS. When a mobile station is receiving link-layer frames continuously from the BS, the link-layer ACKs sent by the mobile stations can be used to predict the channel-state; otherwise, the mobile station sends probe messages periodically so that the BS can predict the corresponding channel-status. Here, it is assumed that the probing interval is one TCP packet transmission time.

6.4.5 Scheduling Schemes

In all the schemes (except *SQ-FF*), TCP packets are buffered in separate queues corresponding to different destination mobile stations. The different scheduling schemes considered here are as follows:

i. *SQ-FF (Single Queue FIFO)*: all TCP packets are buffered in a single queue and are scheduled for transmission on a FIFO basis.

ii. *PD-FF (Per-Destination Queueing with FIFO)*: a timestamp value (corresponding to the time of arrival of the earliest packet) is maintained for each queue. A packet corresponding to the queue for which the channel is expected to be good and for which the timestamp value is minimum is scheduled for transmission.

iii. *PD-RR (Per-Destination Queueing with Round-Robin)*: packets from the queues are scheduled for transmission in a round-robin fashion. A packet is dispatched to RLC/MAC layer for transmission only if the corresponding channel condition is expected to be good.

iv. *PD-LQF (Per-Destination Queueing with Longest Queue First)*: the scheduler schedules packets from the different queues based on the length of the queues. A packet is dispatched to RLC/MAC layer for transmission from a queue if the corresponding queue-length is the highest and if the channel condition is expected to be good for the corresponding destination.

v. *PD-TFS (Per-Destination Queueing with Time-Frame-Based Scheduling)*: a time-frame (in number of TCP packets) is constructed based on the number

of sink mobile stations and their fair share of the link bandwidth. A credit parameter is maintained for each mobile the value of which is increased during each time-frame by an amount determined by the fair share of the mobile and the available link bandwidth. After a TCP packet is successfully transmitted by the BS to a mobile station, the corresponding credit value is decremented by one. For each queue a timestamp value is also maintained.

Let d_i, $d_i(q_i)$, $d_i(credit)$ and R_i denote the destination i, corresponding queue-length, credit parameter and fair share of the link bandwidth, respectively. $P(t)$ denotes the set of queues for which $d_i(q_i) > 0$ and channel-states are expected to be good. In $P(t)$, the queues are logically arranged in the descending order of the $d_i(credit)$ values and, among d_i with the same $d_i(credit)$ value, the one with the highest $d_i(q_i)$ value precedes the others. Again, if both the $d_i(credit)$ and the $d_i(q_i)$ values are the same, the one with the lower timestamp value has priority over the others. Let N_s denote the time frame size in TCP packets and R denote the channel data rate.

This transport-layer packet scheduling policy can then be described as follows:

Procedure Schedule() {

/* initialization done after admission of flows corresponding to d_i */

$a_i = max(1, \lfloor N_s * \frac{R_i}{R} \rfloor)$, $d_i(credit) = a_i$

1. Schedule HOL packet from the first queue in $P(t)$ for transmission.
2. Wait until the packet is successfully transmitted or
   transmission is aborted.
3. If (packet is successfully transmitted)
   decrement $d_i(credit)$
4. Update $P(t)$
   Go to step 1.

}

The BS scheduler updates the value of $d_i(credit)$ for each d_i during each time-window as follows:

$$d_i(credit) = d_i(credit) + a_i$$

6.4.6 Performance Metrics

- *TCP Throughput (T_i)*: It is the long-run average bulk throughput of the protocol and is measured at the destination TCP layer. Here, it is defined as the total number of bytes delivered to the destination application (excluding retransmission and losses) divided by the total connection time.
- *Channel Utilization (U)*: It is defined as $U = \frac{\sum_{i=1}^{N} T_i}{R}$, where R is the total achievable data rate in the wireless channel.
- *Fairness Index F*: It is a function of the variability of the throughput across the TCP connections and has been defined as [87]: $F(x_1, \ldots, x_n) = \frac{(\sum_{i=1}^{N} x_i)^2}{N \cdot \sum_{i=1}^{N} (x_i)^2}$, where x_i = observed throughput of the ith TCP connection ($0 < i \leq N$), normalized to the total achievable throughput in the link, and N = total number of flows sharing the link. But for a wireless network with low bandwidth-delay product this measure is rather 'coarse' (since the value of this metric does not change appreciably for the different scheduling schemes as we observe from simulations) and we adopt a 'finer' measure F which is based on the difference between the maximum and minimum achieved throughput: $F = \frac{1}{(x_{max} - x_{min})^b}$, where $b \in I$.
- *TFP*: It measures the performance tradeoff between the achieved total throughput and the throughput fairness for a particular multipoint to multipoint packet scheduling policy and is defined as follows: $TFP = \left(\sum_{i=1}^{N} T_i\right)^a \times F$. *TFP* is maximized when the product of the total throughput and the fairness is maximized. In our simulation we assume $a = 1$ and $b = 1$. The concept of *TFP* here is similar to that of *Power* in [88].

6.4.7 Simulation Parameters and Simulation Environment

Figure 6.1 shows the topology of our simulation. The buffer size at the BS is assumed to be infinite. The different simulation parameters are packet size, ACK packet size, trunk delay (*TD*), number of concurrent TCP connections (N_c), maximum size of the sending window (W), bandwidth allocated for each flow, RT (Retransmission Timer) backoff parameter (Q), fast retransmit parameter (K), maximum number of transmissions allowed for each RLC/MAC-layer frame (n_{max}) and terminal speed (v).

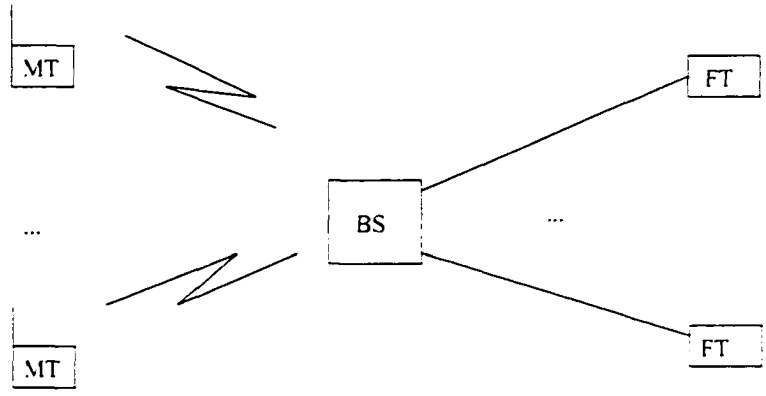

Figure 6.1. *Simulation topology (FT ≡ fixed terminal, MT ≡ mobile terminal, BS ≡ base station).*

The *TD* parameter accounts for the delay experienced by the TCP packets in the wired part of the network and hence affects the *RTT* (Round Trip Time) value of the corresponding TCP connection. If t_{TCP} is the time required to transmit a TCP packet over the wireless link, then $RTT \geq 2 \times TD + t_{TCP}$, if we ignore the transmission delay of the corresponding ACK packet over the wireless channel.

For the *PD-TFS* scheme, we assume that all the receiving mobiles have an equal share of the available channel bandwidth and that there is one TCP connection for each mobile station at a time. The size of the time-frame (in TCP packets) is chosen to be equal to the number of simultaneous TCP connections.

A *C*-coded event-driven simulator is used. The values of some of the simulation parameters used for deriving the simulation results presented in this paper are shown in Table 6.1. Each of the simulations is run until 5 *million* TCP packets are successfully transmitted. To avoid the transient effects, the first 100000 packets are not considered while calculating the performance measures. In the simulator, timeout events can be simulated with very fine precision which may not be possible in an actual implementation. Simulation results are obtained for the different values of W, K, n_{max}, TD and N_c under different channel conditions (i.e., for the different values of P_E and f_dT).

Table 6.1. *Simulation parameters.*

Parameter	Value	Parameter	Value
Channel rate, R	2 *Mbps*	TCP packet size, L	576 *bytes*
Trunk speed, B	2 *Mbps*	Link-layer frame size	24 *bytes*
$RTO_{initial}$	2.5 * TD	RTO_{max}	64 * $RTO_{initial}$
Q	2.00	f_c	900 *MHz*
ACK packet size	24 *bytes*	Simulation length	5 *million* TCP packets

6.5 Simulation Results and Discussions

6.5.1 Effects of Variation in W

TCP connections experiencing the same network conditions but with different values of W have different throughputs. Typical variations in U and $1/F$ with W are shown in Figures 6.2 and 6.3 for some typical values of the simulation parameters. Some simulation results are also enumerated in Table 6.2.

In a particular channel condition, the optimum value of W depends largely on the value of TD. For a particular value of TD, if W is very small, then channel utilization will be low since the network will be underloaded. Again, if W is very large (or TD is small compared to $W \times TD$), frequent timeout may occur since network load in this case will be higher compared to the network capacity. As channel quality deteriorates, the situation becomes even worse; therefore, depending on the RTT value and channel error characteristics, W can be properly chosen to have high channel utilization. Hence *TCP may be enhanced to dynamically adjust the maximum TCP window size, depending on the network delay and error condition, retaining its basic congestion control mechanism.*

Under similar network conditions, with *SQ-FF* and *PD-LQF* based scheduling, the observed throughput fairness among the similar TCP connections is poorer compared to that for the other scheduling schemes considered here. For lower TD values, the fairness index F for *SQ-FF* is observed to be higher than for *PD-LQF*, but as TD increases the situation reverses. The throughput fairness is considerably better with *PD-FF* and *PD-RR* based scheduling. In terms of throughput fairness, these two

Table 6.2. *Variation in U and 1/F for different values of W under different scheduling schemes (for K = 3, N_c = 4, n_{max} = 20, P_E = 0.1, v = 5 km/hr).*

W	SQ-FF		PD-FF		PD-RR		PD-LQF		PD-TFS	
	U	1/F	U	1/F	U	1/F	U	1/F	U	1/F
	TD = 5 ms									
3	0.860	55.622	0.850	21.878	0.854	23.273	0.814	214.094	0.839	1.996
4	0.838	61.379	0.833	22.085	0.823	24.291	0.787	207.550	0.801	2.420
5	0.823	66.478	0.806	21.122	0.802	23.867	0.734	197.090	0.778	3.700
10	0.711	81.531	0.696	18.426	0.673	20.521	0.646	269.154	0.664	3.987
20	0.701	85.589	0.687	19.591	0.664	20.460	0.639	272.791	0.662	4.010
	TD = 15 ms									
3	0.594	37.119	0.670	12.619	0.532	11.877	0.674	25.954	0.676	1.712
4	0.722	52.375	0.776	16.483	0.743	17.886	0.761	51.056	0.773	0.700
5	0.764	65.180	0.794	18.727	0.772	20.794	0.730	64.882	0.762	2.180
10	0.701	94.687	0.698	20.836	0.673	20.813	0.640	84.566	0.665	4.423
20	0.690	101.512	0.690	19.269	0.665	21.254	0.637	84.992	0.661	3.589

schemes offer comparable performance. With the *PD-TFS* scheme, as expected, *F* is significantly higher than in all other schemes.

Simulation results show that, with *SQ-FF* based scheduling, for a particular value of *TD*, throughput fairness deteriorates as W increases. But for all other schemes, the variation in *F* with W may not be monotonic and is dependent on the values of *TD*. It is observed that the throughput fairness decreases monotonically with increasing values of W when *TD* is relatively larger.

Interestingly, for connections with relatively small W in relation to *TD* (e.g., W = 5, *TD* = 15 *ms*), with *PD-TFS* based scheduling, channel utilization is considerably high while it drops off considerably in the case of *SQ-FF* and *PD-RR* based scheduling. On the other hand, for connections with small *window-delay product* (e.g., W = 3, *TD* = 5 *ms*), channel utilization is found to be greater for *SQ-FF* scheme compared to the other schemes.

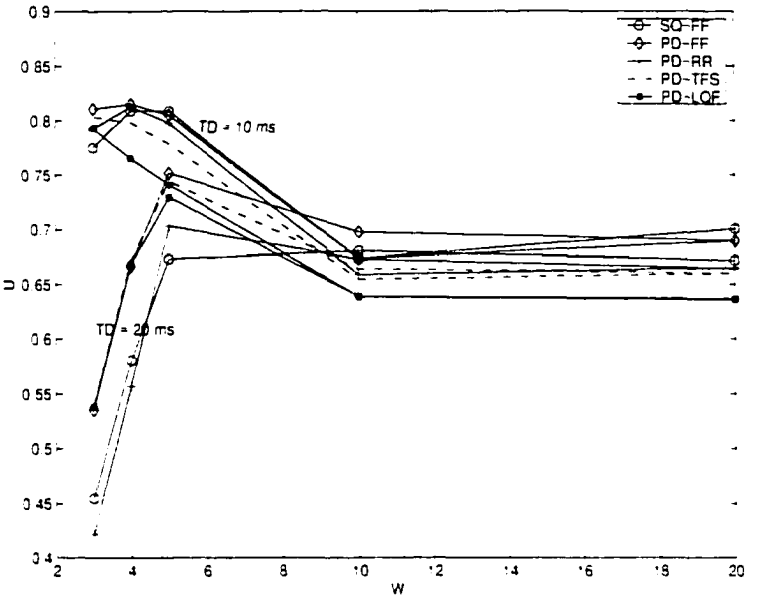

Figure 6.2. *Variation in channel utilization for different values of W under different scheduling schemes (for K = 3, n_{max} = 20, N_c = 4, P_E = 0.1, v = 5 km/hr).*

6.5.2 Effects of Variation in *TD*

Throughput of a TCP connection depends largely on the corresponding RTT value. In a wireless IP network, the RTT value for a TCP connection is determined by *TD* and the time required to transmit a TCP packet over the wireless link.

Simulation results show that, when packet error rate is moderately high (e.g., 10%), as *TD* is increased, for *SQ-FF*-based scheduling, channel utilization drops off more rapidly compared to other schemes (Figure 6.4). For a particular W, as *TD* increases, throughput fainess improves for *PD-LQF* based scheduling (Figure 6.5); but for the other scheduling schemes, the variation in fairness index with *TD* may not be necessarily monotonic and is dependent on other parameters such as W, N and f_dT. It is observed that, among all the schemes, *PD-TFS* scheme offers the maximum throughput fairness under varying values of *TD*.

Some typical simulation results are enumerated in Table 6.3.

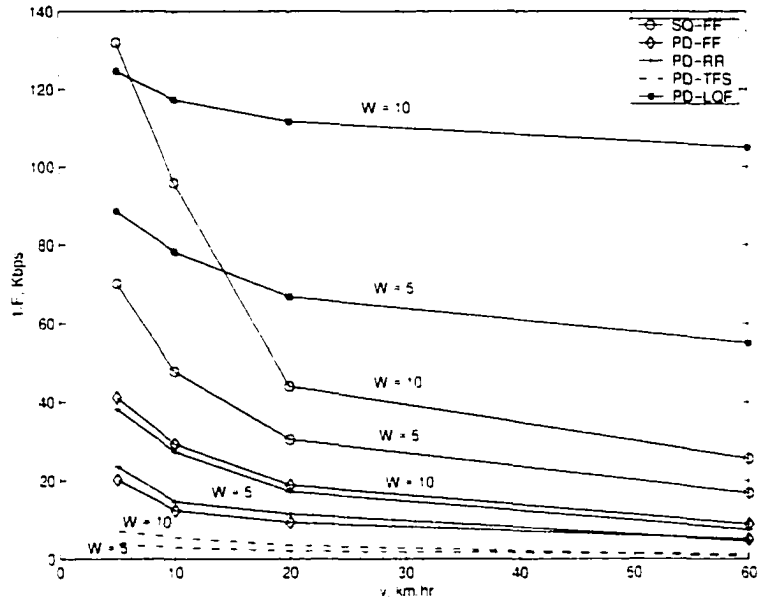

Figure 6.3. *Variation in throughput fairness for different values of W under different scheduling schemes (for K = 3, n_{max} = 20, N_c = 4, TD = 10 ms, P_E = 0.1).*

6.5.3 Effects of Variation in N_c

The variation in channel utilization U with the number of active connections N_c depends largely on the values of W and TD (and hence on window-delay product). For TCP connections with large window-delay product where W is small reltive to TD, when N is small, with *SQ-FF* scheduling, channel utilization is low compared to all other schemes (Figure 6.6). Simulation results for the different values of TD and W show that the variation in U with N_c is very small with *PD-TFS*-based scheduling, compared to other schemes.

For all the scheduling schemes except *PD-LQF*, throughput fairness improves monotonically with increasing N_c (Figure 6.7). The throughput fairness F is the worst for *PD-LQF* scheduling policy. The *SQ-FF* scheme only performs better than *PQ-LQF* and as expected, among all the scheduling schemes, the *PD-TFS* scheme offers the maximum throughput fairness under a varying number of concurrent TCP connections.

Some typical simulation results are shown in Table 6.4.

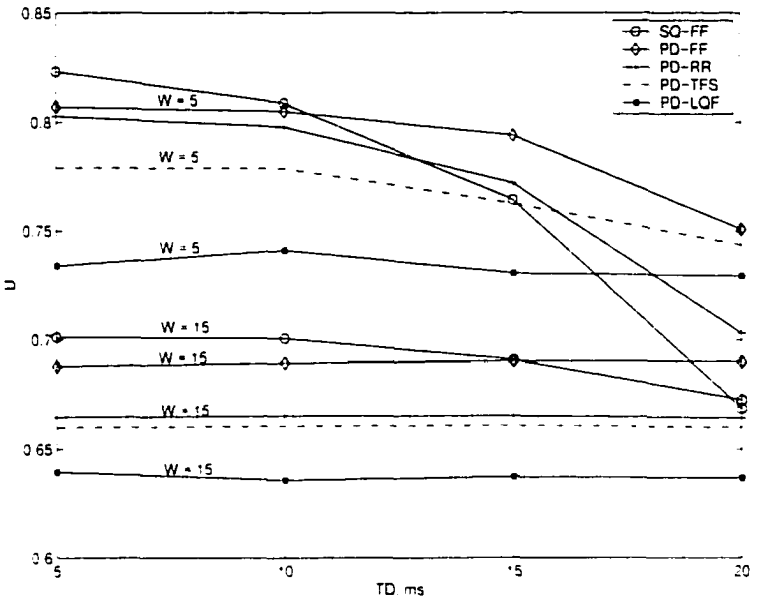

Figure 6.4. *Variation in channel utilization for different values of TD under different scheduling schemes (for K = 3, n_{max} = 20, N_c = 4, P_E = 0.1, v = 5 km/hr).*

6.5.4 Effects of Variation in f_dT

Simulation results show that, for a particular channel quality, as f_dT decreases (i.e., error-correlation increases), independent of the values of the system parameters (i.e., W, K, TD, N_c, n_{max}), the channel utilization improves for all of the scheduling policies considered here. This finding is supportive of some of the earlier results (e.g., in [80]) on the effects of channel error-correlation on TCP throughput. For high packet error rate, better performance is achieved for correlated errors than for *i.i.d.* errors because, for a given value of P_E, correlated errors correspond to fewer error events (each comprising multiple packet errors) and therefore the congestion window is shrunk less frequently than for *i.i.d.* errors.

The throughput fairness improves as f_dT increases (i.e., channel errors become more independent). Consequently, TFP increases as channel errors become more uncorrelated. It is observed that, this trend is independent of the values of other system parameters except in some cases when n_{max} is relatively smaller (e.g., n_{max} = 2) and channel quality is relatively better (e.g., P_E = 0.05). In the latter cases, throughput fairness may not vary monotonically with f_dT. Among all of the scheduling schemes,

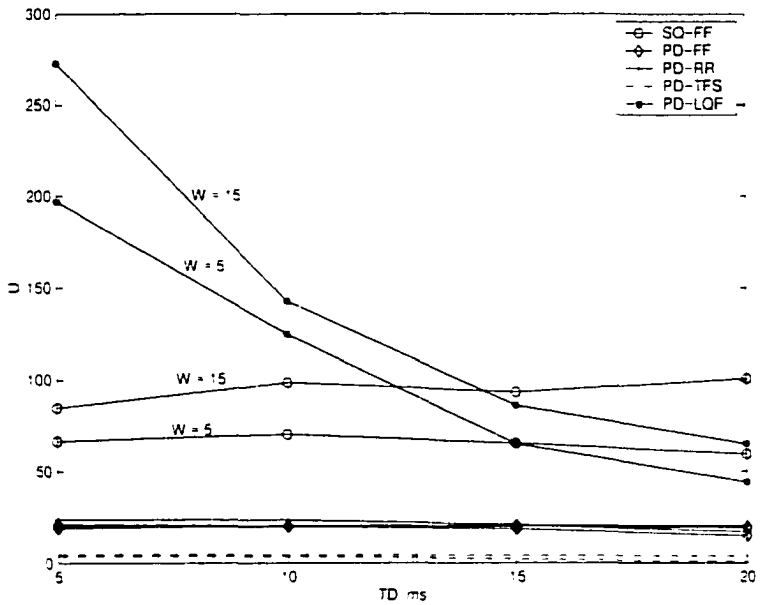

Figure 6.5. *Variation in throughput fairness for different values of TD under different scheduling schemes (for $K = 3$, $n_{max} = 20$, $N_c = 4$, $P_E = 0.1$, $v = 5\ km/hr$).*

the *PD-TFS* scheme offers the best *TFP* (Figure 6.8).

Some typical simulation results are enumerated in Table 6.5.

6.5.5 Effects of Variation in n_{max}

The value of n_{max}, the maximum allowable number of transmissions of a RLC/MAC-layer frame, can significantly affect the total achieved throughput and hence channel utilization, depending on the channel-error correlation pattern. Simulation results show that, for all the scheduling schemes, as the channel errors become more independent, channel utilization drops off rapidly for relatively lower values of n_{max} (e.g., $n_{max} = 2$) (Figure 6.9). As n_{max} is increased, channel utilization improves for the different scheduling schemes when channel errors become less correlated. This is reasonable since, as channel errors become more uncorrelated, while in bad state the channel may turn to good state soon and hence persisting on retransmitting garbled frames is expected to improve RLC/MAC-layer throughput. But, in the case of highly correlated errors, when the channel is in bad state, it is more probable that it will be in the same state in the very near future and hence repeated immediate

Table 6.3. *Variation in U and 1/F for different values of TD under different scheduling schemes (for K = 3, n_{max} = 20, P_E = 0.1, v = 5 km/hr).*

TD	SQ-FF		PD-FF		PD-RR		PD-LQF		PD-TFS	
(ms)	U	1/F	U	1/F	U	1/F	U	1/F	U	1/F
	W=10, N_r=4									
5	0.711	81.531	0.696	18.426	0.673	20.521	0.646	269.154	0.665	6.721
10	0.712	99.174	0.698	19.222	0.673	21.728	0.639	141.927	0.664	3.987
15	0.701	94.687	0.698	20.836	0.673	20.813	0.640	84.566	0.665	4.423
20	0.680	101.127	0.697	18.787	0.673	21.728	0.638	64.658	0.664	3.918
	W=20, N_r=2									
5	0.666	112.368	0.668	41.707	0.656	38.255	0.644	141.743	0.655	6.721
10	0.637	111.172	0.657	38.994	0.645	36.580	0.633	88.144	0.645	5.547
15	0.586	103.939	0.632	30.863	0.617	31.958	0.614	69.388	0.628	2.444
20	0.522	98.613	0.598	27.959	0.571	28.443	0.587	69.688	0.598	3.136

retransmissions of frames may not be useful. Similar results are observed for different values of W and *TD*.

For the different scheduling schemes the throughput fairness does not change significantly with variations in n_{max}. It is observed that when channel quality is relatively better (e.g., P_E = 0.05), for lower values of n_{max} (e.g., n_{max} = 2), the TCP throughput fairness may not vary monotonically with f_dT and improvement in throughput fairness with increasing values of f_dT becomes more evident with increasing n_{max}.

6.5.6 Effects of Variation in K

Simulation results for the five different scheduling polices are obtained for different values of *TD* and W under three different values of K (K = 2, 3 and 5). From the results it is observed that, other parameters remaining the same, for a particular channel quality, the channel utilization and throughput fairness do not vary significantly for the different values of K. The reason is that, since in this TCP model the value of

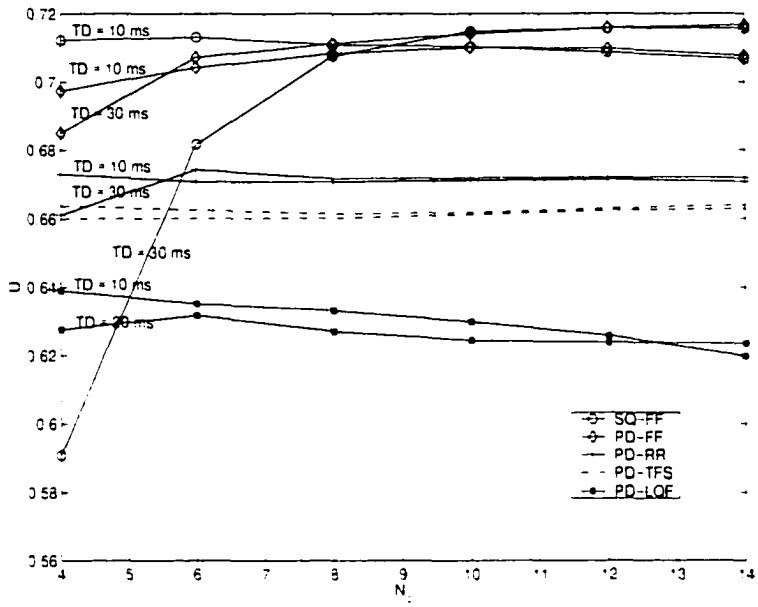

Figure 6.6. *Variation in channel utilization for different values of N_c under different scheduling schemes (for $K = 3$, $n_{max} = 20$, $W = 10$, $P_E = 0.1$, $v = 5$ km/hr).*

W is increased (by 1 until $W(t)$ reaches W_{th} and then by $1/W(t)$) for each duplicate ACK and after receiving the K^{th} duplicate ACK $W(t+)$ is incremented by K, it is more likely that, for a certain value of the maximum window size W, the average window size remains larger for higher values of K. But with a low value of the fast retransmit threshold K, the triggering of the fast retransmission procedure is quicker (although the average transmission window size is likely to be smaller) compared to the case when K is relatively larger. These effects result in a similar 'packet pumping rate' for the senders and hence in a similar throughput performance for the two cases when the other system parameters remain the same.

Some typical simulation results are enumerated in Table 6.6.

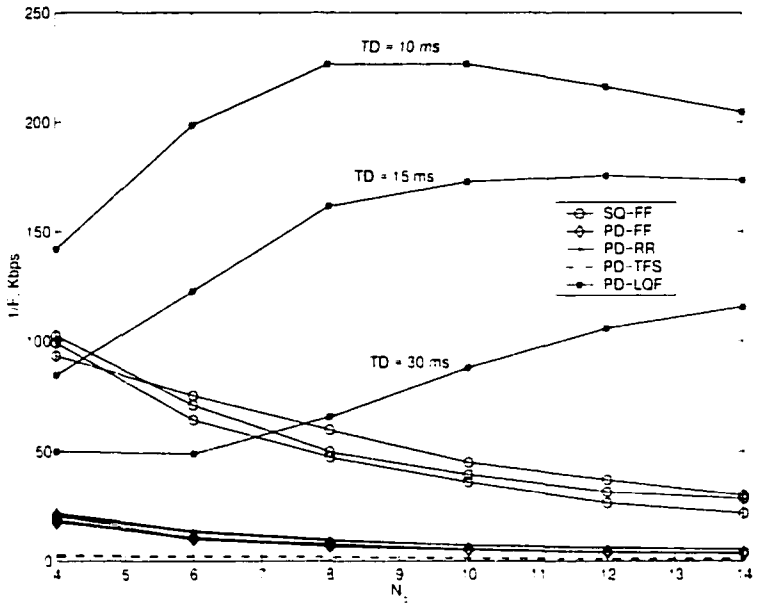

Figure 6.7. *Variation in throughput fairness for different values of N_c (for $K = 3$, $n_{max} = 20$, $W = 10$, $P_E = 0.1$, $v = 5$ km/hr).*

6.6 Integration of TCP Level Packet Scheduling into the GPRS Transmission Protocol Stack

The proposed time-frame-based TCP packet scheduling scheme can be easily integrated into the GPRS protocol stack (Figure 6.10). After receiving a TCP packet from GGSN (Gateway GPRS Support Node), the SNDCP (Subnetwork Dependent Convergence Protocol) module in the SGSN (Serving GPRS Support Node) can optionally perform header compression before passing it to the LLC layer where the packet is segmented into frames for transmission over the wireless link. The packet dispatching can be controlled by a scheduler residing in the SNDCP layer. In the case of frame transmission failure in the LLC layer after some maximum number of transmissions, the scheduler in the SNDCP layer can be notified so that it can dispatch the TCP packet from the next candidate queue.

Table 6.4. *Variation in U and 1/F for different values of N_c under different scheduling schemes (for K = 3, n_{max} = 20, W = 10, TD = 10 ms, P_E = 0.05, v = 5 km/hr).*

	SQ-FF		PD-FF		PD-RR		PD-LQF		PD-TFS	
N_c	U	1/F	U	1/F	U	1/F	U	1/F	U	1/F
4	0.725	28.252	0.716	6.293	0.686	6.219	0.649	142.856	0.675	1.415
6	0.726	20.524	0.722	4.924	0.679	4.672	0.642	232.150	0.670	0.831
8	0.729	13.730	0.726	3.193	0.678	3.070	0.640	252.561	0.671	0.677
10	0.729	10.853	0.727	2.363	0.678	2.714	0.636	244.887	0.672	0.541
12	0.728	9.250	0.727	1.955	0.679	2.129	0.632	231.694	0.673	0.371
14	0.727	7.103	0.726	1.475	0.678	1.644	0.626	214.145	0.674	0.419

6.7 Chapter Summary

The performance evaluation of a number of TCP-level packet scheduling policies is carried out in terms of channel utilization and throughput fairness (among similar competing TCP connections) for multipoint to multipoint data transmission over independently fading wireless channels in a wide area wireless IP network. A semi-reliable RLC/MAC layer is assumed underneath the transport layer. A time-frame-based channel-state aware transport-level packet scheduling scheme has been proposed that imitates wireless fair-queueing in long time scale. A new metric, namely, *Throuput-Fairness Product (TFP)* has been introduced as a performance measure of the different packet scheduling policies. The impacts of different protocol parameters (e.g., TCP parameters) and network characteristics (e.g., wired-IP network delay, error/loss rate, error-correlation pattern) on wireless channel utilization and TCP throughput fairness for the different scheduling schemes have been investigated by simulating the system dynamics under multiple concurrent TCP connections.

The main contribution here is the detailed simulation study of five different transport-level packet scheduling schemes to assess their performance in terms of channel utilization and TCP throughput fairness. The system dynamics is fairly com-

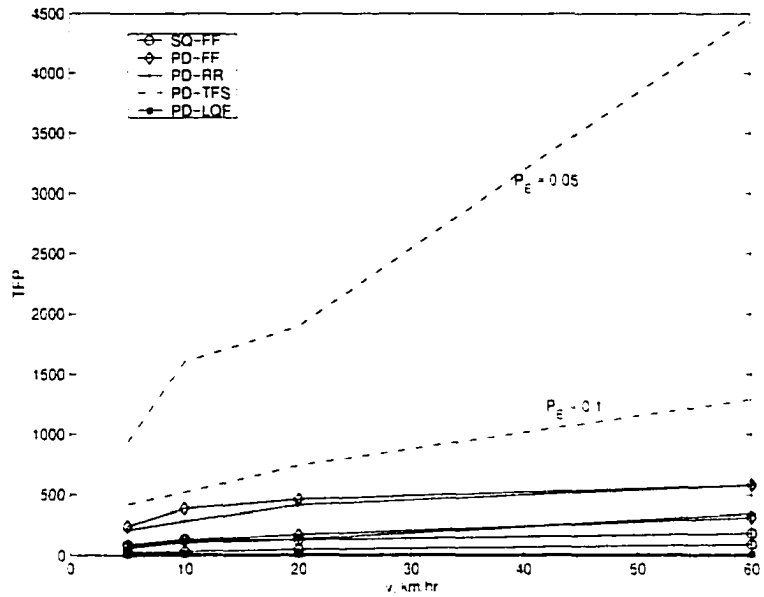

Figure 6.8. *Variation in* TFP *for the different scheduling policies (for* $K = 3$, $n_{max} = 20$, $TD = 10$ *ms*, $W = 5$, $N_r = 4$*).*

plex, involving many parameters and their interdependencies. Some general trends have been identified. It has been observed that in an error-prone wireless network, with an underlying semi-reliable link layer, a reasonably high TCP throughput (e.g., 75%) and throughput fairness among similar wide area TCP connections can be achieved when the protocol parameters (e.g., maximum TCP window size, maximum number of allowable retransmissions at the link layer) and the scheduling scheme are properly chosen. Therefore, one important issue is to develop a mechanism for dynamic adaptation of the maximum TCP window size depending on the network delay and error characteristics. This will enforce a two-level control in the TCP congestion control technique–control of the maximum allowable window size W and control of the instantaneous value of the window size $W(t)$.

Simulation results show that, although in some scenarios the PD-TFS scheme may not provide the maximum throughput performance among all the schemes described here, it provides the best throughput fairness and provides reasonably high channel utilization under a varying number of concurrent connections over a large range of values of the corresponding round trip times even when the TCP packet error rate is

Table 6.5. *Variation in U and 1/F for different values of f_dT under different scheduling schemes (for K = 3, n_{max} = 20, W = 10, TD = 10 ms, N_c = 4).*

	SQ-FF		PD-FF		PD-RR		PD-LQF		PD-TFS	
f_dT	U	1/F	U	1/F	U	1/F	U	1/F	U	1/F
0.0096	0.712	99.174	0.697	18.787	0.673	21.728	0.639	141.927	0.664	3.987
0.0192	0.706	70.181	0.692	12.078	0.668	13.923	0.635	132.878	0.660	2.985
0.0384	0.703	43.031	0.691	8.555	0.666	10.972	0.634	126.842	0.658	2.425
0.1152	0.698	20.099	0.687	5.750	0.662	4.931	0.632	117.857	0.656	0.984

as high as 10%. For TCP connections with long round trip time and relatively small values of maximum window size, the *PD-TFS* scheme outperforms the other schemes in terms of channel utilization as well.

The following provides a brief summary of the key simulation results:

- The optimum value of the maximum TCP window size depends largely on network delay and error condition.
- Among all the scheduling schemes, the *PD-TFS* scheme gives the most throughput fairness and the value of the *TFP* metric is found to be the highest for the *PD-TFS* scheme. Throughput fairness with *SQ-FF* and *PD-LQF* based scheduling is considerably low compared to that for the other schemes. Comparable performance, in terms of throughput fairness, is observed with *PD-FF* and *PD-RR* based schemes.
- Achieved total throughput (and hence overall channel utilization) for connections with a relatively small W relative to *TD* is significantly low for *SQ-FF* and *PD-RR* schemes compared to that for the other scheduling schemes.
- As the number of concurrent TCP connections decreases, channel utilization falls off significantly with *SQ-FF* based scheduling compared to other schemes, especially when the maximum TCP window size is small compared to the window-delay product.
- The throughput fairness is better when the round trip time values of the corresponding connections are relatively large. A consistent increase in the through-

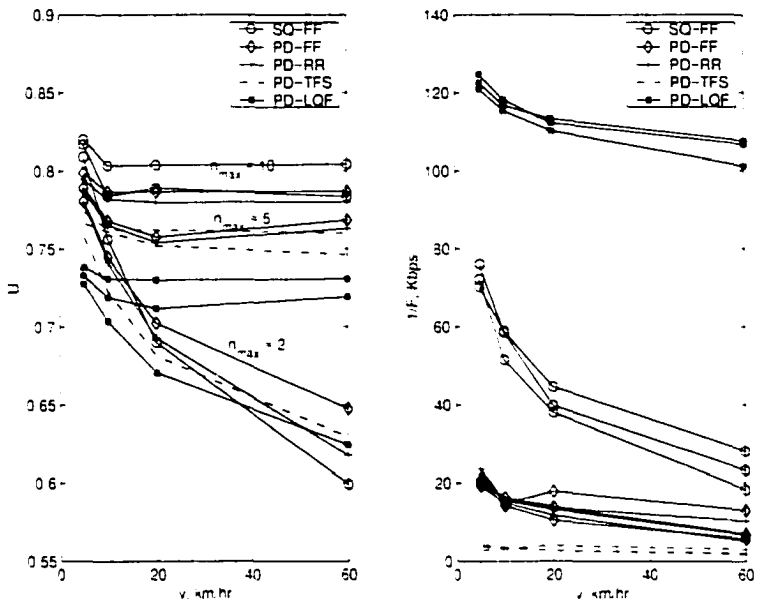

Figure 6.9. *Variation in U and $1/F$ for different values of n_{max} (for K = 3, TD = 10 ms, W = 5, N_c = 4, P_E = 0.1).*

put fairness is observed for all the scheduling schemes, except *PD-LQF*, as the number of concurrent TCP connections increases. Throughput fairness increases monotonically with increasing wired-network delay for *PD-LQF* scheduling.

- For a particular channel quality, other parameters remaining the same, as the terminal speed decreases (i.e., channel error-correlation increases), the channel utilization improves. On the other hand, throughput fairness generally improves as channel error events become more independent.
- When channel errors become more uncorrelated, smaller values of the RLC/MAC-layer retransmission parameter n_{max} (e.g., $n_{max} = 2$) degrades channel utilization.

Table 6.6. *Variation in U and $1/F$ for different values of K under different scheduling schemes (for* TD *= 10 ms, n_{max} = 20, P_E = 0.1, v = 5 km/hr).*

	SQ-FF		PD-FF		PD-RR		PD-LQF		PD-TFS	
K	U	1/F	U	1/F	U	1/F	U	1/F	U	1/F
	W=10, N_c=4									
2	0.712	94.167	0.698	19.222	0.673	20.291	0.639	141.453	0.665	3.120
3	0.712	99.174	0.697	18.787	0.673	21.728	0.639	141.927	0.664	3.987
5	0.712	96.630	0.697	18.538	0.673	20.397	0.639	141.743	0.664	4.175
	W=20, N_c=2									
2	0.636	113.360	0.657	38.956	0.645	36.580	0.633	88.173	0.645	5.526
3	0.636	111.171	0.657	38.994	0.645	36.585	0.633	88.144	0.641	5.547
5	0.636	110.171	0.657	38.972	0.645	36.561	0.633	88.195	0.645	5.535

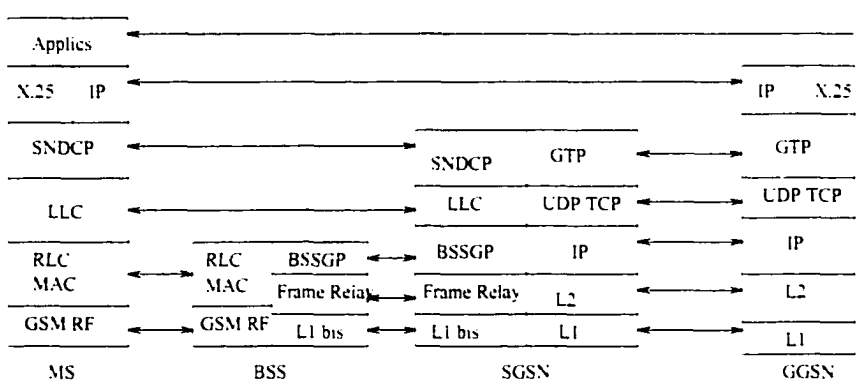

Figure 6.10. *GPRS transmission plane protocol architecture.*

Chapter 7

Link-Level Traffic Scheduling for Providing Predictive QoS in Wireless Multimedia Networks

7.1 Introduction

With increased focus on multimedia networking services and ubiquitous information access, 'multimedia migration' to mobile portable platforms seems to be the most significant issue in the next generation wireless mobile networks. Extending the high speed wireline networking infrastructure with a wireless access mechanism will make seamless end-to-end networking possible with consistent service characteristics [89]. It is expected that with mobility 'built-in' to future network API (Application Programming Interface) specifications, uniform fixed/wireless network API will make terminal mobility transparent to application software [90].

ATM-based technology can provide high speed wireless multimedia communications. In fact, the fine grain multiplexing provided by ATM due to the fixed small *cell*[1] size is well suited to slow-speed wireless links since it leads to lower delay jitter and queueing delays. In addition, the QoS specifiable VCs (Virtual Circuits) in the ATM provide the ability to meaningfully distinguish the data cells sent over the air and not treat all of them according to some generic policy [91]. In a wireless IP network, the same advantages can be derived through using relatively small link-layer frame size along with bandwidth allocation on 'per-flow' basis. From now on, the

[1]The term *cell* here refers to the basic transmission unit in ATM.

terms 'VC' and 'flow' will be used interchangeably.

The different traffic streams with different QoS requirements in a multimedia flow can be transported over different VCs. QoS requirements at the network level (e.g., bandwidth to be reserved, tolerable end-to-end delays and tolerable data loss rate) are directly derived from the transport characterization (e.g., average data rate, sensitivity to real-time skew in data, intermedia/intramedia synchronization) of multimedia flow [92].

Network QoS requirements are specified by parameters such as *CTD* (Cell Transfer Delay), *CDV* (Cell Delay Variation) and *CLP* (Cell Loss Probability). For nrt-VBR (Non-Real-Time Variable Bit Rate), ABR (Available Bit Rate), UBR (Unspecified Bit Rate) and GFR (Guaranteed Frame Rate) traffic, *CLP* is the most important QoS parameter to be guaranteed. Both *CLP* and *CTD* are to be guaranteed for CBR and rt-VBR traffic, and if it is assumed that the *CDV* can be compensated by using buffering and output control at the BS or AP (Access Point), it gets merged with the *CTD* issue. For wireless multimedia networks, the general QoS requirements (e.g., loss and delay requirements) for the different components in a multimedia traffic stream in terms of different ATM service classes can be described as in Table 7.1.

Table 7.1. *Different service classes for wireless multimedia networks.*

Service class	Example application	Bandwidth reqd.(*bps*)	Maximum delay (*ms*)	Target CER
CBR	Voice telephony, digital TV	32 K - 2 M	30-40	$< 10^{-2}$
rt-VBR	Real-time video conferencing	32 K - 2 M (avg.) 128 K - 6 M (peak)	40-90	$< 10^{-3}$
nrt-VBR	Multimedia email Digital video	1 M - 10 M	unbounded	$< 10^{-6}$
ABR, GFR	Web browsing	1 M - 10 M	unbounded	$< 10^{-5}$
UBR	File transfer	1 M - 10 M	unbounded	$< 10^{-8}$

The four basic degrees of QoS provided by a network along with the corresponding

QoS reliability (in a general sense) and the resulting resource utilization are listed in Table 7.2 [93]. Guaranteed QoS is usually achieved by analyzing the worst-case traffic and reserving sufficient resources to provide requested QoS even under the worst network condition. When the flows are bursty, because of the isolation of resources, guaranteed QoS inevitably results in low resource utilization [94]. With statistical QoS, it is promised that no more than a specified fraction of cells will see a performance below a certain specified value. In general, the algorithms providing statistical QoS result in lower resource utilization compared to the algorithms providing predictive QoS. Predictive QoS differs from best-effort QoS in that the latter does not consider the QoS requirements of the applications at all.

Providing predictive QoS has the advantage of not requiring the call to specify its traffic parameters, but the call must belong to one of a predefined set of classes and its traffic should correspond to the traffic characteristics of that class if the guarantees are to be reliable. Measurement-based bandwidth allocation and admission control schemes [95] can be employed to avoid the requirement of tight traffic modeling. Fluctuations in the network load can produce variations in the provided QoS. Measurement-based call admission control combined with the relaxed predictive QoS commitment can achieve a high level of network utilization. Since predictive QoS provides higher resource utilization, it is particularly desirable for a resource (e.g., channel bandwidth) limited wireless environment.

In a wireless mobile network, predictive QoS can be ideal for the scalable multimedia applications which are expected to tolerate some occasional degradations in network performance due to time and space dependent channel errors and user mobility. Predictive QoS control in a wireless multimedia network can be achieved by measurement-based call admission and traffic priority-based dynamic bandwidth allocation in MAC protocol, as will be described here.

In this chapter, a set of burst-level per-VC cell scheduling schemes is presented for transmission of multiservice traffic over TDMA/TDD channels in a WATM (Wireless ATM) network. Performance evaluation of the schemes is carried out using computer simulation, assuming correlated channel fading. The work is based on the premise that multimedia traffic is often bursty and that, in a wireless multimedia network, the different multimedia applications can tolerate occasional degradations in QoS. There-

Table 7.2. *Degree of QoS.*

QoS degree	Realization of QoS	QoS reliability	Resource utilization
guaranteed	guaranteed	highest	lowest
statistical	may be violated (with a certain probability)	< guaranteed	> guaranteed
predictive	network may sometimes fail	< statistical	> statistical
best-effort	QoS is ignored	lowest	highest

fore. burst-level cell scheduling at the link level for dynamic bandwidth allocation to different admitted connections may be ideal for providing predictive QoS while achieving high channel utilization. By integrating with proper traffic conditioning schemes the scheduling framework can be adapted as the MAC policy for wireless IP networks.

The organization of the rest of the chapter is as follows. Section 7.2 describes the primary issues of concern in link-layer traffic scheduling in wireless multimedia networks (e.g., WATM networks) and the relevance of our work. Subsequently, the scheduling algorithms are presented in Section 7.3. Section 7.4 is concerned with the computer simulations of the proposed scheduling schemes. The effects of correlated wireless channel errors and using channel state information in the proposed scheduling framework are described in Section 7.5 along with the simulation results. In Section 7.6, the findings are summarized and the conclusions are stated.

7.2 Link-Layer Traffic Scheduling in Wireless Multimedia Networks

In a wireless network that supports integrated multiservice traffic, the link-layer traffic scheduling protocol is to be designed so that the QoS requirements of the traffic flows from the different mobile stations sharing the limited wireless channel bandwidth are satisfied and, at the same time, high channel utilization can be achieved. Implementation (in hardware and/or software) simplicity and energy efficiency are

treated as other important considerations. Most of the battery power is consumed while a mobile is in transmit and/or receive mode, while least power is consumed in standby mode. Moreover, frequent turnaround between transmit and receive modes, which typically takes between 6 to 30 μs, can cause considerable wastage of 'channel time' and battery power. By using reservation and/or polling-based techniques, by provisioning for broadcasting of the transmission schedule of the mobiles from the BS/AP and by allocating contiguous slots to the mobiles for transmission and reception, the 'bits per joule' rating (i.e. energy efficiency) of the traffic scheduling policy can be significantly increased [96].

In a wireless network, TDMA/TDD-based cell multiplexing schemes provide better flexibility in controlling the available channel bandwidth by dynamically allocating the length of each downlink and uplink period in a TDD frame. In addition, the multipath interference can be mitigated to a reasonable extent by deploying an adaptive antenna array only at the BS/AP. This sort of spatial diversity protection would be necessary in a wireless multimedia network since techniques such as frequency diversity and time diversity, which are usually used for narrowband circuit switched systems, may not be appropriate in a broadband environment [97]. For FDD, to combat multipath fading, diversity antennas are needed to be deployed at both the BS/AP and the mobile station. TDMA/TDD-based cell multiplexing scheme based on centralized scheduling can also simplify the management of multicasting traffic in a *cell*[2].

Here, it is shown that, traffic priority based dynamic per-VC (or per-flow) scheduling for transmission of multimedia traffic over TDMA/TDD channels can provide high channel utilization with a tolerable loss and delay performance in a wireless network. At the same time, an energy-efficient MAC protocol can be realized. Per-VC scheduling is preferred to cell by cell scheduling since the latter can be a significant burden on the BS/AP. The idea here is to reserve some bandwidth (which is changed dynamically) during each frame-time for each class of traffic, and to distribute the excess bandwidth according to the instantaneous needs of the admitted connections.

The burst-level cell scheduling schemes that are being presented to transmit mul-

[2]Here, *cell* refers to the coverage area of a BS/AP. The two different meanings of *cell* will be apparent from the context.

tiservice traffic over TDMA/TDD channels in brodband wireless networks are FCFS-FR (First Come First Served with Frame Reservation). FCFS-FR+. EDF-FR (Earliest Deadline First with Frame Reservation). EDF-FR+ and MTDR (Multitraffic Dynamic Reservation). All of these are centralized schemes with some resource isolation (i.e., bandwidth reservation) for each of the different traffic classes coupled with dynamic allocation. The number of slots allocated to a VC in a frame is determined by the traffic type, the amount of traffic load, TOA (Time of Arrival)/TOE (Time of Expiry) value of the data burst and/or the length of the data burst.

FCFS-FR and EDF-FR are based on TOA and TOE values of data bursts, respectively, while FCFS-FR+ and EDF-FR+ also take queue-lengths corresponding to the different VCs into consideration while performing scheduling computations. MTDR is based on both TOE and queue-length. In the cases of FCFS-FR, FCFS-FR+, EDF-FR and EDF-FR+, for each traffic class, cells corresponding to only one VC are scheduled for transmission during a frame-time if there are sufficient number of cells to transmit during that frame-time. For MTDR, all of the candidate VCs in a traffic class get a minimum share of bandwidth during a frame-time. Therefore, for FCFS-FR, FCFS-FR+, EDF-FR and EDF-FR+, each of the VCs is expected to get a larger number of slots at a time compared to the MTDR case. We present simulation results for different traffic scenarios. The performances of EDF-FR+ and MTDR schemes are also evaluated in the presence of correlated channel errors. In this case, the channel state information is exploited so that it becomes robust to channel impairments to some extent.

A round-robin-based link-level channel-state-dependent packet scheduling (CS-DPS) was proposed in [59] mainly as a means of enhancing TCP performance for a system featuring RTS (Request to Send)-CTS (Clear to Send) messaging. This scheme was enhanced by combining it with CBQ (Class-Based Queueing) in [58] even though the QoS requirements of different multimedia traffic were not considered. In fact, in a high-speed networking scenario, per-cell scheduling may not be computationally feasible and per-VC scheduling is more desirable from an implementation point of view. Moreover, the RTS-CTS-based channel probing used in the above works may often result in discrepancy between the actual channel state and the estimated channel state since links are monitored only occasionally (frequency of which

is determined by a parameter g).

With the TDMA/TDD-based MAC protocol being considered here, each mobile station suffering collision can send a 'probe message' during the RA slot (in a minislot) in each frame-time (which is typically very small). If the BS/AP receives it correctly, it can resume allocating bandwidth to the corresponding mobile station. Thus, it is possible to sample the channel status in a more 'fine-grained' manner in this case by using an efficient contention resolution scheme in the RA minislots and, consequently, a better channel estimation may be achieved.

It is observed that most of the MAC schemes proposed for broadband wireless networks (e.g., WATM networks) in literature take into consideration voice and data traffic and/or simplified video traffic model [96]. The issue of energy-efficiency and the effects of time-varying channel errors on link-layer multitraffic scheduler performance have not been given much thought (except in very few cases, e.g., [96]). The EC-MAC (Energy Conserving Medium Access Control) protocol proposed in [96] uses a 'priority round-robin with dynamic reservation update and error compensation' scheduling where CBR and VBR traffic are prioritized over UBR traffic and within the same traffic type different connections are treated using round-robin mechanism. TOA/TOE values of data bursts are not explicitly taken into account for scheduling.

Our work presented in this chapter complements some of the works on TDMA/TDD-based MAC protocols for dynamic bandwidth allocation in multiservice wireless networks, e.g., MASCARA (Mobile Access Scheme Based on Contention and Reservation for ATM) [65], MDR (Multiservices Dynamic Reservation)-TDMA/TDD protocol [63] and EC-MAC [96]. The work here is different in that the scheduling schemes considered are based on 'soft' resource isolation for different types of traffic and at the same time dynamic allocation based on the instantaneous needs of the connections carrying the same type of traffic. The priority of a connection is determined based on the traffic type, the TOA/TOE value and/or the data burst length. Simulation results reveal that the EDF-FR+ and MTDR schemes can provide reasonably good QoS with high channel utilization in a multiservice traffic environment. FCFS-FR, FCFS-FR+ and EDF-FR schemes provide inferior cell multiplexing performance. In particular, the EDF-FR scheme overprovisions CBR traffic and causes high cell loss for rt-VBR traffic. In addition, schemes such as EDF-FR+ and MTDR are easy to implement and

provide qualitatively better power efficiency (due to contiguous slot allocation and a fewer number of transceiver turnarounds). The novelty of such a scheme lies in its QoS-awareness, simplicity, flexibility, link-state awareness and power-efficiency. The QoS degree considered here is predictive QoS. The performance evaluation is carried out in the presence of long-range dependent (e.g., Pareto) data traffic, short-range dependent (e.g., Poisson) data traffic, H.261-coded bursty video traffic and short-range dependent voice traffic.

In this connection it is worthy to mention some of the packet-level traffic scheduling policies proposed for switches/routers in order to provide guaranteed QoS in wireline integrated services networks [98]. All fair queueing algorithms use weights to compute bandwidth allocations for different streams and the weights are computed from *a priori* knowledge of the traffic streams corresponding to different connections, which is often difficult for diverse type of multimedia traffic. Again, adapting fair queueing for the wireless networks is complicated due to the presence of time and space dependent channel errors [99]. This is also true for link sharing schemes based on CBQ. In addition, these type of schedulers are inefficient for highly bursty traffic, such as for MPEG video, whose throughput requirements change drastically over time. Even though the bandwidth shares in these rate based schedulers can be dynamically adjusted, the way this can be done in a wireless MAC protocol is not well understood.

In another class of scheduling called CBS (Cost Based Scheduling) [100], time-varying cost functions are specified for each traffic class and the objective is to schedule packets from different class to minimize total cost. Each packet's cost at any time can be calculated independently of any other queued packets. The calculation of packet priorities is not trivial, except for simple cost functions, and specifying service constraints using cost functions will be complicated.

7.3 Burst-Level Cell Scheduling Schemes

7.3.1 TDMA/TDD Frame Structure

It is assumed that each TDMA/TDD frame consists of a RA slot (comprised of N_r minislots), UT slots, ACK and DT slots as shown in Figure 7.1. The boundary

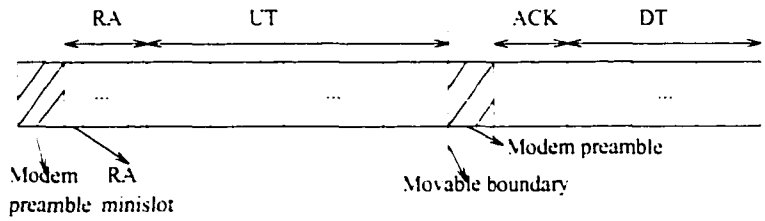

Figure 7.1. *TDMA/TDD frame format.*

between the *UT* and *ACK* slot(s) is not fixed and is determined by the BS/AP based on the symmetry of traffic load in the uplink and downlink directions. For some traffic classes (e.g., video-conferencing and voice) it is more likely that, for each VC in the uplink direction, there will be one VC in the downlink direction. This may not be the case for data traffic. The boundary between *ACK* and *DT* slots is also movable.

7.3.2 Description of the Protocols

The scheduling protocols perform burst-level cell multiplexing at the air interface for the multitraffic cells corresponding to the different VCs. Each VC is identified by a {MTid (Mobile Terminal id), VCid} pair. For each new data burst, the mobile station transmits the reservation request information, such as traffic type, queue-length status and TOA/TOE value (which is assumed to be same for all the cells in a data burst), in one of the minislots during the *RA* subperiod and then it switches to standby mode. A mobile station may have more than one VC active at a time, each of which is presumably carrying traffic of a specific class in a multimedia flow. S-ALOHA-based contention resolution can be used in the reservation of minislots. Some spread-spectrum-based signaling can provide improved performance for the contention resolution mechanism in the reservation minislots.

After the reservation requests are successfully received, the BS/AP registers the corresponding VCs as 'active' for service and stores the reservation information. Then it executes any one of the scheduling algorithms (as described later) to compute the allocation of *UT* slots in the next frame among the 'active' VCs. It transmits the slot allocation information (e.g., total number of *UT* slots, *UT* slots allocated to each user and their positions in the frame, status of the *DT* slots) in the *ACK* slot(s)

along with the acknowledgement information for the UT slots in the current frame after the UT slots pass by in the current frame (Figure 7.1). The mobile stations in the standby mode switch to receive mode at the start of the ACK slot(s). This strategy will cause, at most, two turnarounds between transmit and receive mode in a frame-time compared to three transceiver turn arounds per frame-time as in the EC-MAC [96] protocol. If a new data burst arrives at the mobile terminal while it is active in transmitting the previous data burst, it can send the new buffer information to the BS/AP without going through the reservation process. In this case, the cells corresponding to the previous burst are assumed to be lost and the mobile terminal informs the BS/AP of the new TOA/TOE value of the corresponding data burst.

7.3.3 Scheduling Algorithms

7.3.3.1 Parameters and Notation

The BS/AP keeps the following information for all the admitted VCs in the system: VCid and the corresponding MTid, traffic class (i.e., CBR/rt-VBR /nrt-VBR/ABR/-UBR/GFR), queue-length status and TOA/TOE value. For simplicity, we consider three traffic classes: CBR, rt-VBR and nrt-VBR. Here, the nrt-VBR class is assumed to include also ABR, UBR and GFR traffic. After each slot, the BS/AP updates the TOA/TOE information for each of the 'active' VCs (i.e., VCs for which reservation requests have been successfully transmitted and these VCs are waiting for bandwidth allocation in the next frame).

The following parameters are defined: S_t, S_{ra}, S_{ut}, S_{ack} and S_{dt} denote the total number of slots, the number of RA slots, the number of UT slots, the number of ACK slots and the number of DT slots, respectively, in a TDMA/TDD frame. Then, $S_t = S_{ra} + S_{ut} + S_{ack} + S_{dt}$. R_{CBR}, R_{rt-VBR} and $R_{nrt-VBR}$ denote the number of slots reserved for CBR VCs, rt-VBR VCs and non-real-time (i.e., nrt-VBR/ABR/GFR) VCs, respectively, in each frame-time and $r_{CBR}(t)$, $r_{rt-VBR}(t)$, $r_{nrt-VBR}(t)$ denote the corresponding instantaneous values. VC_i(CBR), VC_i(rt-VBR) and VC_i(nrt-VBR) denote the ith CBR, ith VBR and ith nrt-VBR VC, respectively. N_{CBR}, N_{rt-VBR} and $N_{nrt-VBR}$ denote the number of admitted CBR, rt-VBR and nrt-VBR VCs in the system, respectively, and $n_{CBR}(t)$, $n_{rt-VBR}(t)$, $n_{nrt-VBR}(t)$ denote the instantaneous

number of the 'active' CBR, rt-VBR and nrt-VBR VCs. Q_i denotes the queue-length corresponding to VC_i.

$VC_i(CBR)$s are given priority over VC_i(rt-VBR)s due to their more stringent delay and loss requirements. A minimum amount of bandwidth is reserved during each frame-time for audio, real-time video and non-real-time traffic by setting a nonzero value for R_{CBR}, R_{rt-VBR} and $R_{nrt-VBR}$, respectively. The values of the parameters R_{CBR}, R_{rt-VBR} and $R_{nrt-VBR}$ can be changed dynamically depending on the instantaneous traffic load. The maximum and minimum values of these parameters would depend on the admission control and pricing policies for CBR, rt-VBR and nrt-VBR VCs. The simplest adaptation policy is to toggle them between some fixed value of $R_{CBR}/R_{rt-VBR}/R_{nrt-VBR}$ and 0, depending on whether there are any 'active' CBR/rt-VBR/nrt-VBR VCs when scheduling computations are performed. The values of the parameters R_{CBR}, R_{rt-VBR} and $R_{nrt-VBR}$ can be varied adaptively depending on the QoS satisfaction of the 'active' VCs using some measurement-based scheme. For a fixed value of S_t, as has already been mentioned, the value of S_{uta} depends on the traffic symmetry in the uplink and downlink traffic. If we assume that for each of the VC_i(CBR)/VC_i(rt-VBR)s in the uplink direction, there is a corresponding VC_i(CBR)/VC_i(rt-VBR) in the downlink direction, then for CBR, rt-VBR and nrt-VBR traffic present in both directions, S_{uta} can be approximated by $(S_t - S_{ra} - S_{ack}) * N_{(nrt-VBR).uplink}/N_{(nrt-VBR).downlink}$ if we assume that all the nrt-VBR VCs carry similar traffic load.

7.3.3.2 Algorithms

It is assumed that the scheduling computations for frame-time t are performed just after the RA slot in the previous frame-time. The algorithms for determining the number of slots to be allocated to each of the 'active' VCs are described below.

FCFS-FR:

Let $P_{CBR}(t)$, $P_{rt-VBR}(t)$ and $P_{nrt-VBR}(t)$ denote the set of 'active' VC_i(CBR), VC_i(rt-VBR) and VC_i(nrt-VBR)s, respectively, during tth frame-time (i.e., one frame-time later than the scheduling computation time).

Let $P(t) = P_{CBR}(t) \cup P_{rt-VBR}(t) \cup P_{nrt-VBR}(t)$. In $P_{CBR}(t)$, $P_{rt-VBR}(t)$, $P_{nrt-VBR}(t)$ and in $P(t)$ the VC numbers are ordered with ascending TOA values of the corre-

sponding data bursts. The ordering among the VCs with the same TOA value is assumed to be random.

Procedure FCFS_FR() {

1. If ($n_{CBR}(t) = 0$) {
   $r_{CBR}(t) = 0$. $Z(t) = R_{CBR}$
   }
   Else {
   $r_{CBR}(t) = R_{CBR}$. $Z(t) = 0$. $i = 0$
   While ($i < |P_{CBR}(t)|$ and $r_{CBR}(t) > 0$) {
   Allocate $min(Q_i(t), r_{CBR}(t))$ slots to VC_i in $P_{CBR}(t)$
   $r_{CBR}(t) = r_{CBR}(t) - min(Q_i(t), r_{CBR}(t))$
   $i++$
   } /* While */
   $Z(t) = Z(t) + r_{CBR}(t)$
   } /* Else */
2. If ($n_{rt-VBR}(t) = 0$) {
   $r_{rt-VBR}(t) = 0$. $Z(t) = Z(t) + R_{rt-VBR}$
   }
   Else {
   $r_{rt-VBR}(t) = R_{rt-VBR}$. $i = 0$
   While ($i < |P_{rt-VBR}(t)|$and $r_{rt-VBR}(t) > 0$) {
   Allocate $min(Q_i(t), r_{rt-VBR}(t))$ slots to VC_i in $P_{rt-VBR}(t)$
   $r_{rt-VBR}(t) = r_{rt-VBR}(t) - min(Q_i(t), r_{rt-VBR}(t))$
   Update Q_i corresponding to VC_i accordingly
   $i++$
   } /* While */
   $Z(t) = Z(t) + r_{rt-VBR}(t)$
   } /* Else */
3. If ($n_{nrt-VBR}(t) = 0$) {
   $r_{nrt-VBR}(t) = 0$. $Z(t) = Z(t) + R_{nrt-VBR}$
   }
   Else {

$r_{nrt\text{-}VBR}(t) = R_{nrt\text{-}VBR}$. $i = 0$

While ($i < |P_{nrt\text{-}VBR}(t)|$ and $r_{nrt\text{-}VBR}(t) > 0$) {

Allocate $min(Q_i(t), r_{nrt\text{-}VBR}(t))$ slots to VC_i in $P_{nrt\text{-}VBR}(t)$

$r_{nrt\text{-}VBR}(t) = r_{nrt\text{-}VBR}(t) - min(Q_i(t), r_{nrt\text{-}VBR}(t))$

Update Q_i corresponding to VC_i accordingly

$i++$

} /* While */

$Z(t) = Z(t) + r_{nrt\text{-}VBR}(t)$

} /* Else */

4. $i = 0$

While ($i < |P(t)|$ and $Z(t) > 0$) {

Allocate $min(Q_i(t), Z(t))$ slots to VC_i in $P(t)$

$Z(t) = Z(t) - min(Q_i(t), Z(t))$

Update Q_i corresponding to VC_i accordingly

$i++$

} /* While */

}

FCFS-FR+:

FCFS-FR+ is similar to FCFS-FR except that among the VCs with the same TOA value, the one with the highest Q_i value is prioritized over the others.

EDF-FR:

Let $P_{CBR}(t)$, $P_{rt\text{-}VBR}(t)$ and $P_{nrt\text{-}VBR}(t)$ denote the set of 'active' VC_i(CBR), VC_i(rt-VBR) and VC_i(nrt-VBR)s, respectively, during tth frame-time where the VC numbers are arranged in the ascending order of the TOE values of the corresponding data bursts. Then, the allocation algorithm is similar to FCFS-FR.

EDF-FR+:

EDF-FR+ is similar to EDF-FR except that among the VCs with the same TOE value, the one with the highest Q_i value is prioritized over the others.

MTDR:

Let $P_{CBR}(t)$, $P_{rt\text{-}VBR}(t)$ and $P_{nrt\text{-}VBR}(t)$ denote the set of 'active' $VC_i(CBR)$, VC_i(rt-VBR) and VC_i(nrt-VBR), respectively, during tth frame-time where the VC numbers are arranged in the ascending order of the TOE values of the corresponding

data bursts. The VCs with the same TOE value are arranged in the descending order of the Q_i values.

Procedure MTDR() {

1. Initialize: $a = b = 0$

If $(n_{CBR}(t) = 0)$, $r_{CBR}(t) = 0$ Else $n_{CBR}(t) = R_{CBR}$

If $(n_{rt\text{-}VBR}(t) = 0)$, $r_{rt\text{-}VBR}(t) = 0$ Else $r_{rt\text{-}VBR}(t) = R_{rt\text{-}VBR}$

If $(n_{nrt\text{-}VBR}(t) = 0)$, $r_{nrt\text{-}VBR}(t) = 0$ Else $r_{nrt\text{-}VBR}(t) = R_{nrt\text{-}VBR}$

If $(n_{CBR}(t) \neq 0)$, $a = \lfloor \frac{S_{uta} - r_{rt\text{-}VBR}(t) - r_{nrt\text{-}VBR}(t)}{n_{CBR}(t)} \rfloor$

Else if $(n_{rt\text{-}VBR}(t) \neq 0)$, $b = \lfloor \frac{S_{uta} - r_{nrt\text{-}VBR}(t)}{n_{rt\text{-}VBR}(t)} \rfloor$

2. Compute $Z(t) = S_{uta} - n_{CBR}(t).a - n_{rt\text{-}VBR}(t).b + I_A.r_{rt\text{-}VBR}(t) + r_{nrt\text{-}VBR}(t)$

where $I_A = \begin{cases} 1 & \text{if } n_{CBR}(t) > 0 \\ 0 & \text{otherwise} \end{cases}$

3. If $(n_{CBR}(t) \neq 0)$ {

For each $VC_i(CBR)$ {

If $(Q_i(t) < a)$ {

Allocate Q_i slots to $VC_i(CBR)$

$Z(t) = Z(t) + (a - Q_i(t))$

$Q_i(t) = 0$

} /* If */

Else allocate a slots to VC_i and $Q_i(t) = Q_i(t) - a$

} /* For */

/* allocate slots on a round-robin basis among the VCs in $P_{CBR}(t)$ one slot each time */

$i = 0$

While $(P_{CBR}(t) \neq \emptyset$ *and* $Z(t) > r_{rt\text{-}VBR}(t) + r_{nrt\text{-}VBR}(t))$ {

Allocate 1 slot to $VC_i(CBR)$ in $P_{CBR}(t)$

Decrement $Z(t)$ and $Q_i(t)$

If $(Q_i(t) = 0)$ $P_{CBR}(t) = P_{CBR}(t) - \{VC_i(CBR)\}$

$i = (i + 1) \bmod |P_{CBR}(t)|$

} /* While */

} /* If */

Else if $(n_{rt\text{-}VBR}(t) \neq 0)$ {

For each $VC_i(rt - VBR)$ {

If ($Q_i(t) < b$) {
Allocate Q_i slots to $VC_i(rt - VBR)$
$Z(t) = Z(t) + (b - Q_i(t))$
$Q_i(t) = 0$
} /* If */
Else allocate b slots to VC_i and $Q_i(t) = Q_i(t) - b$
} /* For */
/* allocate slots on a round-robin basis among the VCs
in $P_{rt-VBR}(t)$ one slot each time */
$i = 0$
While ($P_{rt-VBR}(t) \neq \emptyset$ *and* $Z(t) > r_{nrt-VBR}(t)$) {
Allocate 1 slot to $VC_i(rt - VBR)$ in $P_{rt-VBR}(t)$
Decrement $Z(t)$ and $Q_i(t)$
If ($Q_i(t) = 0$) $P_{rt-VBR}(t) = P_{rt-VBR}(t) - \{VC_i(rt - VBR)\}$
$i = (i + 1) \bmod |P_{rt-VBR}(t)|$
} /* While */
} /* Else if */
$i = 0$
While ($P_{nrt-VBR}(t) \neq \emptyset$ *and* $Z(t) > 0$) {
Allocate 1 slot to $VC_i(nrt - VBR)$ in $P_{nrt-VBR}(t)$
Decrement Z(t) and $Q_i(t)$
If ($Q_i(t) = 0$) $P_{nrt-VBR}(t) = P_{nrt-VBR}(t) - \{VC_i(nrt - VBR)\}$
$i = (i + 1) \bmod |P_{nrt-VBR}(t)|$
} /* While */
}

The scheduling algorithm first attempts to allocate all the slots (excluding those reserved for rt-VBR and nrt-VBR traffic) uniformly to all the 'active' CBR VCs. The left-over slots (if any) are then allocated to those 'active' CBR VCs requiring more slots in a round-robin fashion starting from the one with the highest Q_i value among the VC(s) with the lowest TOE value. The r_{rt-VBR} slots and the slots left-over (if any, excluding the $r_{nrt-VBR}$ slots) are allocated to the rt-VBR VCs. The $r_{nrt-VBR}$ slots and the left over slots (if any) are then allocated to the 'active' nrt-VBR VCs

in a round-robin fashion starting with the one having the highest Q_i value among the VC(s) having the lowest TOE value. The algorithm can easily be extended for a larger number of traffic classes. It is to be noted that, for a = b = 0, it reduces to simple round-robin based scheduling within each traffic class.

After the number of slots to be allocated to each VC is determined, the allocations during the frame-time are made so that each mobile station gets contiguous slot allocation. The slot allocations during the frame-time are determined according to the priority based on the TOE values and the corresponding traffic class (with CBR class having the highest priority and the nrt-VBR class the lowest). In the case that two VCs in the same traffic class have the same TOE value, the one with the higher Q_i value is prioritized over the other. Next, the scheduler checks whether the TOE value reaches zero for any one of the VCs to which slots have been allocated. If it happens, then, in order to avoid wasting bandwidth, the slot(s) are allocated to the VC for which the next immediate slot has been allocated during this frame-time.

7.3.3.3 Computational Complexity

Each time a new VC is admitted into the system, the BS/AP controller updates the relevant data-structure (e.g., linked list) for the admitted VCs at the cost of $O(n)$ operations, where n is the total number of admitted VCs. The scheduling computations are to be done once in each frame-time. For all the above algorithms, the computational complexity is $O(n)$. In fact, for practical values of n the computation time would not be significant at all.

7.4 Simulation of the Scheduling Schemes

7.4.1 Traffic Models and Performance Measures

Voice traffic generated by a vocoder with fast speech activity detector [101] is classified as CBR traffic. During minispurts, the voice source generates continuous bitstreams. A typical traffic trace of 3000 s for this model is shown in Figure 7.2. Real-time video-conferencing traffic is considered here as the representative of the rt-VBR class.

Video-conference sequences, generated by different hardware coders and different

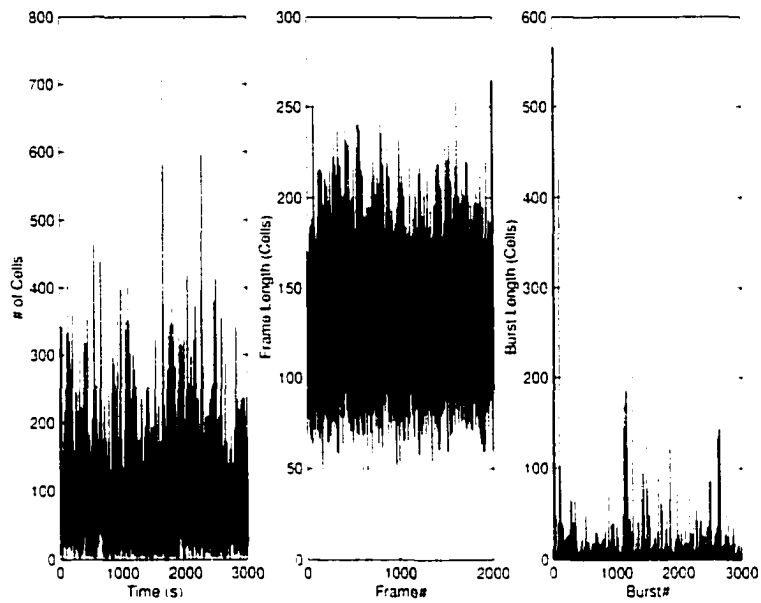

Figure 7.2. *Traffic trace generated from (a) a voice source with fast speech activity detector, (b) a GBAR source and (c) an LRD model (Pareto distribution with mean = 10, $a = 1.1$).*

coding algorithms, generally have negative binomial marginal distributions and geometric autocorrelation functions and fit a DAR (Discrete Autoregressive) model well when several sources are multiplexed. But for a single source in isolation, DAR model does not work well, instead GBAR (Gamma-Beta Autoregressive) model gives a more accurate representation [67].

The GBAR(1) process described in [67] is used as the source model to generate the number of cells per video frame. A typical traffic trace showing the number of cells generated in consecutive video frames (up to 2000 frames) is given in Figure 7.2. Data burst duration of non-real-time multimedia traffic and asynchronous bursty computer data traffic corresponding to each nrt-VBR VC is modeled as a Pareto distributed process.

The performance measures considered are:

Average CBR/rt-VBR/nrt-VBR Cell Loss Probability ($CLP_{CBR}/CLP_{rt-VBR}/CLP_{nrt-VBR}$): it is the ratio of the number of CBR/rt-VBR/nrt-VBR cells dropped to the total number of CBR/rt-VBR/nrt-VBR cells generated. A cell that cannot

be transmitted within the tolerable delay (given by TOE value of a video cell) is discarded by the MT.

Average CBR/rt-VBR Cell Transfer Delay (CTD_{CBR}/CTD_{rt-VBR}): it is the time between the generation of a CBR/rt-VBR cell and the transmission of the same cell.

Channel Utilization (U): it is the ratio of the number of slots used for WATM cell transmission to the total number of slots available for transmission.

Average nrt-VBR Throughput ($T_{nrt-VBR}$): it refers to the volume of nrt-VBR data transmitted per second.

It is assumed here that predictive QoS can be provided to the voice, video and data traffic if the values of CLP_{CBR}, CLP_{rt-VBR} and $CLP_{nrt-VBR}$ are of $O(10^{-2})$, $O(10^{-3})$ and $O(10^{-6})$, respectively, and the values of CTD_{CBR} and CTD_{rt-VBR} are less than 10 *ms* and 15 *ms*, respectively.

7.4.2 Simulation Environment

A *C*-coded event-driven simulator is used to evaluate the performance of the scheduling schemes. There can be more than one VC admitted per mobile station at a time. All the cells comprising a burst have the same TOA/TOE value. For the nrt-VBR traffic type, data bursts are generated (after an exponentially distributed interarrival time) either after the current burst has been transmitted or has 'expired'. The amount of available buffers at a mobile station is assumed to be infinite. Since the performance measures of the scheduling algorithms are of major concern, the contention resolution process among the reservation requests during the *RA* subperiods is disregarded and it is assumed that all the contending mobile stations can successfully send their reservation requests during this period.

For all the results presented here, it is assumed that traffic is symmetric in both the mobile to BS/AP and the BS/AP to mobile directions. Therefore, the BS/AP scheduler allocates equal number of slots for uplink and downlink transmission during a frame-time. The cell delay variation (i.e., jitter) for CBR and rt-VBR traffic is not considered since it is bounded by the corresponding TOE values and the effects of jitter can be eliminated by buffering at the destination. The frame-length is assumed to be fixed. Fixed length frames provide an easier synchronization.

In all the cases, the simulator is run for 2 *million* frame-time (with the parameters

Table 7.3. *Simulation parameters.*

Parameter	Value	Parameter	Value
Channel rate	20 *Mbps*	Slot size	22 μs
No. of ACK slot	1	Principal talkspurt length	1.00 s
Principal silence gap length	1.35 s	Minispurt length	0.275 s
Minigap length	0.05 s	Voice coding rate	32 *Kbps*
Mean no. of cells/video frame	104.9	Variance of no. of cells/video frame	882.09
Correlation coefficient, ρ	0.984	Video interframe rate	40 ms
Mean nrt-VBR burst length	0.48 *KB*	Pareto shape parameter	1.1
Mean nrt-VBR burst interarrival time	1 s	TOE of a voice cell	32 ms
TOE of a video cell	40 ms	TOE of a data cell	5 s

shown in Table 7.3), which corresponds to 1320 s for a frame-size of 30 *slots*. Simulation statistics are flushed after 500000 frame-time to avoid the transient effects. The performance results presented here apply for both uplink and downlink traffic. The channel utilization is calculated without taking control overheads into account.

7.4.3 Simulation Results and Discussions

rt-VBR traffic only: Figure 7.3 shows the variations in U, $CTD_{rt\text{-}VBR}$ and $CLP_{rt\text{-}VBR}$ with $N_{rt\text{-}VBR}$ in a single traffic (rt-VBR only) scenario for the assumed simulation parameters. With FCFS-FR, FCFS-FR+ and EDF-FR scheduling, for 6 concurrent rt-VBR VCs, $CLP_{rt\text{-}VBR} \approx 3 \times 10^{-3}$ and $CTD_{rt\text{-}VBR} < 8$ ms for all these three cases. For the same $N_{rt\text{-}VBR}$, $CLP_{rt\text{-}VBR} \approx 5 \times 10^{-4}$ and $CLP_{rt\text{-}VBR} \approx 3 \times 10^{-5}$ for EDF-FR+ and MTDR scheduling schemes. In all the cases, $U \approx 0.84$. For $N_{rt\text{-}VBR} = 7$, $CLP_{rt\text{-}VBR} \approx 4 \times 10^{-2}$ for all the first three scheduling schemes whereas $CLP_{rt\text{-}VBR} \approx 7 \times 10^{-3}$ for EDF-FR+ based scheduling with $U \approx 0.95$ and $CTD_{rt\text{-}VBR} \approx 15$ ms. In the case of MTDR, $CLP_{rt\text{-}VBR} \approx 2 \times 10^{-2}$, $U \approx 0.97$ and $CTD_{rt\text{-}VBR} \approx 17$ ms. Even with a lower traffic load, $CLP_{rt\text{-}VBR}$ is higher for FCFS-

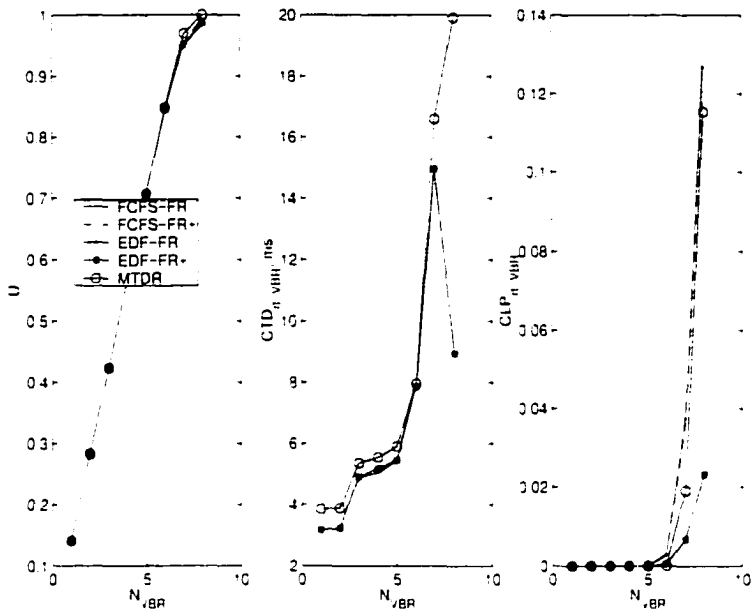

Figure 7.3. *Variation of U, $CTD_{rt\text{-}VBR}$ and $CLP_{rt\text{-}VBR}$ in a single traffic (rt-VBR only) environment.*

FR, FCFS-FR+ and EDF-FR schemes compared to the $CLP_{rt\text{-}VBR}$ for EDF-FR+ and MTDR schemes. Therefore, EDF-FR+ and MTDR schemes provide better cell multiplexing performance than the other three schemes.

Optimum frame-size depends on factors such as traffic load, traffic burstiness and traffic QoS requirements. It has been observed that with the assumed TOE values for CBR/rt-VBR data bursts, as the frame-size is decreased, the loss and delay performances improve except for very high traffic load. On the other hand, the smaller the frame-size, the more the control overhead will be. From several simulation experiments, we observe that frame size of 30 *slots* is a good choice and therefore use this frame-size to obtain all the simulation results presented here.

CBR and rt-VBR traffic: Some simulation results for the case of CBR and rt-VBR traffic mix are shown in Figures 7.4 and 7.8. The average cell loss and delay measures and the channel utilization depend on the proportion of each traffic in the total traffic load and on the value of the reservation parameters. It is observed that as the proportion of CBR traffic decreases, channel utilization increases with a consequent increase in the throughput of rt-VBR traffic.

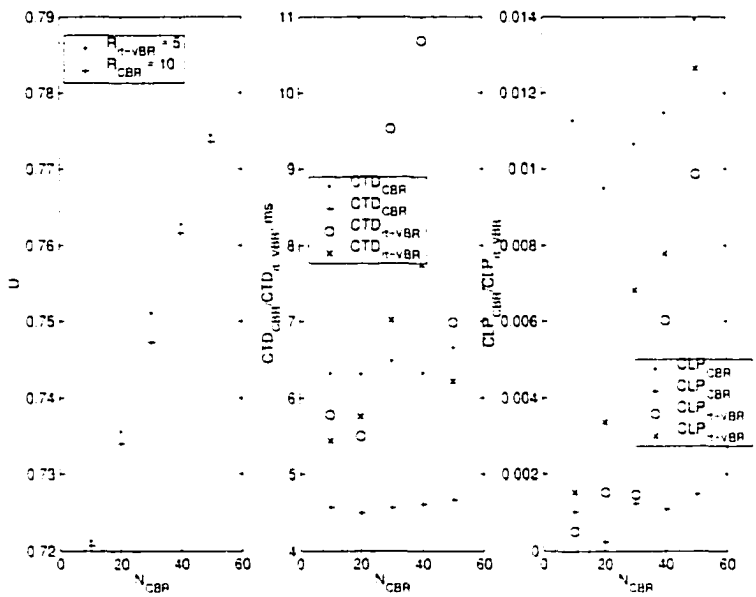

Figure 7.4. *FCFS-FR: Variation of U, $CTD_{CBR}/CTD_{rt\text{-}VBR}$ and $CLP_{CBR}/CLP_{rt\text{-}VBR}$ in a multitraffic (CBR and rt-VBR) environment (for $N_{rt\text{-}VBR} = 5$).*

In the case of FCFS-FR, with $N_{rt\text{-}VBR} = 5$ and $N_{CBR} = 30$, for $R_{CBR} = 0$ and $R_{rt\text{-}VBR} = 5$, $CLP_{CBR} \approx 1 \times 10^{-2}$ and $CLP_{rt\text{-}VBR} \approx 1 \times 10^{-3}$ with a resulting channel utilization of about 0.75. As the traffic load is increased further, the cell loss rates increase rapidly. For example, as N_{CBR} is increased to 50, $CLP_{CBR} \approx 1.4 \times 10^{-2}$ and $CLP_{rt\text{-}VBR} \approx 1 \times 10^{-2}$. As $R_{rt\text{-}VBR}$ is decreased and/or R_{CBR} is increased, $CLP_{rt\text{-}VBR}$ increases and CLP_{CBR} decreases. For example, with $R_{CBR} = 6$ and $R_{rt\text{-}VBR} = 0$, $CLP_{CBR} \approx 1 \times 10^{-3}$ and $CLP_{rt\text{-}VBR} \approx 7 \times 10^{-3}$ when $N_{rt\text{-}VBR} = 5$ and $N_{CBR} = 30$.

For the same traffic scenario and the same values for the reservation parameters, FCFS-FR+ provides slightly better $CLP_{rt\text{-}VBR}$ but at the expense of higher CLP_{CBR}. This is intuitive since the rt-VBR data bursts, which are expected to be relatively larger than the CBR data bursts, are prioritized for the same TOA value.

Simulation results show that the EDF-FR scheme overprovisions CBR traffic and, as a result, rt-VBR traffic suffers relatively high cell loss. For instance, even with $R_{rt\text{-}VBR} = 13$, $CLP_{CBR} \approx 2 \times 10^{-3}$ and $CLP_{rt\text{-}VBR} \approx 2 \times 10^{-3}$ when $N_{rt\text{-}VBR} = 5$

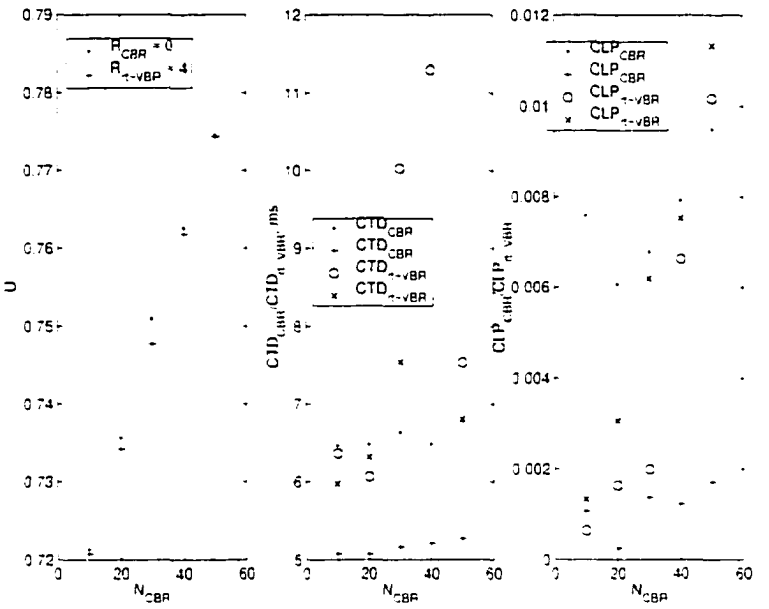

Figure 7.5. *FCFS-FR+: Variation of U, CTD_{CBR}/CTD_{rt-VBR} and CLP_{CBR}/CLP_{rt-VBR} in a multitraffic (CBR and rt-VBR) environment (for $N_{rt-VBR} = 5$, $R_{rt-VBR} = 0$).*

and $N_{CBR} = 20$. As CBR traffic load is increased, CLP_{rt-VBR} increases considerably. For example, $CLP_{CBR} \approx 5 \times 10^{-3}$ and $CLP_{rt-VBR} \approx 6 \times 10^{-3}$ when N_{CBR} is increased to 40. With $N_{rt-VBR} = 6$ and $N_{CBR} = 10$, for $R_{rt-VBR} = 11$, $CLP_{CBR} \approx 1 \times 10^{-2}$ and $CLP_{rt-VBR} \approx 8 \times 10^{-3}$ with a resulting channel utilization of about 0.85. As R_{rt-VBR} increases, both CTD_{CBR} and CLP_{CBR} increase.

CLP_{rt-VBR} is significantly lower with EDF-FR+ based scheduling than for the FCFS-FR, FCFS-FR+ and EDF-FR schemes. With $N_{rt-VBR} = 5$ and $N_{CBR} = 45$, for $R_{rt-VBR} = 4$, $CLP_{CBR} \approx 7.6 \times 10^{-3}$ and $CLP_{rt-VBR} \approx 1.2 \times 10^{-3}$ with a resulting channel utilization of about 0.77. As N_{CBR} is increased to 60, CLP_{CBR} and CLP_{rt-VBR} increase to about 1.0×10^{-2} and 2.4×10^{-3}, respectively, with a resulting channel utilization of about 0.79. As R_{rt-VBR} is increased, CLP_{rt-VBR} decreases at the expense of increased CLP_{CBR}. Under a similar traffic load, for MTDR-based scheduling with $R_{rt-VBR} = 4$, $CLP_{CBR} \approx 1 \times 10^{-2}$ and $CLP_{rt-VBR} \approx 3 \times 10^{-3}$. As N_{CBR} is increased to 80, CLP_{CBR} and CLP_{rt-VBR} increase to about 1.4×10^{-2} and

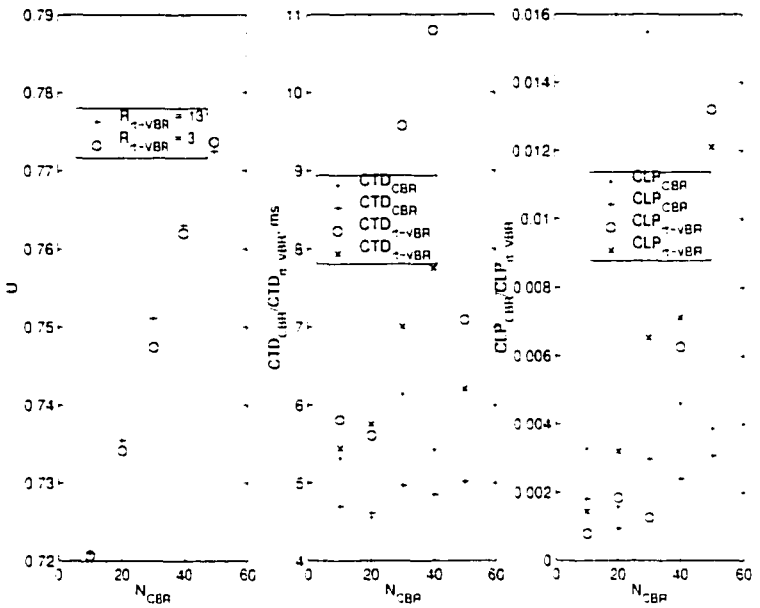

Figure 7.6. *EDF-FR: Variation of U, CTD_{CBR}/CTD_{rt-VBR} and CLP_{CBR}/CLP_{rt-VBR} in a multitraffic (CBR and rt-VBR) environment (for $N_{rt-VBR} = 5$, $R_{CBR} = 0$).*

5×10^{-3}, respectively, with a resulting channel utilization of about 0.82. As N_{CBR} is further increased, CLP_{rt-VBR} increases rapidly. With EDF-FR+ scheduling, for $N_{CBR} = 80$ and $N_{rt-VBR} = 5$, $CLP_{CBR} \approx 1.7 \times 10^{-2}$, $CLP_{rt-VBR} \approx 3.7 \times 10^{-3}$ and $U \approx 0.81$.

Simulation results show that, for MTDR-based scheduling, with $N_{rt-VBR} = 6$, $N_{CBR} = 50$ and $R_{rt-VBR} = 5$, $CLP_{CBR} \approx 1.7 \times 10^{-2}$ and $CLP_{rt-VBR} \approx 8.5 \times 10^{-3}$ with a resulting channel utilization of about 0.91. Under similar conditions, with EDF-FR+ scheme, $CLP_{CBR} \approx 1.5 \times 10^{-2}$, $CLP_{rt-VBR} \approx 5 \times 10^{-3}$ and $U \approx 0.90$. Therefore, EDF-FR+ provides better traffic multiplexing performance than does MTDR under different traffic load conditions.

In all of the above cases, CTD_{CBR} and CTD_{rt-VBR} are found to be well below the corresponding QoS limits. The variation in CTD_{rt-VBR} with traffic load is more random compared to that of CTD_{CBR}. This phenomenon may be attributed to the correlated and bursty nature of the video traffic and the dynamics of the scheduling

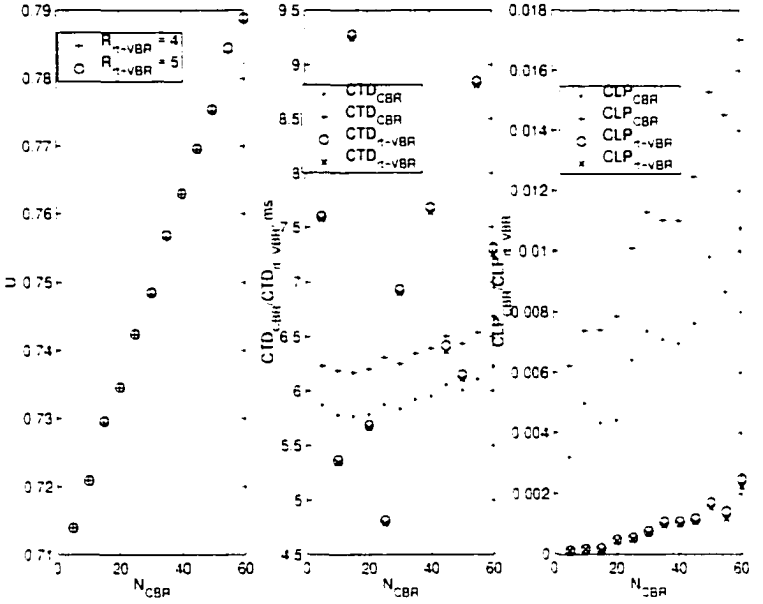

Figure 7.7. *EDF-FR+: Variation of U, CTD_{CBR}/CTD_{rt-VBR} and CLP_{CBR}/CLP_{rt-VBR} in a multitraffic (CBR and rt-VBR) environment (for $N_{rt-VBR} = 5$, $R_{CBR} = 0$).*

policies.

CBR, rt-VBR and nrt-VBR traffic: Figures 7.9 and 7.10 show some performance results for EDF-FR+ and MTDR schemes in a multitraffic (CBR, rt-VBR and nrt-VBR traffic mix) scenario. For a fixed number of admitted CBR VCs and rt-VBR VCs, as the number of admitted nrt-VBR VCs is increased, the channel utilization increases. With both EDF-FR+ and MTDR schemes, for the assumed simulation parameters, with $N_{CBR} = 30$, $N_{rt-VBR} = 5$ and $N_{nrt-VBR} = 20$, the soft QoS requirements for all the admitted traffic can be guaranteed with a resulting channel utilization of about 0.76. In fact, due to very large variance in the nrt-VBR burst size (resulting from the Pareto distribution), it is almost impossible to meet the very low loss requirement of nrt-VBR traffic while achieving high channel utilization when the delay of the nrt-VBR traffic is to be bounded. By limiting the burst size to some maximum value, better channel utilization can be achieved. The channel utilization also depends on the ratio of the real-time traffic load to total traffic load. It has been observed that, as the ratio decreases, channel utilization increases and, consequently,

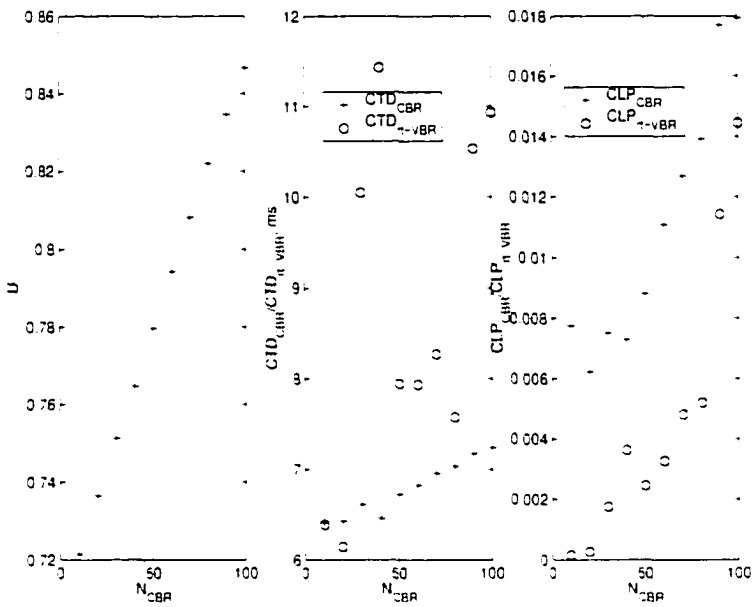

Figure 7.8. *MTDR: Variation of U, $CTD_{CBR}/CTD_{rt\text{-}VBR}$ and $CLP_{CBR}/CLP_{rt\text{-}VBR}$ in a multitraffic (CBR and rt-VBR) environment (for $N_{rt\text{-}VBR} = 5$, $R_{CBR} = 0$, $R_{rt\text{-}VBR} = 5$).*

$T_{nrt\text{-}VBR}$ increases.

Some simulation results are presented in Figure 7.6 for short-range dependent non-real-time traffic. In this case, it is assumed that the burst-lengths are exponentially distributed with a mean duration of 0.1 *s* with traffic rate of 1 *Mbps* and that the maximum tolerable delay of a burst is 5 *s*. It is observed that with both EDF-FR+ and MTDR schemes, channel utilization as high as 0.80 is achieved (for $N_{CBR} = 30$ and $N_{rt\text{-}VBR} = 5$) with tolerable QoS to all traffic streams when data traffic is not as bursty as in the LRD case (Figure 7.10).

7.5 Effect of Correlated Channel Errors and Channel-State Aware Scheduling

In a typical WATM network operating at a data rate as high as 20 *Mbps*, the TDMA/TDD frame-size is expected to have very small time duration. For exam-

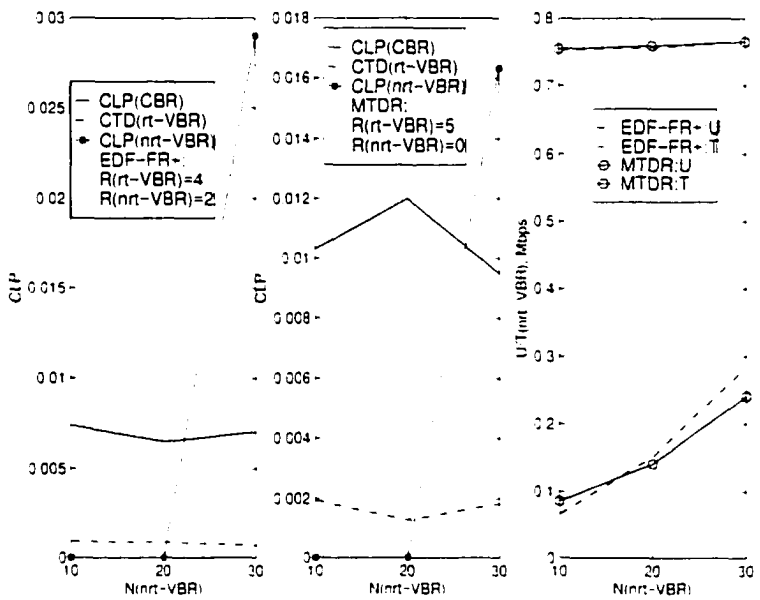

Figure 7.9. *EDF-FR+ and MTDR: Variation of CLP_{CBR}, CLP_{rt-VBR} with $N_{nrt-VBR}$ in a multitraffic (CBR, rt-VBR and LRD nrt-VBR) environment (for $N_{rt-VBR} = 5$, $N_{CBR} = 30$, for EDF-FR+: $R_{rt-VBR} = 4$, $R_{nrt-VBR} = 2$, for MTDR: $R_{rt-VBR} = 5$, $R_{nrt-VBR} = 0$).*

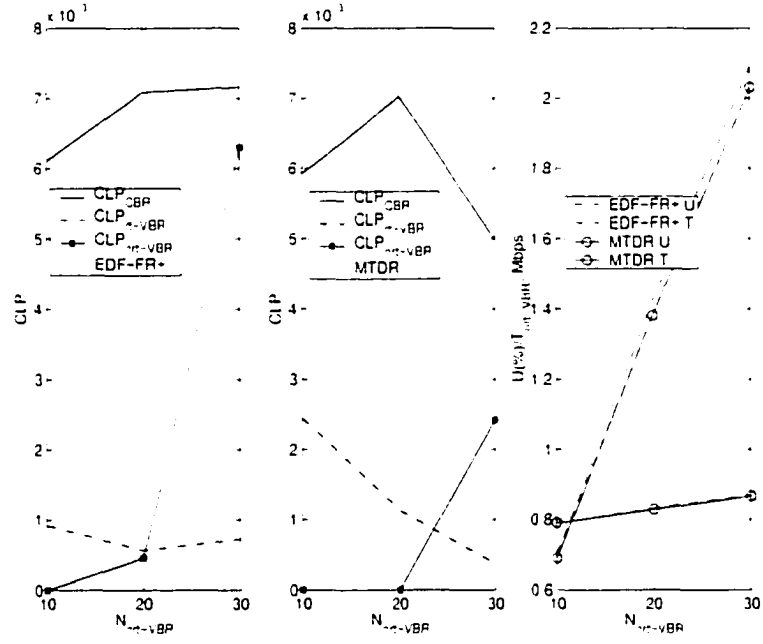

Figure 7.10. *EDF-FR+ and MTDR: Variation of CLP_{CBR}, CLP_{rt-VBR} with $N_{nrt-VBR}$ in a multitraffic (CBR, rt-VBR and non-LRD nrt-VBR) environment (for $N_{CBR} = 30$, $N_{rt-VBR} = 5$, $R_{rt-VBR} = 4$, $R_{CBR} = R_{nrt-VBR} = 0$).*

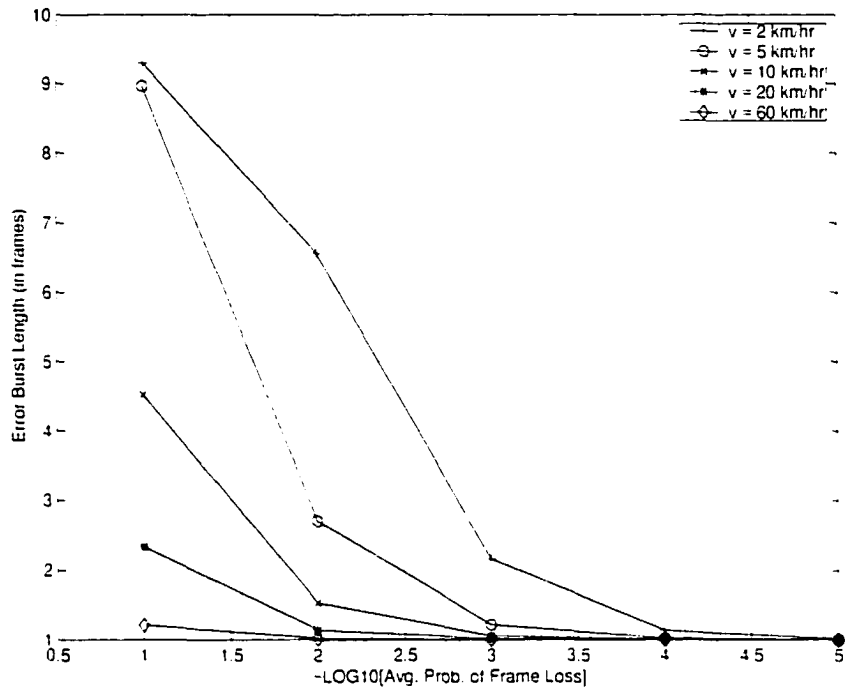

Figure 7.11. *Variation of average burst error length with P_E and v.*

ple. the size of a TDMA/TDD frame consisting of 30 cells is 0.66 *ms*. assuming that a WATM cell is of size 55 *bytes*. Therefore. it is more likely that in a bursty fading channel one or more of the frames will suffer error during a fade period.

For a certain value of average frame error rate P_E (particularly for a value in the range of 10^{-2} to 10^{-3}). the error burst-length increases (i.e.. the fading process becomes more correlated) as the user mobility (and hence f_dT) decreases. Other parameters remaining the same. as P_E increases. error burst length increases rapidly for low-mobility users (Figure 7.11).

By delaying transmissions of the traffic from a mobile station for which the transmission channel is expected to be noisy and by allowing the mobile stations to transmit for which the channel condition is expected to be good. better effective link utilization can be achieved. The channel status information can be exploited by the BS/AP scheduler as follows: after the uplink traffic corresponding to a VC is not ACKed by the BS/AP. the corresponding mobile station assumes that the cells have been lost and it starts transmitting 'probe' messages periodically during *RA* period. A few of the *RA* minislots can be reserved for these 'probe' messages. As soon as the BS/AP

receives one of these probe messages correctly, it assumes that the channel condition has improved and resumes scheduling the traffic corresponding to these VCs again for transmission.

EDF-FR+ and MTDR schemes are simulated using the above channel model where the channel evolution in each TDMA/TDD frame corresponding to each mobile is independent and is determined by the two paremeters p and q, which, in turn, are determined by the normalized Doppler frequency f_dT and the average frame loss rate P_E. SR-ARQ (Selective-Repeat ARQ)-based error recovery scheme is assumed. If a transmission attempt fails due to channel errors, the corresponding cells(s) are kept waiting in the transmit data buffer until they are successfully transmitted or the waiting time reaches the TOE value of the corresponding data burst.

The variations in U and CLP_{rt-VBR} with N_{rt-VBR} in a rt-VBR traffic only scenario is shown in Figure 7.12 for EDF-FR+ and MTDR. For $P_E = 0.01$ (i.e., channel fading margin = 19.978 dB) and $f_dT = 0.00308$, which corresponds to a user velocity of 1 km/hr for carrier frequency equal to 5 GHz, $CLP_{rt-VBR} \approx 1 \times 10^{-3}$ and $CLP_{rt-VBR} \approx 8 \times 10^{-4}$ for EDF-FR+ and MTDR, respectively, when $N_{rt-VBR} = 6$. The resulting channel utilization is about 0.84 for both the cases. For the same value of P_E and f_dT and with $N_{rt-VBR} = 7$, $CLP_{rt-VBR} \approx 7 \times 10^{-3}$ and $CLP_{rt-VBR} \approx 2 \times 10^{-2}$ for EDF-FR+ and MTDR, respectively, and the resulting channel utilization is about 0.95. Thus, as long as channel utilization is below certain limit (e.g., 0.84), both the schemes can provide tolerable loss and delay performances for an average frame error rate as high as 1%. Since with low user-mobility the channel fading events are more correlated, for a certain value of P_E, as v decreases, CLP_{rt-VBR} increases. Some typical simulation results for EDF-FR+ are enumerated in Table 7.4.

For CBR and rt-VBR traffic mix with $R_{rt-VBR} = 4$ and $R_{CBR} = 0$, for $P_E = 0.01$ and $v = 1$ km/hr, $CLP_{CBR} \approx 1 \times 10^{-2}$ and $CLP_{rt-VBR} \approx 7.5 \times 10^{-4}$ for EDF-FR+ when $N_{rt-VBR} = 5$ and $N_{CBR} = 30$ (Figure 7.13). In this case, the resulting channel utilization is about 0.75. Under similar traffic conditions, for the same value of the reservation parameters, $CLP_{CBR} \approx 9 \times 10^{-3}$ and $CLP_{rt-VBR} \approx 9 \times 10^{-4}$ for the MTDR scheme. As N_{rt-VBR} is increased to 50, $CLP_{rt-VBR} \approx 1.8 \times 10^{-3}$ and $CLP_{rt-VBR} \approx 2.4 \times 10^{-3}$ for EDF-FR+ and MTDR schemes, respectively. In this case, $U \approx 0.77$ and CTD_{CBR}/CTD_{rt-VBR} are well below the tolerable QoS limits. It

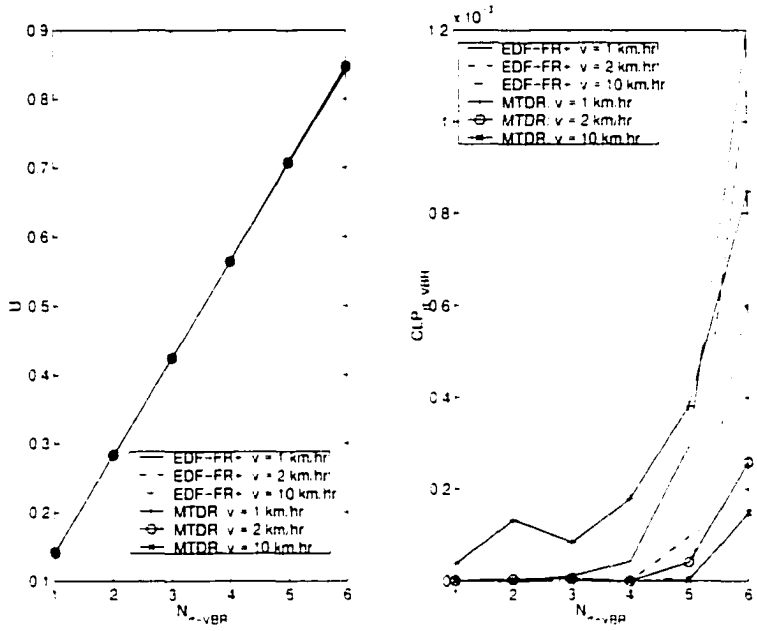

Figure 7.12. *EDF-FR+, MTDR: Effect of channel-error correlations on U and $CLP_{rt\text{-}VBR}$ in a single traffic (rt-VBR only) scenario (for $P_E = 0.01$)*

is observed that increasing the value of $R_{rt\text{-}VBR}$ causes large degradation in $CLP_{rt\text{-}VBR}$. Some typical simulation results for EDF-FR+ are enumerated in Table 7.5.

For CBR, rt-VBR and nrt-VBR traffic mix, with EDF-FR+ scheduling channel utilization of about 0.75 can be achieved with tolerable QoS for all traffic streams when $N_{CBR} = 25$ and $N_{rt\text{-}VBR} = 5$ for P_E as high as 0.01 (Figure 7.14). Under similar traffic load and for the same P_E, it is observed that MTDR provides inferior performance than EDF-FR+. For both the schemes, in the case of non-LRD traffic, the channel utilization increases significantly. With $N_{CBR} = 40$ and $N_{rt\text{-}VBR} = 5$, $U \approx 0.90$ and $CLP_{nrt\text{-}VBR} < 10^{-6}$ for the EDF-FR+ scheme in the case of non-LRD nrt-VBR traffic. In this case, $T_{nrt\text{-}VBR} \approx 2.7$ *Mbps*. Under similar traffic load channel utilization as high as 0.80 is achieved for MTDR scheme as long as $CLP_{nrt\text{-}VBR} < 10^{-6}$.

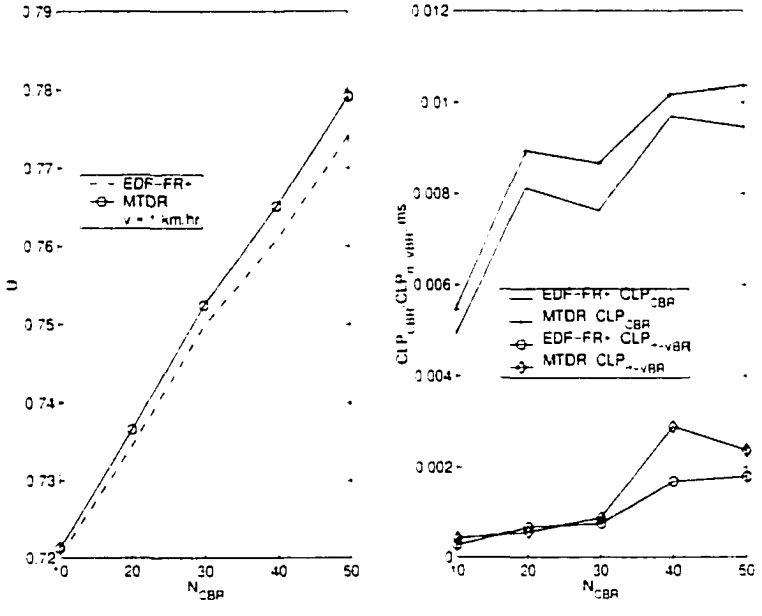

Figure 7.13. *EDF-FR+, MTDR: Effect of channel-error correlations on U and CLP_{rt-VBR} in a multitraffic (CBR and rt-VBR) scenario (for P_E = 0.01, N_{rt-VBR} = 5, R_{rt-VBR} = 4, R_{CBR} = 0).*

Table 7.4. *Typical performance results for EDF-FR+ in a GBAR rt-VBR only traffic scenario (P_E = 0.01).*

v (km/hr)	N_{rt-VBR}	U	CTD_{rt-VBR} (ms)	CLP_{rt-VBR}
1	5	0.71	7.79	$O(1 \times 10^{-4})$
1	6	0.84	8.55	$O(1 \times 10^{-3})$
1	7	0.98	11.87	$O(7 \times 10^{-3})$
2	5	0.71	7.78	$O(1 \times 10^{-4})$
2	6	0.84	10.00	$O(1 \times 10^{-3})$
2	7	0.95	15.00	$O(7 \times 10^{-3})$
10	5	0.71	7.78	$O(6 \times 10^{-6})$
10	6	0.84	9.98	$O(6 \times 10^{-4})$
10	7	0.95	15.51	$O(6 \times 10^{-3})$

7.6 Chapter Summary

For multiplexing bursty multimedia traffic (which are often difficult to model) over WATM air interface, traffic priority-based dynamic cell scheduling schemes are viable to achieve high channel utilization while providing predictive QoS to different traffic streams.

A set of centralized burst-level cell scheduling schemes: FCFS-FR (First Come First Served with Frame Reservation), FCFR-FR+, EDF-FR (Earliest Deadline First with Frame Reservation), EDF-FR+ and MTDR (Multitraffic Dynamic Reservation) have been investigated for the transmission of multiservice traffic over TDMA/TDD channels in broadband mobile (e.g., WATM) networks. In these schemes, the number of slots allocated to a VC (Virtual Circuit) is changed dynamically, depending on the traffic type, system traffic load, TOA/TOE value of the data burst and data burst length. The performances of the schemes have been evaluated by computer simulation using realistic voice, video and data traffic models and their QoS requirements in a wireless mobile multimedia network.

Among the burst-level cell scheduling schemes described here, the FCFS-FR,

Table 7.5. *Typical performance results for EDF-FR+ in a CBR and GBAR rt-VBR traffic scenario (v = 1 km/hr. R_{rt-VBR} = 4. R_{CBR} = 0).*

P_E	N_{CBR}	N_{rt-VBR}	U	CTD_{CBR} (ms)	CTD_{rt-VBR} (ms)	CLP_{CBR}	CLP_{rt-VBR}
0.01	30	5	0.75	5.9	5.0	$O(8 \times 10^{-3})$	$O(8 \times 10^{-4})$
0.01	40	5	0.76	6.0	8.4	$O(1 \times 10^{-2})$	$O(2 \times 10^{-3})$
0.001	30	5	0.75	5.8	4.7	$O(6 \times 10^{-3})$	$O(5 \times 10^{-4})$
0.001	40	5	0.76	5.9	8.2	$O(8 \times 10^{-3})$	$O(1 \times 10^{-3})$

FCFS-FR+ and EDF-FR schemes for TDMA/TDD-based wireless access are not suitable to provide tolerable QoS in a multitraffic environment while achieving high channel utilization. EDF-FR+ and MTDR schemes provide better cell multiplexing performance in different traffic scenarios. Such schemes are easy to implement and can result in an energy efficient TDMA/TDD medium access control protocol for broadband wireless access.

Simulation results show that the EDF-FR+ scheme provides better performance than the MTDR scheme for properly chosen reservation parameters. In addition, establishing synchronization at the BS/AP is expected to be relatively simpler for the EDF-FR+ scheme since, during each frame-time, it is more likely that fewer number of mobiles are involved in transmission/reception. In both these schemes, the reservation parameters, namely, R_{CBR}, R_{rt-VBR} and $R_{nrt-VBR}$, are the tuning knobs that can be dynamically adjusted to vary the QoS required for the different traffic streams.

Simulation results show that, for a TDMA/TDD frame size of 0.66 ms and link speed of 20 $Mbps$, reasonably good QoS (in terms of long-term cell loss rate and cell transfer delay) can be obtained for bursty voice, GBAR video and LRD data traffic even with a simple binary adaptation policy. Improved performance may be achieved by employing some measurement-based scheme to adjust the values of R_{CBR}, R_{rt-VBR} and $R_{nrt-VBR}$, according to the instantaneous traffic load but at the expense of more processing overhead. It is also observed that providing bounded delay to LRD traffic while simultaneously achieving very high channel utilization may not be possible.

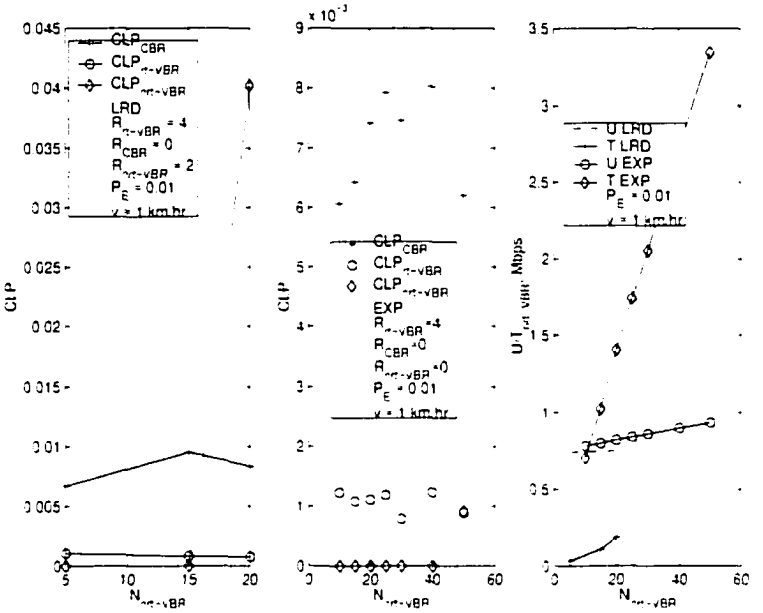

Figure 7.14. *Performance of EDF-FR+ in an integrated traffic scenario for LRD and non-LRD nrt-VBR (shown as EXP) traffic (for P_E = 0.01, v = 1 km/hr).*

Higher channel utilization is achieved for short-range dependent data traffic.

Burst-level cell scheduling scheme, such as EDF-FR+, can be easily adapted as a MAC policy for wireless access to the emerging DS Internet by integrating it with some traffic conditioning schemes at the mobile end. The BS/AP, functioning as an ingress node to the wireline network, can integrate the traffic admission policy with the wireless MAC protocol. The BS/AP scheduler can also use ECN (Explicit Congestion Notification) [85] information from downstream routers for scheduling data bursts at the mobile stations. Traffic conditioning at the network edge (i.e., in mobile stations) along with a centralized scheduling scheme can provide predictable QoS to multimedia traffic in a DS wireless network.

Chapter 8

Conclusions

8.1 Contributions of the Dissertation

Several of the RLC/MAC-layer and transport-layer protocol design problems in wireless packet based networking have been addressed in this dissertation. These include retransmission control in a multichannel distributed MAC framework, dynamic rate adaptation and error control at the link level, inter-layer protocol dependencies, service fairness (for both uplink and downlink transmission) and service integration at the air interface. The contributions of this dissertation are summarized below.

Multichannel wireless access protocols are envisioned to provide high data rate services in the next generation tetherless data networks. A retransmission control scheme for a multichannel S-ALOHA protocol in micro-cellular wireless environment has been proposed and analyzed. With properly chosen protocol parameters, the proposed scheme can provide high system capacity for a very wide range of system load. The delay and throughput performances have been evaluated by exact Markov analyses in the case of error-free and *i.i.d.* (independent and identically distributed) Rayleigh fading channels. A capture model has been incorporated in the analysis. The performance results for the single channel systems with constant probability retransmission control can be found from the general model as special cases. The effects of the different protocol parameters on system capacity have been investigated. For correlated Rayleigh fading channels, computer simulation has been used for performance evaluation. The performance of the retransmission control scheme based on local adaptation has also been studied.

Optimal dynamic rate allocation among mobile stations for variable rate packet

data transmission in a cellular wireless network is an NP-complete problem; therefore, sub-optimal solutions to this problem are sought for. Again, interference calculation is non-trivial in the case of a packet-switched cellular CDMA network under variable rate transmission with heterogeneous traffic load in different cells. A sub-optimal two-step dynamic rate selection procedure has been proposed for uplink packet data transmission in cellular VSF (Variable Spreading Factor) WCDMA systems. A novel 'mean-sense' approach for inter-cell interference calculation has been employed assuming homogeneous traffic load in the different cells. Two different error control alternatives for this variable rate packet transmission environment have been presented and their performances have been analyzed for three different channel models. The performance of the proposed two-step rate selection procedure is fairly close to that of the optimal rate allocation based on exhaustive search. A general Markov model for performance evaluation of the proposed dynamic rate allocation procedure (in the ideal case) has been presented. In addition, a BS-assisted distributed medium access control scheme employing dynamic rate adaptation has been proposed and the performance of the scheme has been evaluated.

Physical layer mechanisms to enhance wireless channel reliability can impact the performance of higher layer protocol techniques in a non-trivial manner. The performance implications of retransmission diversity packet combining on link-layer and transport-layer protocol performance have been investigated for three different heuristic-based RLC/MAC-layer access control schemes in a CDMA S-ALOHA network under frequency selective Rayleigh fading. The transport-layer protocol here implements a two-level error recovery mechanism for reliable data transmission. Two different transport-layer timer control mechanisms have been considered. Performance evaluation has been carried out for these access control schemes also for single-level error recovery in the case of moderately delay and loss sensitive data traffic. In addition, the impacts of some physical layer parameters on system performance have been discussed. All these results enable us to get insight into the identification of proper higher layer protocol mechanisms and physical layer design choices in a CDMA-based wireless packet data networking scenario.

A unified TDMA-based centralized wireless access scheme has been proposed for performing the statistical multiplexing of bursty data sources in a wireless packet

data network. This scheme combines dynamic bandwidth allocation with admission control and packet conditioning (at the mobile stations) to provide fair bandwidth distribution among bursty data flows with different profile rates (or subscription levels) in an error-prone environment.

The dynamic bandwidth allocation policy is *credit*-based and both the burst-level and the packet-level bandwidth allocations are considered. The performance of the scheme has been evaluated using computer simulations for different total subscription levels, for different compositions of flows with different profile rates and for different channel quality with different channel-error correlation pattern. The simulation results have shown that the throughput variability among flows with the same level of subscription is considerably small except for LRD (Long Range Dependent) flows with very high traffic burstiness. The relative throughput fairness among flows with different profile rates can also be achieved. The *post facto* loss and delay values (i.e., observed average packet delay and average packet loss values) for the flows depend on the corresponding delay tolerance limits of the data bursts, TDMA frame-length and the wireless link utilization level.

The energy-efficiency of the wireless access scheme has been evaluated in terms of the average *transmitter usage time* and the average *receiver usage time* in the mobile stations for both the burst-level and the packet-level bandwidth allocation. The proposed scheme can be used in an adaptive QoS framework for dynamically adjusting the QoS for flows in order to accommodate wireless channel errors and user mobility.

The performance evaluation of a number of TCP level packet scheduling policies has been carried out in terms of channel utilization and throughput fairness (among similar competing TCP connections) in a wide area wireless IP network. A semi-reliable RLC/MAC layer has been assumed underneath the transport layer. A time-frame-based channel-state aware transport-level packet scheduling scheme has been proposed that imitates wireless fair-queueing in long time scale. A new metric, namely, the *Throuput-Fairness Product (TFP)* has been introduced as a performance measure of the different packet scheduling policies. The impacts of different protocol parameters (e.g., TCP parameters) and network characteristics (e.g., wired-IP network delay, error/loss rate, error-correlation pattern) on wireless channel utilization

and TCP throughput fairness under different TCP-level packet scheduling schemes have been investigated by simulating the system dynamics under multiple concurrent TCP connections.

One important revelation from this work is that mechanisms for dynamic adaptation of TCP parameters (e.g., maximum transmission window size) based on the network delay and error characteristics would be required to achieve a high TCP throughput in wide area wireless IP networks. These mechanisms would enforce two-level control in the TCP congestion control technique.

A set of centralized burst-level bandwidth allocation schemes, namely, *FCFS-FR* (First Come First Served with Frame Reservation), *FCFR-FR+*, *EDF-FR* (Earliest Deadline First with Frame Reservation), *EDF-FR+* and *MTDR* (Multitraffic Dynamic Reservation), have been investigated for the transmission of multiservice traffic over TDMA/TDD channels in broadband wireless (e.g., WATM) networks. In these schemes, the number of slots allocated to a VC (or flow) is changed dynamically depending on the traffic type, system traffic load, TOA (Time of Arrival)/TOE (Time of Expiry) value of the data burst and data burst length. The performances of the schemes have been evaluated through computer simulation using realistic voice, video and data traffic models and their QoS requirements in a wireless mobile multimedia network.

Simulation results show that, the EDF-FR+ and MTDR schemes outperform the other schemes and can provide high channel utilization with predictive QoS guarantee in a multiservice traffic environment even in the presence of bursty channel errors. EDF-FR+ is found to provide better cell multiplexing performance than MTDR scheme. Such a scheme is easy to implement and can result in an energy efficient TDMA/TDD medium access control protocol for broadband wireless access. Burst-level cell scheduling schemes, such as EDF-FR+, can be easily adapted as MAC protocols in the emerging DS (Differentiated Services) enhanced wireless IP networks.

8.2 Future Work and Work in Progress

The introduction of WCDMA in the packet-switched wireless networking scenario gives rise to many interesting and challenging research problems. A brief description

of some of these issues follows.

- **Integrated Rate Control, Error Control and Access Control in Multidimensional Multi-Code WCDMA Systems**

  In a DS-CDMA system, multidimensional multi-code (MDMC) signaling can offer some advantages over conventional multi-code transmission and single-code transmission with variable spreading factor (VSF). In this case, transmission is based on multiple orthogonal parallel channels and the effects of self-interference among the parallel channels (which become more severe as the number of channels increases, especially in multipath fading channels) are minimized. In addition, a constant envelope signal is generated using precoding channels.

  The performance of MDMC signaling for dynamic rate adapation for uplink data transmission in cellular DS/CDMA systems needs to be evaluated. The concept of single-cell rate adaptation may be extended to joint multi-cell rate adaptation, which would presumably, offer better performance than single-cell rate adaptation. The effects of 'cell-breathing' on link-level throughput variations under dynamic rate adaptation need to be investigated.

- **Call Admission Control Protocol Design for Multirate WCDMA Networks**

  In order to use the radio resources efficiently while providing QoS to the packet data service users, efficient and robust admission control algorithms need to be devised for cellular WCDMA networks which have intrinsic support for variable rate transmission at the link layer. This again requires a realistic modeling of 'packet error process' in cellular WCDMA networks, which would be used subsequently in the call admission and call rerouting through BS selection.

- **Exploiting Channel Artifact due to Fading and Interference for TCP Performance Enhancement in WCDMA-Based Cellular Wireless IP Networks**

  TCP performance in cellular WCDMA environment is one of the most critical issues, especially in the presence of highly variable interference resulting from bursty packet data traffic. In this regard, investigating mechanisms that are able to exploit the channel status information in the TCP congestion control method for performance enhancement would be worthwhile. TCP performance

needs to be assessed in such an environment, in the presence of microdiversity and macrodiversity, while taking user mobility also into consideration.

In addition, variable rate transmission at the link level may significantly affect the end-to-end performance (i.e., TCP performance) resulting from the increased round trip delay variability. Efficient interference management mechanisms need to be developed to enhance TCP performance in such scenarios.

- **Call Control Signaling and Mobility Management in Multimedia Wireless IP Networks**

  For successful migration of multimedia services in the wireless IP domain, call control and mobility management issues (in the control plane) need to be resolved. In this regard, ITU-T H.323 and IETF SIP (Session Initiation Protocol) are the two competing multimedia call control protocol standards that are being envisioned to be used in wireless IP networks ([102]-[104]). The relative performance evaluation of these two schemes in a WCDMA-based cellular IP scenario and the development of efficient mobility management schemes for H.323/SIP calls would be of much interest.

Our current works are addressing several of these problems.

Bibliography

[1] Kaveh Pahlavan and Allen H. Levesque. *Wireless Information Networks*. John Wiley & Sons. Inc.. 1995

[2] M. Zeng. A. Annamalai and V. K. Bhargava. "Recent advances in cellular wireless communications." *IEEE Communications Magazine*. vol. 37. no. 9. Sept. 1999. pp. 128-138.

[3] Mun Choon Chan and Thomas Y.C. Woo. "Next-generation wireless data services." *IEEE Personal Communications*. vol. 6. no.1. Feb. 1999. pp. 20-33.

[4] R. V. Nobelen. N. Seshadri. J. Whitehead and S. Timiri. "An adaptive radio link protocol with enhanced data rates for GSM evolution." *IEEE Personal Communications*. Feb.. 1999. pp. 54-64.

[5] G. Brasche and B. Walke. "Concepts. services and protocols of the new GSM phase 2+ general packet radio service." *IEEE Communications Magazine*. Aug. 1997. pp. 94-104.

[6] Raj Ayala. Kalyan Basu and Steve Elliot. "Internet technology based infrastructure for mobile multimedia services." *Proc. IEEE WCNC'99*. pp. 109-113.

[7] Leonard J. Cimini. Jr.. Justin C.-I. Chuang and Nelson R. Sollenberger. "Advanced cellular Internet service (ACIS)." *IEEE Communications Magazine*. Oct. 1998. pp. 150-159.

[8] Li Fung Chang. Vijay Varma and Ray Cheng. "Architecture alternatives for PCS-to-Internet protocol interworking." *Proc. IEEE WCNC'99*. pp. 766-770.

[9] Roman Pichna. Tero Ojanperä. Harri Posti and Jouni Karppinen. "Wireless Internet - IMT-2000/wireless LAN interworking." *Journal of Communications and Networks*. vol. 2. no. 1. Mar. 2000. pp. 46-57.

[10] Norman Abramson. "Wideband random access for the last mile." *IEEE Personal Communications*. Dec. 1996. pp. 29-35.

[11] M. A. Marsan and D. Roffinela. "Multichannel local area networks protocols." *IEEE Journal on Selected Areas in Communications*. vol. 1. no. 5. Nov. 1983. pp. 885-897.

[12] M. S. Alam. "Retransmission probability selection of multichannel CSMA/CD protocols for LANs". *Computer Communications*. vol. 15. no. 10. Dec. 1992. pp. 621-629.

[13] D. Raychaudhuri and K. Joseph. "Performance evaluation of slotted ALOHA with generalized retransmission backoff." *IEEE Transactions on Communications*. vol. 38. no. 1. Jan. 1990. pp. 117-122.

[14] D. G. Jeong and W. S. Jeon. "Performance of an exponential backoff scheme for slotted ALOHA protocol in local wireless environment." *IEEE Transactions on Vehicular Technology*. vol. 44. no. 3. Aug. 1995. pp. 470-479.

[15] W. S. Jeon and D. G. Jeong. "Contention-based reservation protocol for WDM local lightwave networks with nonuniform traffic pattern." *IEICE Transactions on Communications*. vol. E82-B. no. 3. Mar. 1999. pp. 521-531.

[16] I. E. Pountourakis and E. D. Sykas. "Analysis, stability and optimization of Aloha-type protocols for multichannel networks." *Computer Communications*. vol. 15. no. 10. Dec. 1992.

[17] W. Xu. A. Chockalingam and L. Milstein. "Performance analysis of multichannel wireless access protocol in the presence of bursty packet losses." *Proc. IEEE MILCOM'98*. Bedford. MA. Oct. 1998.

[18] Chiew T. Lau and Cyril Leung. "Capture models for mobile packet radio networks." *IEEE Transactions on Communications*. vol. 40. no. 5. May 1992. pp. 917-925.

[19] M. Zorzi. R. R. Rao and L. B. Milstein. "On the accuracy of a first-order Markov model for data transmissions on fading channels." *Proc. IEEE ICUPC'95*. Nov. 1995. pp. 211-215.

[20] A. Chockalingam. Michele Zorzi. L. B. Milstein and P. Venkataram. "Performance of a wireless access protocol on correlated Rayleigh fading channels with capture." *IEEE Transactions on Communications*. vol. 46. no. 5. May 1998. pp. 644-655.

[21] M. Zorzi and R. R. Rao. "On channel modeling for delay analysis of packet communications over wireless links." *36th Annual Allerton Conference*. Allerton House. Monticello. IL. Sept. 1998. URL: http://aloha.ucsd.edu/~zorzi/publ.html

[22] E. Dahlman et al.. "UMTS/IMT-2000 based on wideband CDMA." *IEEE Communications Magazine*. vol. 36. pp. 70-80. Sept. 1998.

[23] M. Oğuz Sunay and Jussi Kähtävä. "Provision of variable data rates in third generation wideband DS-CDMA systems." *Proc. IEEE WCNC'99*. pp. 505-509.

[24] O. Sallent and R. Augustí. "A proposal for an adaptive S-ALOHA access system for a mobile CDMA environment." *IEEE Transactions on Vehicular Technology*. vol. 47. no. 3. Aug. 1998. pp. 977-985.

[25] C.-L. I and K. K. Sabnani. "Variable spreading gain with adaptive control for true packet switching wireless network." *Proc. IEEE ICC'95*. pp. 725-730.

[26] Alfred V. Aho and Jeffrey D. Ullman. *Foundations of Computer Science (C Edition)*. Computer Science Press. 1998.

[27] C.-L. I and R. D. Gitlin. "Multi-code CDMA wireless personal communications networks." *Proc. IEEE ICC'95*. pp. 1060-1063.

[28] E. Dahlman and K. Jamal. "Wide-band services in a DS-CDMA based FPLMTS system." *Proc. IEEE VTC'96*. pp. 1656-1660.

[29] W. W. Peterson and E. J. Weldon. *Error Correcting Codes*. Cambridge, MA: M.I.T. Press. 1972.

[30] M. B. Pursley. "Performance evaluation for phase-coded spread-spectrum multiple-access communication - Part I: system analysis." *IEEE Transactions on Communications*. vol. 25. Aug. 1977. pp. 795-799.

[31] J. G. Proakis. *Digital Communications*. 2nd ed. New York: McGraw-Hill. 1989.

[32] E. A. Geraniotis. "Throughput and packet error probability tradeoffs in cellular direct-sequence and hybrid spread-spectrum networks." *Proc. IEEE ICC'87*. pp. 40.4.1-40.4.6.

[33] Rec. ITU-R M.1225. *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000*. 1997.

[34] A. J. Viterbi. A. M. Viterbi and E. Zehavi. "Other-cell interference in cellular power-controlled CDMA." *IEEE Transactions on Communications*. vol. 42. Feb./Mar./Apr. 1994. pp. 1501-1504.

[35] Erik Dahlman. Per Beming. Jens Knutsson. Fredrik Ovesjö. Magnus Persson and Christiaan Robol. "WCDMA - the radio interface for future mobile multimedia communications." *IEEE Transactions on Vehicular Technology*. vol. 47. no. 4. Nov. 1998. pp. 1105-1118.

[36] Antun Samukic. "UMTS universal mobile telecommunications system: developments of standards for the third generation." *IEEE Transactions on Vehicular Technology*. vol. 47. no. 4. Nov. 1998. pp. 1099-1104.

[37] Peter W. de Graaf and J. S. Lehnert. "Performance comparison of a slotted ALOHA DS/SSMA network and a multichannel narrowband slotted ALOHA network." *IEEE Transactions on Communications*. vol. 46. no. 4. Apr. 1998. pp. 544-552.

[38] M. Saito. H. Okada. T. Sato. T. Yamazato. M. Katayama and A. Ogawa. "Throughput improvement of CDMA ALOHA systems." *IEICE Transactions on Communications*. vol. E80-B. no. 1. Jan. 1997.

[39] D. Chase. "Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets." *IEEE Transactions on Communications*. vol. 33. 1985. pp. 385-393.

[40] S. Kallel. "Analysis of a type II hybrid ARQ scheme with code combining." *IEEE Transactions on Communications*. vol. 38. 1990. pp. 1133-1137.

[41] S. Souissi and S. B. Wicker. "A diversity combining DS/CDMA system with convolutional encoding and Viterbi decoding." *IEEE Transactions on Vehicular Technology*. vol. 44. May 1995. pp. 304-312.

[42] Amir M. Y. Bigloo. T. Aaron Gulliver and Vijay K. Bhargava. "A slotted frequency-hopped multiple-access network with packet combining." *IEEE Journal on Selected Areas in Communications*. vol. 14. no. 9. Dec. 1996. pp. 1859-1865.

[43] A. Annamalai and Vijay K. Bhargava. "Adaptive retransmission diversity with packet combining for slotted DS/CDMA packet radio networks." *Wireless Personal Communications*. vol. 11. no. 3. Dec. 1999. pp. 269-291.

[44] M.C. Chuah. B. T. Doishi. S. Dravida. R. P. Ejzak and S. Nanda. "Performance comparisons of two retransmission protocols for CDMA." *Proc. IEEE VTC'96*. pp. 272-276.

[45] Mooi Choo Chuah. On-Ching Yue and Antonio DeSimone."Performance of two TCP implementations in mobile computing environments." *Proc. IEEE GLOBECOM'95*. pp. 339-344.

[46] Lazaros Merakos and Dimitri Kazakos. "On retransmission control policies in multiple-access communication networks." *IEEE Transactions on Automatic Control*. vol. AC-30. no. 2. Feb. 1985.

[47] Ronald L. Rivest. "Network control by Bayesian broadcast." *IEEE Transactions on Information Theory*. vol. IT-33. . pp. 323-327. May 1987.

[48] S. Haykin. *Digital Communications*. John Wiley & Sons. Inc.. 1988.

[49] W. Stallings. *High-Speed Networks: TCP/IP and ATM Design Principles*. Prentice Hall. Inc.. 1998.

[50] S. Lin and D. J. Costello. *Error Control Coding: Fundamentals and Applications*. Englewood Cliffs. NJ: Prentice-Hall. 1983.

[51] Sudhir Ramakrishna and Jack M. Holtzman. "Interaction of TCP and data access control in an integrated voice/data CDMA system." *ACM/Baltzer Mobile Networks and Appls.*. vol. 3. 1998. pp. 409-417.

[52] Mohsen Soroushnejad and Evaggelos Geraniotis. "Multi-access strategies for an integrated voice/data CDMA packet radio network." *IEEE Transactions on Communications*. vol. 43. no. 2/3/4. Feb./Mar./Apr. 1995. pp. 934-945.

[53] S. Blake. D. Black. M. Carlson. E. Davies. Z. Wang and W. Weiss. "An architecture for differentiated services." RFC 2475. Dec. 1998.

[54] Indu Mahadevan and Krishna M. Sivalingam. "Quality of service Architectures for wireless networks: intserv and diffserv models." *Proc. International*

Workshop on Mobile Computing. ISPAN'99. Perth. Australia. June 1999. URL: http://www.eecs.wsu.edu/~imahadev/publications.html

[55] Indu Mahadevan and Krishna M. Sivalingam. "Quality of service in wireless networks using enhanced differentiated services approach." *Proc. IEEE ICCCN'99.* Boston. Oct. 1999. URL: http://www.eecs.wsu.edu/ imahadev/publications.html

[56] Thyagarajan Nandagopal. Songwu Lu and Vaduvur Bharghavan. "A unified architecture for the design and evaluation of wireless fair queueing algorithms." *Proc. ACM/IEEE MobiCom'99.* pp. 132-142.

[57] S. Floyd and V. Jacobson. "Link-sharing and resource management models for packet networks." *IEEE/ACM Transactions on Networking.* vol. 3. no. 4. Aug. 1995.

[58] C. Fragouli. V. Sivaraman. M. B. Srivastava. "Controlled multimedia wireless link sharing via enhanced class-based queuing with channel-state-dependent packet scheduling." *Proc. IEEE INFOCOM'98.* pp. 572-580. URL: http://www.cs.ucla.edu/~vijay/

[59] P. Bhagwat. P. Bhattacharya. A. Krishna and S. Tripathi. "Enhancing throughput over wireless LANs using channel state dependent packet scheduling." *Proc. IEEE INFOCOM'96.* vol. 3. pp. 1133-1140. URL: http://www.cs.umd.edu/~pravin/publications/publist.htm

[60] Songwu Lu. Vaduvur Bharghavan and R. Srikant. "Fair scheduling in wireless packet networks." *IEEE/ACM Transactions on Networking.* vol. 7. no. 4. Aug. 1999. pp. 473-489.

[61] T. S. Eugene Ng. Ion Stoica and Hui Zhang. "Packet fair queueing algorithms for wireless networks with location-dependent errors." *Proc. IEEE INFOCOM'98.* URL: http://www-cgi.cs.cmu.edu/afs/cs.cmu.edu/user/hzhang/WWW/HuiZhang.html

[62] Parameswaran Ramanathan and Prathima Agrawal. "Adapting packet fair queueing algorithms to wireless networks." *Proc. of ACM/IEEE MobiCom'98*

[63] D. Raychaudhuri and N. D. Wilson. "ATM-based transport architecture for multiservices wireless personal communication networks." *IEEE Journal on Selected Areas in Communications.* vol. 12. no. 8. Oct. 1992. pp. 1401-1414.

[64] M. J. Karol. Z. Liu and K. Y. Eng. "An efficient demand-assignment multiple access protocol for wireless packet (ATM) networks." *ACM/Baltzer Wireless Networks.* vol. 3. Dec. 1995. pp. 269-279.

[65] N. Passas et al.. "MAC protocol and traffic scheduling for wireless ATM networks." *ACM/Baltzer Mobile Networks and Appls..* vol. 3. no. 3. Sept. 1998. pp. 275-291.

[66] Jyh-Cheng Chen. Krishna M. Sivalingam and Prathima Agrawal. "Performance

comparison of battery power consumption in wireless multiple access protocols." *ACM/Baltzer Wireless Networks*. 5 (1999). pp. 445-460.

[67] D. P. Heyman. "The GBAR source model for VBR videoconferences." *IEEE/ACM Transactions on Networking*. vol. 5. no. 4. Aug. 1997. pp. 554-560.

[68] Kihong Park. Gitae Kim and Mark Crovella. "On the relationship between file sizes. transport protocols and self-similar network traffic." *Proc. Int. Cnf. on Network Protocols*. pp. 171-180. Oct. 1996. URL: http://www.cs.purdue.edu/homes/park/

[69] Christian Wietfeld and Ulrich Gremmelmaier. "Seamless IP-based service integration across fixed/mobile and corporate/public networks." *Proc. IEEE VTC'99 (Spring)*. pp. 1930-1934.

[70] Thomas F. La Porta. Luca Salgarelli and Gerard T. Foster. "Mobile IP and wide area wireless data." *Proc. IEEE WCNC'99*. pp. 1528-1532.

[71] Stefen Savage. Neal Cardwell. David Wetherall and Tom Anderson. "TCP congestion control with a misbehaving receiver." *ACM SIGCOMM Computer Communication Review*. vol. 29. no. 5. Oct. 1999. pp. 71-77.

[72] Anurag Kumar. "Comparative performance analysis of versions of TCP in a local network with a lossy link." *IEEE Transactions on Networking*. vol. 6. no. 4. Aug. 1998. pp. 485-498.

[73] Sally Floyd and Kelvin Fall. "Promoting the use of end-to-end congestion control in the Internet." *IEEE/ACM Transactions on Networking*. May 1993. URL: http://www-nrg.ee.lbl.gov/floyd

[74] R. Cáceres and Liviu Iftode. "Improving the performance of reliable transport protocols in mobile computing environments." *IEEE Journal on Selected Areas in Communications*. vol. 13. no. 5. June 1995. pp. 850-857.

[75] A. Bakre and B. R. Badrinath. "I-TCP: indirect TCP for mobile hosts." *Proc. 15th IEEE Int. Conf. Distributed Computing Systems (ICDCS)*. May 1995. pp. 136-143.

[76] M. Mathis. J. Mahdavi. S. Floyd and A. Romanow. "TCP selective acknowledgement options." *Internet RFC 2018*. 1996.

[77] H. Balakrishnan. S. Seshan and R. Katz. "A comparison of mechanisms for improving TCP performance over wireless links." *IEEE/ACM Transactions on Networking*. vol. 5. Dec. 1997. pp. 756-769.

[78] H. Balakrishnan and R. Katz. "Explicit loss notification and wireless web performance". *Proc. IEEE GLOBECOM'98. Internet Mini Conference*. Sydney. Australia.

[79] A. DeSimone. M. C. Chuah and O. -C. Yue. "Throughput performance of

transport-layer protocols over wireless LANs." *Proc. IEEE GLOBECOM'93*, pp. 542-549.

[80] Michele Zorzi and Ramesh R. Rao. "The effect of correlated errors on the performance of TCP." *IEEE Communications Letters*, vol. 1, Sept. 1997, pp. 127-129.

[81] A. Chockalingam and Gang Bao. "Performance of TCP/RLP protocol stack on correlated fading DS-CDMA wireless links." *Proc. IEEE VTC'98*, pp. 363-367.

[82] Yong Bai, Gang Wu and Andy T. Ogielski. "TCP/RLP coordination and interprotocol signaling for wireless Internet." *Proc. IEEE VTC'99*, pp. 1945-1951.

[83] Jackson W. K. Wong and Victor C. M. Leung. "Improving end-to-end performance of TCP using link-layer retransmissions over mobile internetworks." *Proc. IEEE ICC'99*, pp. 324-328.

[84] Christina Parsa and J. J. Garcia-Luna-Aceves. "Improving TCP performance over wireless networks at the link layer." to appear in *ACM/Baltzar Mobile Networks and Applications*. URL: http://www.cse.ucsc.edu/research/ccrg/publications.html

[85] S. Kalyanaraman, D. Harrison, S. Arora, K. Wanglee and G. Guarriello. "A one-bit feedback enhanced differentiated services architecture." Internet Draft - work in progress. http://www.ietf.org/internet-drafts/draft-shivkuma-ecn-diffserv-01.txt

[86] S. Floyd and V. Jacobson. "Random early detection gateways for congestion avoidance." *IEEE/ACM Transactions on Networking*, 3(3), Sept. 1993, pp. 115-156.

[87] T. R. Henderson, E. Sahouria, S. McCanne and R. H. Katz. "On improving the fairness of TCP congestion avoidance." *Proc. IEEE GLOBECOM'98*, pp. 539-544.

[88] A. Giessler, J. Haanle, A. Konig and E. Pade. "Free buffer allocation by simulation." *Computer Networks*, vol. 1, no. 3, July 1978, pp. 191-204.

[89] W. Honcharenko, Jan P. Kruys, David Y. Lee and Nitin J. Shah. "Broadband wireless access." *IEEE Communications Magazine*, Jan. 1997, pp. 20-26.

[90] D. Raychaudhuri. "Wireless ATM: an enabling technology for multimedia personal communication." *ACM/Baltzer Wireless Networks*, vol. 2, 1996, pp. 163-171.

[91] P. Agrawal et al., "SWAN: a mobile multimedia wireless network." *IEEE Personal Communications Magazine*, Apr. 1996, pp. 18-33.

[92] K. Ravindran and Ralf P. Steinmetz. "Object-oriented communication structures for multimedia data transport." *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 7, Sept. 1996, pp. 1360-1375.

[93] Yin Bao. "Quality of service control for real-time multimedia applications in packet switched networks." Ph.D. dissertation, submitted to the Department of

Computer and Information Sciences. University of Delaware. May 1998. URL: http://www.eecis.udel.edu/~bao/phdthesis.html

[94] Jim Kurose. "Open issues and challenges in providing quality of service guarantees in high speed networks." *ACM SIGCOMM Computer Communication Review.* 23(1). Jan. 1993. pp. 6-15.

[95] S. Jamin. P. B. Danzig. S. J. Shenker and L. Zhang. "A measurement-based admission control algorithm for integrated services packet networks." *IEEE/ACM Transactions on Networking.* Feb. 1997. URL: http://netweb.usc.edu/jamin/admctl/ton96.ps.Z

[96] J. -C. Chen. K. M. Sivalingam. Prathima Agrawal and R. Acharya. "Scheduling multimedia services in a low-power MAC for wireless and mobile ATM networks." *IEEE Transactions on Multimedia.* vol. 1. no. 2. June 1999. pp. 187-201.

[97] A. Acampora. "Wireless ATM: a perspective on issues and prospects." *IEEE Personal Communications Magazine.* Aug. 1996. pp. 8-17.

[98] R. Guerin and V. Peris. "Quality-of-service in packet networks: basic mechanisms and directions." *Computer Networks (Special Issue on Internet Telephony).* vol. 31. no. 3. Feb. 1999. pp. 169-189.

[99] Vaduvur Bharghavan. Songwu Lu and Thyagarajan Nandagopal. "Fair queuing in wireless networks: issues and approaches." *IEEE Personal Communications.* Feb. 1999. pp. 44-53.

[100] J. M. Peha and F. A. Tobagi. "A cost-based scheduling algorithm to support integrated services." *Proc. IEEE INFOCOM'91.* pp. 741-753.

[101] S. Nanda. David J. Goodman and Uzi Timor. "Performance of PRMA: a packet voice protocol for cellular systems." *IEEE Transactions on Vehicular Technology.* vol. 40. no. 3. Aug. 1991. pp. 584-599.

[102] Wanjiun Liao. "Mobile Internet telephony: mobile extensions to H.323." *Proc. IEEE INFOCOM'99.* New York. Mar. 1999.

[103] Wanjiun Liao. "VoIP mobility in IP/cellular network interworking." *IEEE Communications Magazine.* Apr. 2000. pp.70-75.

[104] Elin Wedlund and Henning Schulzrinne. "Mobility support using SIP." *Proc. 2nd Int. Worksshop on Wireless Mobile Multimedia (WOWMOM).* Aug. 1999. Seattle. Washington. USA. pp. 76-82.

[105] K. Pahlavan. P. Krishnamurthy. A. Hatami. M. Ylianttila. J.-P. Makela. R. Pichna and J. Vallström. "Handoff in hybrid mobile data networks." *IEEE Personal Communications.* vol. 7. no. 2. Apr. 2000. pp. 34-47.

Appendix A

Evaluation of $p_{(l,M,L)}$

The probability of successful transmisssion in a channel for L (> 0) linearly spaced power levels in the receiver when l stations contend in M channels is:

$$
\begin{aligned}
p_{(l,M,L)} &= \frac{1}{L} \cdot \sum_{i=1}^{L} \sum_{j=0}^{l-1} \binom{l-1}{j} \left(\frac{1}{M}\right)^{j} \cdot \left(1 - \frac{1}{M}\right)^{l-1-j} \cdot \left(\frac{i-1}{L}\right)^{j} \\
&= \frac{1}{L} \cdot \sum_{i=1}^{L} \left(\frac{i-1}{ML} + 1 - \frac{1}{M}\right)^{l-1} \\
&= \frac{1}{L} \cdot \sum_{i=1}^{L} \left[1 + \frac{i-(1+L)}{ML}\right]^{l-1}
\end{aligned}
$$

Appendix B

Evaluation of α in (3.8)

Suppose that a cell is loaded with the effective number of users, x_i $(i = 0, 1, \ldots)$, where x_0 is associated with the (desired) center cell and $\{x_i\}$ $(i \geq 1)$ with other neighboring cells. In cellular power controlled-CDMA [34], the inter-cell interference is affected by path loss and shadowing, modeled as a multiplicative loss of ξ_i, which yields the composite of inter-cell interference as $\xi_1 x_1 + \xi_2 x_2 + \ldots$. Due to the law of large numbers, it fairly approaches an expectation, namely, $\mathbf{E}\{\xi_1 x_1 + \xi_2 x_2 + \ldots\}$, which is

$$\mathbf{E}\{\xi_1 x_1 + \xi_2 x_2 + \ldots\} \longrightarrow \eta \mathbf{E}\{x\}. \tag{B.1}$$

Here, x is found in (3.7) subject to $n = G$ due to the equal average attempts G per cell, and the frequency reuse factor η can be computed by the area-average of ξ_i in [34]. Since $\alpha = \mathbf{E}\{v_m\}/v_1$, we have

$$\alpha = \frac{\mathbf{E}\{x\}}{G} \qquad \text{subject to } n = G. \tag{B.2}$$

which is well approximated by (3.8).

For the evaluation of α, (3.8) is substituted into (3.6) with the constraint of $n = G$ $(x_0 = x)$, and then the sub-optimal rate selection is performed using the two-step search procedure, and α is determined by (3.8) using the value of x that corresponds to the sub-optimal rate selection. In this way the value of α corresponding to a value of G is obtained that reflects the effect due to the rate adaptation in other cells in an average sense.

Appendix C

Implementation of the second-step of the rate search procedure

Procedure FindBeta() {

1. $x_o = m_o n$. $x_{o-} = m_{o-} n$. $x_{o+} = m_{o+} n$. $\Delta_l = x_o - x_{o-}$. $\Delta_r = x_{o+} - x_o$. $\delta = 1$
2. If ($\beta(x_{o-}) > \beta(x_o)$) search-direction = RIGHT
   Else search-direction = LEFT
3. If (search-direction == RIGHT) {
   $\beta^* = \beta(x_o)$. $temp = \beta^*$. $i = 0$
   Do {
   $temp = \beta^*$
   $\Delta = (++i) \times \delta$
   $n_{m_{o+}} = \Delta$. $n_{m_o} = n_{m_o} - \Delta$
   $x_o = n_{m_{o+}} \times m_{o+} + n_{m_o} \times m_o$
   $\beta^* = \beta(x_0)$
   } (While ($\beta^* > temp$) and ($\Delta < \Delta_r$))
   Output $temp$. $n_{m_{o+}}$ and n_{m_o}
   }
   Else {
   $\beta^* = \beta(x_o)$. $temp = \beta^*$. $i = 0$
   Do {
   $temp = \beta^*$
   $\Delta = (++i) \times \delta$
   $n_{m_{o-}} = \Delta$. $n_{m_o} = n_{m_o} - \Delta$

$x_o = n_{m_{o-}} \times m_{o-} + n_{m_o} \times m_o$

$\mathcal{F}^* = \mathcal{F}(x_o)$

} (While ($\mathcal{F}^* > temp$) and ($\Delta < \Delta_l$))

Output $temp$. $n_{m_{o-}}$ and n_{m_o}

}

}

Appendix D

Evaluation of $\beta(i)$ in (3.20)

When n_m stations transmit in a slot (at rate r_m each), the probability that all the transmissions from s_m stations are successful is $B(n_m, s_m, P_c^m(x-m))$ which results in ms_m successful frame transmissions. Since in case of SR-based error control, there are some number of correct frames even when transmission in a timeslot as a whole fails, each of the $n_m - s_m$ users contributes δ_m number of frames to the total throughput where δ_m is given by

$$\delta_m = \sum_{l=0}^{m-1} l \binom{m}{l} P_c^l(x-m)[1-P_c(x-m)]^{m-l} = mP_c(x-m) - mP_c^m(x-m). \quad \text{(D.1)}$$

Therefore, the total number of successfully transmitted RLC/MAC-layer frames can be well approximated to $ms_m + \delta_m(n_m - s_m)$.

Using this formulation for all the rates ($m = 1, 2, \ldots, \varphi$) yields (3.20).