Quality of Service Provisioning for Multimedia Transmissions in Packet-Switched Wireless Communication Systems

by

Hui Chen

B.Sc., Nanjing University, Nanjing, China, 1999
M.Sc., Nanjing University, Nanjing, China, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August, 2010

Abstract

This thesis covers several research problems in the area of multimedia transmissions over wireless communication systems. The objective is to enhance the quality of service (QoS) provisioning for multimedia transmissions over wireless communication systems. Recent advances in signal processing techniques have enabled wireless networks to have multi-packet reception (MPR) capability at the physical layer, where it is possible to receive more than one packet when concurrent transmissions occur. In this thesis, a multi-reservation, multiple access (MRMA) scheme for wireless multimedia networks, based on an MPR channel model, is proposed to differentiate the QoS provisioning for multimedia traffic and best effort traffic on MPR-enabled channels. Another advance in recent wireless technologies is the cross-layer optimization over wireless links. In this thesis, a cross-layer QoS provisioning method is proposed. Specifically, the data frame drop ratio (FDR) in the medium access control layer and the bit error rate (BER) in the physical layer are jointly optimized to provide a better data frame loss ratio (FLR) seen by the higher layers. In this thesis, a cross-layer enhanced scheduling (CEPS) method is proposed for providing QoS optimization for multimedia traffic in multi-code code-division multiple access (MC-CDMA) systems. CEPS also provides flexible fairness provisioning among different traffic flows. Then, the cross-layer QoS FLR in the MC-CDMA systems is further optimized by accounting for not only the predesigned QoS requirements, the historical QoS experienced,

and the current buffer status, but also the statistical traffic models of different traffic flows. Results show that the system performance and the QoS provisioning can be greatly improved by the cross-layer QoS consideration and the proposed optimizations. The thesis also adopts the cross-layer QoS consideration in adaptive modulation and coding (AMC)-enabled wireless systems. A QoS-based cross-layer scheduling scheme (QoS-CLS) is proposed for AMC-based systems. It can guarantee the QoS provisioning for multimedia traffic flows and/or share the QoS among different traffic flows by making scheduling and coding/modulation decisions based on the buffer status, traffic status, channel status, and other information for different traffic flows. Results show that QoS-CLS can greatly enhance the system performance.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deep gratitude to my research supervisor Dr. Victor C.M. Leung and Dr. Henry C.B. Chan for all these years of guidance and help throughout my PhD research work. Their great efforts in teaching and research provided me with valuable support and good ideas. Without their guidance, this thesis would not be possible.

I also want to thank all the members of my supervisory committee for their comments and time, Dr. Vikram Krishnamurthy and Dr. Jane Wang.

Thanks to Dr. Fei (Richard) Yu, Dr. Min Chen, Dr. Zhanping (Walter) Yin, Dr. Qixiang (Kevin) Pang, Dr. Zhibing (Harry) Chen, Dr. Ki-Dong Lee, Dr. Syed Hussain Ali, Jie Zhang, Ying Wai (Ray) Lam, Yangwen Liang, Haoming Li, with whom I have the great opportunity to collaborate during the past years.

Special thanks to my wife Jie Zhang and my parents Yuanfang and Hongzhang for all their support during these years.

Co-Authorship Statement

I am the first author and principal contributor of all manuscript chapters. All chapters are co-authored with Dr. Henry C.B. Chan and Dr. Victor C.M. Leung, who supervise the present thesis research. Chapter 2 is also co-authored with Dr. Richard F. Yu, who helped to review the journal paper draft and provided positive feedback on the a couple of detailed design in the protocol. Chapter 3 is also co-authored with Jie Zhang, who helped in contributing to parts of the simulation modeling.

Chapter 1

Introduction

The Internet has grown dramatically in the last decade and many interesting multimedia applications have emerged in business, social, entertainment, personal management, and other fields. People's lifestyles have changed greatly and communication services are becoming especially important, since they serve as the fundamental element for accessing Internet-based multimedia applications. In such an environment, packet-switched wireless communication systems have attracted great interest from both service providers and subscribers, since the wireless technologies can allow users to reach communication services from anywhere, without needing physical attachments to immobilized equipment. While wireless technologies provide great convenience and flexibility, wireless communication systems are still faced with the challenge to provide competitive communication services over a limited amount of radio spectrum resources for users to enjoy the plethora of multimedia applications via shared transmission media. On one hand, service providers need to control the channel access to distribute the total bandwidth among users wisely and efficiently so that users will enjoy a good quality of service (QoS) when subscribing to different types of multimedia services. On the other hand, service providers also need to maximize the system utilization with limited spectrum resources, to support as many users as possible who subscribe to bandwidth-consuming multimedia applications.

This thesis focuses on multimedia transmissions over wireless links. Note that transmis-

sions over wireless links are usually considered as the bottleneck of wireless communication systems, because of the limited wireless spectrum resources. As a result, QoS provisioning on wireless links is essential for the end-to-end QoS for different applications running on top of wireless communication systems. The objective of this thesis is to provide flexible and guaranteed QoS for multimedia transmissions, while enhancing the capacity of wireless links. While numerous wireless technologies have been proposed to achieve higher data rates over wireless links, the ability to manipulate and coordinate the transmissions on the data link layer must also be considered to achieve the desired QoS, especially for bandwidth-consuming multimedia transmissions. This thesis focuses on the protocol design of wireless links and cross-layer optimizations for wireless channels to provide effective and efficient guaranteed or differentiated QoS for different transmissions. Specifically, a multiple access protocol is proposed for wireless channels with multi-packet reception (MPR) capabilities, so that the QoS of voice, video, and data transmissions can be differentiated. Also, a cross-layer QoS consideration is proposed and applied to the scheduling and the transmission parameter optimizations in code-division multiple access (CDMA) networks and to wireless channels with adaptive modulation and coding (AMC) capabilities. Thus, the QoS of different transmissions can be guaranteed or differentiated, according to their QoS requirements, while the channel utilization is optimized.

1.1 Quality of Service of Multimedia Transmissions

Traditional circuit-switched voice or best effort data transmissions require only simple coordinations and a small effort for QoS provisioning. Nevertheless, the traffic flows of Internet-based multimedia applications require more bandwidth and have complicated and diverse QoS re-

quirements. Wireless communication systems often classify transmissions into different types to provide better service. For example, in the Third Generation Partnership Project (3GPP), traffic flows for different applications are classified into four groups, according to the traffic characteristics and the QoS requirements [1]. The conversational class represents applications, such as real-time voice communications, which have very strict delay and jitter requirements. The streaming class represents applications or services like playback video, which have less stringent delay requirements but still need to have very little jitter. The interactive class represents services like web browsing, which have almost no delay requirements, though the content must be delivered error-free and the traffic has a specific request/response pattern. The background class represents best effort data services such as email and telemetry, which have QoS requirements that are similar to those of the interactive class, but without a specific traffic pattern. Some other systems, such as the Institute of Electrical and Electronics Engineers (IEEE) 802.11e wireless local area networks (WLANs) [2], also have their own definitions of traffic classes or types to enable the system to serve different transmissions with different QoS requirements more easily. In [3], International Telecommunication Union–Telecommunication standardization sector (ITU-T) proposed the end-user multimedia QoS categories. It also provided the key QoS parameters for multimedia QoS and the recommended target values for different categories as shown in Table 1.1. While the traffic classifications in different systems are based on the end-to-end QoS requirements of specific applications, QoS provisioning on wireless links for different classes of traffic flows is especially critical for wireless communication systems and has been widely studied. One of the major differences between wired and wireless communication systems is that wireless links are generally less reliable than wired links. ITU-T also provided a relationship between the end-to-end QoS and the QoS for a specific network

section (say the wireless link) in [4].

Transmissions over wireless links are coordinated by schemes involving the physical layer and the medium access control layer (MAC, a sub-layer of the data link layer). Packets arriving from the upper layer are divided and encapsulated into data frames before being transmitted over wireless links. For multimedia transmissions, packet delivery can be time-sensitive and if a data frame is missing its transmission deadline, the packet should be dropped to preserve valuable network resources. For traffic that is not delay-sensitive, even though data frames can wait for a long time before they are transmitted, new data frames may arrive continuously causing the buffer to overflow and some data frames to be dropped. The fraction of dropped data frames is referred to as the frame drop ratio (FDR), and is one of the most important QoS parameters for the MAC layer. In this thesis, the packet-drop ratio (PDR) is sometimes used for the QoS parameter on the MAC layer. It has the same definition as FDR and is more relevant to the higher layers. The average data frame delay, or the packet delay, is another important QoS parameter for the MAC layer. The effective data link layer throughput can also be used to measure the performance of the system. In any case, the QoS parameter for the physical layer is the bit error ratio (BER), or the signal-to-interference-plus-noise ratio (SINR). These parameters can be used to measure the error probability of transmitting one bit or one symbol. In this thesis, all of these QoS parameters and their relationships to different wireless links are considered, in proposing new transmission schemes to improve QoS provisioning over wireless links.

1.2 Multiple Access with Multi-Packet Receptions

In wireless communication systems, the MAC protocol allows several user terminals connected to the same wireless link to transmit over the link and share its capacity. It runs on top of the physical layer multiplexing technologies and deals with issues such as addressing, assigning multiplex channels to different users, and avoiding collisions. Conventional multiple access protocols, such as ALOHA [5] and carrier sense multiple access (CSMA) [6], are aimed to serve best effort communications such as best effort data transmissions. Thus, the focus is on effective channel throughput rather than the QoS of specific traffic flows. Nevertheless, as more and more multimedia transmissions appear in wireless communication systems, QoS provisioning at the MAC layer becomes increasingly important. There is a large body of work that addresses this issue. Packet reservation multiple access (PRMA) [7–10], the collision resolution and dynamic allocation (CRDA) MAC protocol [11] and other work [12–14] combine random access and reservation techniques to provide better QoS for multimedia traffic. All these schemes are based on the single packet reception model, where errors always happen if more than one transmissions occur simultaneously on the channel. Many new physical layer techniques have been developed recently to enable the MPR capability of wireless links. With these advanced techniques, one or more user terminals can proceed with data transmissions simultaneously without interference between each other that might otherwise cause serious transmission errors. Such physical layer techniques include CDMA [15–18], multi-input-multi-output (MIMO) [19], orthogonal frequency-division multiple access (OFDMA) [20], among others. Compared to single packet reception wireless channels, MPR channels allow simultaneous low-data-rate transmission from several users, to greatly decrease the access delay from the low-data-rate users. Also, since

collisions can be avoided in the MPR channel, it offers benefits for contention-based multiple access.

Conventional multiple access protocols are not fit for wireless communication systems with the new MPR techniques. Consequently, developing efficient multiple access protocols for fully exploiting the MPR capability of the wireless links while providing efficient QoS for different multimedia transmissions becomes a new challenge. [21] and [22] proposed two MAC protocols for the MPR channel, where the state of the traffic load is estimated and a set of users is scheduled to access the channel to maximize the channel utilization. [23, 24] analyzed the performance of the traditional ALOHA protocol on the MPR channel. [16] and [17] employed both time-division multiple access (TDMA) and CDMA in the same wireless link. [18] adopted the packet reservation multiple access protocol in CDMA networks, and [25] and [26] proposed effective scheduling algorithms for CDMA networks to enhance the QoS of multimedia traffic.

One important contribution of this thesis is the proposal of a new MAC scheme, multi-reservation multiple access (MRMA) [27], for wireless channels with MPR capability.

1.3 Scheduling Algorithm for Multi-Code CDMA

One of the typical wireless technologies with the MPR capability is the multi-code CDMA (MC-CDMA) [30]. As the MC-CDMA can flexibly provide diverse transmission rates to a variety of devices over a shared wireless channel, it has been extensively studied in both academic and industry research [31–36]. It has been adopted as the Universal Mobile Telecommunications System (UMTS) [37] and high-speed downlink packet access (HSDPA) [38, 39] standard and will continue to be used in many current and future wireless communication systems [40, 41]. Al-

though the MC-CDMA can support simultaneous multimedia transmissions, it faces challenges in providing multimedia traffic with the required QoS. This goal is accomplished by multiple access protocols, or scheduling algorithms, that fairly and efficiently provide the required QoS for all traffic flows (the FDR or PDR and packet delay).

In infra-structured wireless communication systems, a base station or an access point usually serves as the central controller for all wireless user terminals within its coverage area. While MAC schemes provide coordination for uplink transmissions among different users, the scheduling algorithm that runs on the central controller coordinates the down-link transmissions. In some cases, the scheduling algorithm can also be used for up-link transmissions as a complement to MAC schemes, as long as all of the transmission information can be obtained by the central controller and the central controller can notify user terminals about scheduling decisions. The scheduling algorithm determines the service sequences for different transmissions so that specific QoS provisioning can be achieved. Conventional scheduling algorithms for best effort data service include round-robin [42, 43], fair queuing (max-min fair) [44], proportionally fair scheduling [45], and maximum throughput [46], etc.

In CDMA networks (and many other MPR systems), apart from ordering the transmissions of different flows, scheduling algorithms also have radio resource management functions for controlling the interference between users so that the physical layer QoS requirements can be satisfied. In most multiple access control protocols for CDMA networks, the radio resource management function and the transmission ordering function are independent of each other. Specifically, the radio resource management function aims to support QoS by keeping the physical layer BER (or SINR) below (or above) a certain threshold. The transmission ordering function seeks to serve the traffic flows fairly or with priority, while maximizing the link

throughput or minimizing the data frame drops [47–49].

Compared to the common CDMA, where a scheduled user always sends transmissions at a constant rate, the MC-CDMA system enables scheduled users to transmit at multiple rates using multiple codes. Thus, the scheduling algorithm must not only determine which users can transmit, but also determine the rate of transmission. Various scheduling algorithms have been proposed for MC-CDMA [25, 26, 50]. To support multimedia traffic, sophisticated scheduling algorithms are used in wireless communication systems to guarantee the physical layer QoS (SINR or BER) and to provide differentiated MAC layer QoS FDR for different traffic flows.

As a complement to the MAC scheme, the scheduling algorithm is another key element in QoS provisioning in the MC-CDMA and other wireless communication systems. Another important contribution of this thesis is in the proposal for a new scheduling scheme for the MC-CDMA, called cross-layer enhanced scheduling (CEPS) [51, 52].

1.4 Cross-Layer Optimization of Quality of Service

In the traditional network infrastructure, communication tasks are divided into sub-tasks and handled in separated layers independently. The layered architecture makes the network planning and implementation easier and more efficient. Nevertheless, while numerous wireless technologies have been proposed recently, the cross layer technologies are suggested to enhance the wireless link capacity by sharing the knowledge among different layers and applying adaptation techniques [55–59].

Previous work on multiple access or scheduling algorithms has usually been aimed to separately meet the physical layer QoS requirement, in terms of SINR or BER, and the data link

layer QoS requirements, in terms of FDR (or PDR) and delay. The physical layer parameter, BER, and the MAC layer (or the data link layer) parameter, FDR, have similar effects on the system, in terms of the QoS seen at the higher layer. This is due to the fact that data frames with uncorrectable bit errors are also dropped. In fact, given the data frame structure, BER can be further translated into the frame-error rate (FER), the fraction of data frames that are lost due to channel errors. We define the FLR as the fraction of total data frames that are lost due to both data frame transmission errors (represented by the FER) and missed transmission deadlines or buffer overflow (represented by the FDR). Similarly, if the data link layer has a re-transmission capacity and can re-transmit an error data frame, the re-transmission will lead to more data frame drops, since the chance of data frames missing the deadline or the chance of buffer overflow is increased. In such cases, BER will affect the FDR directly, and FLR will be equal to FDR.

In MC-CDMA networks, since scheduling algorithms affect both frame error rate (FER) (by radio resource management) and FDR (by transmission ordering), these physical and MAC layer QoS parameters should be considered jointly; i.e., to optimize the overall FLR, instead of FER (or BER) and FDR separately. We present a simple example to illustrate the cross-layer QoS issue, as studied in this thesis. Consider the case where one slot is available in the current TDMA frame, which can be used to transmit multiple data frames (e.g., using CDMA), and when two, three, or four data frames are transmitted simultaneously in the slot, the corresponding FER are 0, 0.2, and 0.6, respectively. At the MAC layer, four data frames must be transmitted in the frame; otherwise, they will be dropped because of the delay violation. From the view of the physical layer, transmitting two data frames in the slot is better, so as to keep the FER reasonably low, where the FLR is $2/4 = 0.5$ and the throughput is 2 (data

frames). On the other hand, from the view of the MAC layer, transmitting all four data frames is better, as the FLR is 0.6 and the throughput is $4 \times (1 - 0.6) = 1.6$ (data frames). From the view of the cross-layer consideration, transmitting three data frames and dropping one is better for achieving the best FLR of 0.4 (i.e., $(3 \times 0.2 + 1)/4$), where the throughput is also maximized at 2.4 (i.e., $4 \times (1 - 0.4)$). From this simple example, it can be seen that the cross-layer optimization of the FLR results in a better performance, by accounting for the MAC layer queue statuses or the delay requirements. Moreover, the optimization of FLR actually optimizes the channel throughput as well. This cross-layer QoS consideration can also be used with other wireless transmission technologies.

The cross-layer QoS consideration can be used in scheduling algorithms and other transmission parameter optimizations to enhance the QoS provisioning of the wireless link. The CEPS, proposed in this thesis, is employed to enhance the QoS provisioning and system capacity. Also, a cross-layer optimization of the maximum number of simultaneous transmissions in MC-CDMA networks is proposed in this thesis, based on the cross-layer QoS consideration [60].

1.5 Opportunistic Scheduling and Adaptive Modulation and Coding

To further study the cross-layer QoS consideration in scheduling algorithms for multimedia transmissions over wireless communication systems, we studied the wireless communication systems with AMC, where the QoS provisioning of multimedia transmissions is also a challenge.

A notable feature of wireless communication, in contrast to wired communication, is that

every user may have a different channel status while sharing the same wireless channel. The channel status is determined by the fading of the signal through the paths between antennas of the sender and receiver. This is also affected by the interference received by the receiver's antenna. In traditional wireless communication systems such as 802.11e and UMTS, the physical layer QoS BER is guaranteed. If the BER cannot be guaranteed even with the maximum transmission power, the scheduling algorithm can temporarily block the user's transmission and give channel access to some other users with better channel statuses.

In recent years, advances in wireless communications have enabled the same transmission over a wireless link to use different transmission rates, based on different channel statuses. The better the channel status, the higher the transmission rate, which can be achieved while the physical layer QoS is guaranteed. A typical example of such technology is AMC. In AMC, based on the channel status, the system can choose different coding (forward error coding) and modulation schemes. While some coding and modulation pairs result in very low transmission rates, they can tolerate a relatively poor physical layer channel status, and other coding and modulation pairs can achieve very high transmission rates but require very good channel statuses. The system uses training sequences to detect the channel status before making decisions about the coding and modulation schemes. AMC is already widely used in today's wireless communication systems. Because of its capability to maximize system efficiency over the varying wireless channels, it has been adopted at the physical layer in several important standards, including 3GPP [38, 69], 3GPP2 [70, 71], HIPERLAN/2 [72], IEEE 802.11a [73, 74], IEEE 802.15.3 [75], and IEEE 802.16 [76], and has been combined with different advance technologies such as multi-carrier code division multiple access, MIMO, and cooperative networks.

Scheduling methods that operate in AMC-based systems can improve the overall channel

throughput by selecting the user with the best channel status to transmit. Such a strategy is called opportunistic scheduling, or multiuser diversity [77–82]. Nevertheless, the opportunistic scheduling may result in unfairness and QoS degradation problems, particularly for real-time traffic [83]. Hence, the design of a fair and efficient scheduling algorithm is strongly needed to support multimedia communications over AMC-based systems. Several authors have proposed ways for addressing this problem using different strategies, with the aim to fairly provision the MAC layer QoS FDR for different traffic flows, while keeping the physical layer QoS SINR or BER guaranteed [84, 85].

This thesis makes a significant contribution in proposing a novel scheduling algorithm for AMC-based systems [86]. The cross-layer consideration, introduced above, is used to optimize the channel throughput, while guaranteeing the QoS provisioning for different traffic flows.

1.6 Summary of Contributions

This thesis examines several state-of-the-art research problems in the area of wireless communication systems. The objective is to support multimedia transmissions with QoS provisioning in wireless communication systems. The major contributions of this thesis are divided into four chapters, as summarized below:

- Chapter 2 proposes a novel multiple access scheme over MPR channels, called MRMA. MRMA addresses some new issues of the reservation protocol for the MPR channel, such as adjusting the random access scheme, designing a new slot allocation mechanism to take advantage of the MPR capability of the channel, preventing lost reservations and infinite reservations, and accounting for both data frame drops and data frame errors for

realistic FLR. A Markov model for analyzing the performance of the proposed MRMA protocol is proposed in Chapter 2. The proposed Markov model is suitable for analyzing the access performance of multimedia traffic as it incorporates bi-state traffic sources (e.g., mini-sources for video), in general, and on-off voice sources, in particular. The model is also used to validate the simulation model, which is used to carry out detailed performance evaluations. Chapter 2 presents simulation and analytical results that quantify the performance of MRMA for integrated multimedia (voice, video, and data) traffic.

MRMA can achieve very good QoS provisioning and keep the operation simple and straightforward, which are important aspects for access control in wireless communication systems. MRMA can differentiate the QoS of voice, video, and best effort data traffic flows, and the PLR and packet access delay of voice and video transmissions can be guaranteed. Thus, multimedia transmissions can be better supported in a practical system.

MRMA is an effective MAC scheme for wireless channels with MPR capacity, when compared to other, conventional MAC schemes. Its specific design can efficiently exploit the capacity of wireless channels and manipulate channel access wisely. It can be used in practical systems, such as CDMA networks. Furthermore, many insights can be gained for MAC scheme designs in future wireless communications, stemming from the many interesting observations made during the design phase.

- Chapter 3 proposes the CEPS scheme for scheduling multimedia cross-layer enhanced packet traffic over the uplink MC-CDMA networks. In CDMA networks, since scheduling algorithms affect both FER and FDR, these physical and MAC layer QoS parameters

should be considered jointly, i.e., to optimize the overall FLR, instead of FER and FDR, separately. In CEPS, the scheduling algorithm makes decisions by considering the joint FLR. Also, in CEPS, the total FLR experienced by the system is shared among all traffic flows according to predesigned weights. By estimating the data frame losses that may happen for the current frame and for later frames, according to the current buffer status of different traffic flows, CEPS optimizes the sum of the weighted FLR for all traffic flows. Also, the historical FLR experienced by different traffic flows is considered by CEPS in providing the proportional fairness. Chapter 3 presents both the essential optimization objective and the detailed practice steps of the scheduling algorithm.

CEPS is a great enhancement of the system performance in terms of channel utilization and the maximum admissible traffic flows, due to its cross-layer QoS consideration. The enhancement can provide the more effective capacity for the wireless communication system, which is crucial for wireless communication systems with limited wireless radio spectrum resources. CEPS verifies the effectiveness of the cross-layer QoS consideration. The rest of this thesis is focused on how the cross-layer QoS consideration can be used to improve the performance of other systems and in other aspects.

Compared to other scheduling algorithms, CEPS can give more effective QoS provisioning for multimedia transmissions. It can provide QoS guarantee to specific multimedia traffic flows and share the rest of the bandwidth fairly among others. This feature can greatly improve the users' experience when subscribing to multimedia applications over wireless communication systems. CEPS also has a low complexity such that it can easily be applied to practical MC-CDMA systems for enhancing the system performance.

- Chapter 4 proposes a cross-layer method for optimizing the QoS provisioning of multimedia traffic in an MC-CMDA system by dynamically adjusting the maximum number of simultaneous data frame transmissions in every frame. The optimization is based on the current buffer status and the traffic status of specific traffic flows. The problem is formalized as a Markov Decision Process (MDP) and solved by linear programming. The result is stored in the system upon new traffic flow being admitted and based on old traffic flows remaining, and applied at the beginning of every frame. Chapter 4 also presents the theoretical analysis for the QoS parameter in terms of the FLR, average data frame delay, and channel throughput. The MDP solution is computationally infeasible if the size of the whole system is too large. Thus, Chapter 4 presents an approximation method for the cross-layer optimization. According to the approximation method, only the traffic status is considered in the optimization, and the buffer status is considered statistically.

  The optimization method proposes maximizing the channel capacity by choosing an appropriate number for the maximum number of simultaneous transmissions in one time slot, so that more multimedia transmissions can be admitted into the system. It also verifies the effectiveness of the cross-layer QoS consideration. By using the optimization method in a practical MC-CDMA system, the whole system is expected to provide better QoS for multimedia applications, since the capacity is greatly enhanced. With the proposed approximation method, even very large systems can be optimized with the cross-layer QoS consideration.

- Chapter 5 proposed a scheme, called QoS-CLS, to dynamically determine the scheduled traffic flow and the modulation and coding pairs used for each transmission time interval.

Instead of optimizing the FDR and BER separately, QoS-CLS optimizes the FLR by considering both the physical layer and the MAC layer information. For FLR-sensitive traffic flows, the FLR can be guaranteed. For best effort data traffic flows, the FLR of the system is distributed according to pre-specified weights or priorities. To relax the computational complexity, Chapter 5 also presents two approximation methods. The first method divides the buffer states of a specific traffic flow into a small number of groups and tracks the buffer states in terms of groups. In the other method, optimization is decomposed into two parts with each being less computationally complex, compared to the original optimization.

The scheduling in the AMC-based system is of interest to research. A tradeoff exists between the channel capacity optimization and the fair QoS provisioning. Both are important for multimedia transmission over wireless communication systems. Compared to other work, QoS-CLS can support these two goals in one optimization while achieving a good balance. It gives a good solution for the scheduling problem for AMC-based systems. First, a good QoS provisioning is implemented, and the FLR and delay of specific transmissions can be guaranteed. The other transmission can share the channel fairly so that the FLR can be achieved proportionally for predesigned weights. Second, while the above QoS provisioning is achieved, the capacity of the wireless link is also optimized. Thus, the number of transmissions admitted into the system can be maximized. QoS-CLS also verifies the effectiveness of the application of the cross-layer QoS consideration in AMC-based systems. QoS-CLS can achieve much better channel throughput because of its cross-layer QoS consideration. The proposed approximation methods can alleviate

the computational burden of optimization in practical systems.

1.7 Thesis Organization

This thesis is presented in the manuscript-based format, which is one of the two thesis formats specified by the Faculty of Graduate Studies at the University of British Columbia (UBC). Each of the inner chapters (Chapters 2 to 5) is written in the style of a journal manuscript, based on manuscripts that have been published, accepted, submitted, or under preparation for submission. Moreover, each chapter has its own bibliography as required for the manuscript-based thesis format. The remainder of the thesis is organized as follows. Chapter 2 introduces the multiple reservation multiple access scheme proposed for MPR channels. Chapter 3 proposes the cross-layer enhanced packet scheduling scheme for MC-CDMA systems that schedules data frames on the basis of historical joint data frame losses and estimated joint data frame losses in current and future frames. Chapter 4 presents the QoS-based cross-layer scheduling for MC-CDMA systems that optimizes the maximum number of simultaneous transmissions for achieving a better system FLR. Chapter 5 presents the optimization method proposed for adaptive modulation and the coding-based system, and its approximation methods. The method optimizes the channel throughout while guaranteeing the FLR for specific traffic flows and sharing the FLR among other best effort data traffic flows. Chapter 6 summarizes the findings of the thesis and discusses possible future research directions.

Table 1.1: ITU-T Recommendations for Multimedia QoS Target

Medium	Application	Typical data rates	Key performance parameters and target values			
			One-way delay	Delay variation	Information loss	Other
Audio	Conversational voice	4-64 kbit/s	< 150 ms preferred < 400 ms limit	< 1 ms	< 3% packet loss ratio (PLR)	
Audio	Voice messaging	4-32 kbit/s	< 1 s for playback < 2 s record	< 1 ms	< 3% PLR	
Audio	High quality streaming audio	16-128 kbit/s	< 10 s	<< 1 ms	< 1% PLR	
Video	Videophone	16-384 kbit/s	< 150 ms preferred < 400 ms limit		< 1% PLR	Lip-synch: < 80 ms
Video	Streaming	16-384 kbit/s	< 10 s		< 1% PLR	

Bibliography

[1] 3GPP, "Quality of service (QoS) concept and architecture," *3GPP TS23.107*, v6.4.0, Mar. 2006.

[2] IEEE-802.11WG, "IEEE 802.11e standard draft/D8.0: Draft supplement to standard for telecommunications and information exchange between systems LAN/MAN specific requirements part 11: MAC enhancements for quality of service (QoS)," Feb. 2004.

[3] ITU-T, "ITU-T Recommendation G. 1010: End-user multimedia QoS categories," *ITU-T G.1010*, Nov. 2001

[4] ITU-T, "ITU-T Recommendation Y. 1541: Network performance objectives for IP-based services," *ITU-T Y.1541*, Feb. 2006

[5] N. Abramson, "THE ALOHA SYSTEM: Another alternative for computer communications," *Proc. of AFIPS Joint Computer Conference*, pp. 281-285, 1970.

[6] F. Tobagi and L. Kleinrock, "Packet switching in radio channels: Part II–The hidden terminal problem in carrier sense multiple-access and the busy-tone solution," *IEEE Trans. Commun.*, vol. 23, no. 12, pp. 1417-1433, Dec. 1975.

[7] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Trans. Commun.*, vol. 37, no. 8, pp. 885-890, Aug. 1989.

[8] S. Elnoubi and A. M. Alsayh, "A packet reservation multiple access (PRMA)-based algorithm for multimedia system," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 215-222, Jan. 2004.

[9] S. Nanda, D. J. Goodman and U. Timor, "Performance of PRMA: A packet voice protocol for cellular systems," *IEEE Trans. Veh. Technol.*, vol. 40, no. 3, pp. 544-598, August 1991.

[10] E. D. Re, R. Fantacci, G. Giambene, and W. Sergio, "Performance analysis of an improved PRMA protocol for low Earth orbit-mobile satellite systems," *IEEE Trans. Veh. Technol.*, vol. 48, no. 3, pp. 985-1001, May 1999.

[11] L. Lenzini, M. Luise and R. Reggiannini, "CRDA: A collision resolution and dynamic allocation MAC protocol to integrate data and voice in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 6, pp. 1153-1163, June 2001.

[12] M. J. Karol, Z. Liu and K. Y. Eng, "Distributed queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks," *Proc. of IEEE ICC'95*, pp. 1224-1231, June 1995.

[13] X. Qiu and V. O. K. Li, "Dynamic reservation multiple access (DRMA): A new multiple access scheme for personal communication systems (PCS)," *ACM/Kluwer Wireless Networks*, vol. 2, no. 2, pp. 117-128, June 1996.

[14] H. C. B. Chan, J. Zhang and H. Chen, "A dynamic reservation protocol for LEO mobile satellite systems," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 3, pp. 559-573, April 2004.

[15] N. D. Wilson, R. Ganesh, K. Joseph and D. Raychaudhuri, "Packet CDMA versus Dynamic TDMA for multiple access in an integrated voice/data PCN," *IEEE J. Sel. Areas Commun.*, vol. 11, no. 6, pp. 870-884, Aug. 1993.

[16] T. Weber, J. Schlee, S. Bahrenburg, P. W. Baier, J. Mayer and C. Euscher, "A hardward demonstrator for TD-CDMA," *IEEE Trans. Veh. Technol.*, vol. 51, no. 5, pp. 877-892, Sept. 2002.

[17] C. Yeh, "A TCDMA protocol for next generation wireless cellular networks with bursty traffic and diverse QoS requirements," *Proc. PIMRC'02*, pp. 2142-2147, Sept. 2002.

[18] A. E. Brand and A. H. Aghvami, "Performance of a joint CDMA/PRMA protocol for mixed voice/data transmission for third generation mobile communication," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 9, pp. 1698-1707, Dec. 1996.

[19] J. Salz, "Digital transmission over cross-coupled linear channels," *AT&T Technical Journal*, vol. 64, no. 6, pp. 1147-1159, July-August 1985.

[20] H. Yin and S. Alamouti, "OFDMA: A broadband wireless access technology," *Proc. of IEEE Sarnoff Symposium, 2006*, pp. 1-4, March 2006.

[21] Z. Qing and L. Tong, "A multiqueue service room MAC protocol for wireless networks with multipacket reception," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 125-137, Feb. 2003.

[22] Z. Qing and L. Tong, "A dynamic queue protocol for multiaccess wireless networks with multipacket reception," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2221-2231, Nov. 2004.

[23] S. Ghez, S. Verdü, and S. C. Schwartz, "Stability properties of slotted ALOHA with multipacket reception capability," *IEEE Trans. Autom. Control*, vol. 33, no. 7, pp. 640-649, July 1988.

[24] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite user slotted ALOHA with multipacket reception," *IEEE Trans. Information Theory*, vol. 51, no. 7, pp. 2636-2656, July 2005.

[25] V. Huang and W. Zhuang, "QoS-orientied packet scheduling for wireless multimedia CDMA communications," *IEEE Trans. Mobile Comput.*, vol. 3, pp. 73-85, Jan.-Mar. 2004.

[26] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 146-158, April 1999.

[27] H. Chen, F. Yu, H. C. B. Chan and V. C. M. Leung, "A novel multiple access scheme over multi-packet reception channels for wireless multimedia networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1501-1511, April 2007.

[28] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834-844, July 1988.

[29] 3GPP, "Physical layer - general description," *3GPP TS25. 201*, v3.4.0, June 2002.

[30] C. Lin and R. D. Gitlin, "Multi-code CDMA wireless personal communication networks," *Proc. of IEEE Commun.*, pp. 1060-1064, Jun. 1995.

[31] A. Hamid, R. Hoshyar, and R. Tafazolli, "Joint rate and power adaptation for MC-CDMA over tempo-spectral domain," *Proc. IEEE WCNC2008*, pp. 969-973, Mar. 2008.

[32] Z. Han, G. Su, A. Kwasinski, M. Wu, and K. J. R. Liu, "Multiuser distortion management of layered video over resource limited downlink multicode-CDMA," *IEEE Trans. Wireless Communi.*, vol. 5, no. 11, pp. 3056-3067, Nov. 2006.

[33] B. S. Thian, Y. Wang, T. T. Tjhung, and L. W. C. Wong, "A hybrid receiver scheme for multiuser multicode CDMA systems in multipath fading channels," *IEEE Trans on Vehicular Tech.*, vol. 56, no. 5, pp. 3014-3023, Sept. 2007.

[34] C. S. Chang and K. C. Chen, "Medium access protocol design for delay-guaranteed multicode CDMA mutimedia networks," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1159-1167, Nov. 2003.

[35] Y. Ma, J. Jin and D. Zhang, "Throughput and channel access statistics of generalized selection multiuser scheduling," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 2975-2987, August 2008.

[36] C. D. Iskander, "Performance of multicode DS/CDMA with noncoherent M-ary orthogonal modulation in the presence of timing errors," *IEEE Trans. Vehicular Techno.*, vol. 57, no. 6, pp. 3867-3874, Nov. 2008.

[37] 3GPP, "Spreading and modulation (FDD)," *3GPP TS25.213*, v3.4.0, Dec. 2000.

[38] W. Xiao, A. Ghosh, D. Schaeffer, and L. Downing, "Voice over IP (VoIP) over Cellular: HRPD-A and HSDPA/HSUPA," *Proc. of IEEE VTC2005*, pp. 2785-2789, Sept. 2005.

[39] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming multimedia users in high-speed downlink packet access (HSDPA)," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 844-855, August 2006.

[40] J. Zhou and M. Gurcan, "An improved multicode CDMA transmission method for Ad Hoc networks," *Proc. of IEEE WCNC 2009*, pp. 1-6, April 2009.

[41] L. Luo, J. Zhang and Z. Shi, "Novel block-interleaved multi-code CDMA system for UWB communications," *Proc. of IEEE ICUWB 2007*, pp. 648-652, Sept. 2007.

[42] L. B. Le, E. Hossain, and A. S. Alfa,"Service differentiation in multirate wireless networks with weighted round-robin scheduling and ARQ-based error control," *IEEE Trans. Commun.*, vol. 54, no. 2, pp. 208-215, Feb. 2006

[43] P. Y. Kong, K. C. Chua and B. Bensaou, "Multicode-DRR: A packet-scheduling algorithm for delay guarantee in a multicode-CDMA network," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2694-2704, Nov. 2005.

[44] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: Issues and approaches," *IEEE Pers. Commun.*, vol. 6, no. 1, pp. 44-53, Feb. 1999.

[45] G. barriac, and J. Holtzman, "Introducing delay sensitivity into the proportional fair algorithm for CDMA downlink scheduling," *Proc. of IEEE 7th Int'l. Symp. Spread Spectrum Techniques and Apps.*, vol. 3, pp. 652-656, Sept. 2002.

[46] H. Zeng et al., "Packet scheduling algorithm considering both the delay constraint and user throughput in HSDPA," *Proc. of Int'l. Conf. Commun., Circuits and Sys.*, vol. 1, pp. 387-92, May 2005

[47] M. A. Arad and A. Leon-Garcia, "Scheduled CDMA: A hybrid multiple access for wireless ATM networks," *Proc. of IEEE Pers., Indoor & mobile Radio Commun. 1996*, pp. 913-917, Oct. 1996.

[48] O. Gurbuz and H. Owen, "Dynamic resource scheduling strategies for QoS in W-CDMA," *Proc. of IEEE GLOBECOM 1999*, pp. 183-187, Dec. 1999.

[49] D. Zhao, X. Shen and J. W. Mark, "Radio resource management for cellular CDMA systems supporting heterogeneous service," *IEEE Trans. Mobile Computing*, vol. 2, no. 2, pp. 147-160, Apr.-Jun. 2003.

[50] P. Y. Kong, K. C. Chua, and B. Bensaou, "A novel scheduling scheme to share dropping ratio while guaranteeing a delay bound in a multicode-CDMA network," *IEEE/ACM, Trans. Networking*, vol. 11, no. 6, pp. 994-1006, Dec. 2003.

[51] H. Chen, H. C. B. Chan and V. C. M. Leung, "Cross-layer enhanced real-time packet scheduling over CDMA networks," *Proc. of IEEE ICON2006*, pp. 1-6, Sept. 2006.

[52] H. Chen, H. C. B. Chan, V. C. M. Leung and J. Zhang, "Cross-layer enhanced uplink packet scheduling for multimedia traffic over MC-CDMA networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 986-992, Feb. 2010.

[53] Q. Liu, S. Zhou and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142-1153, May 2005.

[54] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, May 2005.

[55] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantee in adaptive wireless networks," *IEEE J. Selected Areas Commun.*, vol. 23, no. 5, pp. 1056-1066, May 2005.

[56] F. Yu, V. Krishnamurthy and V. C. M. Leung, "Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless CDMA networks," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 542-555, Feb. 2006.

[57] L. Alonso, and R. Aqusti, "Automatic rate adaptation and energy-saving mechanisms based on cross-layer information for packet-switched data networks," *IEEE Commun. Magaz.*, vol. 42, no. 3, pp. S15-S20, Mar. 2004.

[58] Y. Li and G. Zhu, "M-gated scheduling and cross-layer design for heterogeneous services over wireless networks," *IEEE Trans. Vehicular Technol.*, vol. 58, no. 4, pp. 1983-1997, May 2009.

[59] V. Cirvino, V. Tralli and R. Verdone, "Cross-layer radio resource allocation for multicarrier air interfaces in multicell multiuser environments," *IEEE Trans. Vehicular Technol.*, vol. 58, no. 4, pp. 1864-1875, May 2009.

[60] H. Chen, H. C. B. Chan, and V. C. M. Leung, "Two cross-layer optimization methods for transporting multimedia traffic over multicode CDMA networks," *Proc. of IEEE WCNC'07*, pp. 288-293, March 2007.

[61] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer combining of adaptive modelation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, May 2005.

[62] J. Ramis, L. Carrasco, and G. Femenias, "A two-dimensional markov model for cross-layer design in AMC/ARQ-based wireless networks," *Proc. of IEEE GLOBECOM 2008*, pp. 1-6, Nov. 2008.

[63] M. Schwartz, *Broadband Integrated Networks*: Prentice Hall, 1996.

[64] J. N. Daigle and J. D. Langford, "Models for analysis of packet-voice communication systems," *IEEE J. Sel. Areas Commun.*, vol. 4, no. 6, pp. 847-855, Sep. 1986.

[65] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 446-459, Aug. 1993.

[66] A. T. Anderson and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 719-732, Jun. 1998.

[67] E. P. C. Kao, *An introduction to stochastic processes*, Duxbury Press, 1996.

[68] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley & Sons, New York, 1994.

[69] 3GPP, "Physical layer aspects of UTRA high speed downlink packet access," *3GPP TR 25.848*, V4.0.0, 2001.

[70] 3GPP2, "Physical layer standard for CDMA2000 spread spectrum systems," *3GPP2 C.S0002-0*, v1.0, July 1999.

[71] J. Yang, N. Tin and A. K. Khandani, "Adaptive modulation and coding in 3G wireless systems," *Proc. of IEEE VTC2002*, vol. 1, pp. 544-548, Sept. 2002.

[72] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, no. 5, pp.172-180, May 2002.

[73] IEEE Standard 802.11 Working Group, *IEEE 802.11a Physical Layer Specifications*, July 1999.

[74] F. Peng, J. Zhang and W. E. Ryan, "Adaptive modulation and coding for IEEE 802.11n," *IEEE Proc. WCNC 2007*, March 2007.

[75] H. Hu, Y. Zhang and J. Luo, "Distributed antenna systems open architecture for future wireless communications," Auerbach Publications, CRC Press, May 2007.

[76] IEEE Standard 802.16 Working Group, *IEEE standard for local and metropolitan area networks Part 16: Air Interface for Fixed Broadbandwireless Access Systems*, 2002.

[77] X. Liu, E. K. P. Chong and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Selected Areas Commun.*, vol. 19, no. 10, pp. 2053-2064, Oct. 2001.

[78] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling for wireless systems with multiple interfaces and multiple constraints," *Proc. of the 6th ACM/SIGCOM MSWiM*, pp. 11-19, Sep. 2003.

[79] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," *Proc. of IEEE GLOBECOM 2003*, pp. 533-537, Dec. 2003.

[80] S. H. Ali and V. C. M. Leung, "Mobility assisted opportunistic scheduling for downlink transmissions in cellular data networks," *Proc. of IEEE WCNC2005*, pp. 1213-1218, Mar. 2005.

[81] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," *Proc. of Euro. Wireless*, pp. 2244-2248, Sept. 2004.

[82] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming users in high-speed downlink packet access (HSDPA)," *Proc. of GLOBECOM2004*, pp. 4043-4047, Nov. 2004.

[83] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Communications*, vol. 9, no. 5, pp. 76-83, Oct. 2002.

[84] A. Golaup, O. Holland, and A. Aghvami, "A packet scheduling algorithm supporting multimedia traffic over the HSDPA link based on early delay notification," *Proc. 1st Int'l. Conf. Multimeda Services Access Networks*, pp. 78-82, June 2005.

[85] B. Al-Manthari, H. Hassanein and N. Nasser, "Packet scheduling in 3.5G high-speed downlink packet access networks: Breadth and depth," *IEEE Network*, vol. 21, no. 1, pp. 41-46, Jan.-Feb. 2007.

[86] H. Chen, H. C. B. Chan, and V. C. M. Leung, "A cross-layer scheduling scheme for wireless multimedia transmissions with adaptive modulation and coding," *Proc. of IEEE Broadnets2008*, pp. 249-256, Sep. 2008.

[87] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR: A hign efficiency-high data rate personal communication wireless system," *Proc. of IEEE VTC2000*, pp. 1854-1858, May 2000.

[88] K. W. Choi, D. G. Jeong and W. S. Jeon, "Packet scheduler for mobile communications systems with time-varying capacity region," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 1034-1045, March 2007.

[89] C. Cicconetti, L. Lenzini, E. Mingozzi, and G. Stea, "An efficient cross layer scheduler for multimedia traffic in wireless local area networks with IEEE 802.11e HCCA," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 3, pp. 31-46, July 2007.

[90] E. Biglieri, G. Caire, and G. Taricco, "Limiting performance of block-fading channels with multiple antennas," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1273-1289, May 2001.

[91] G. L. Stüber, *Principles of Mobile Communication*, 2nd ed. Norwell, MA: Kulwer, 2001.

[92] Y. L. Guan and L. F. Turner, "Generalized FSMC model for radio channels with correlated fading," *IEE Proc. Commun.*, vol. 146, no. 2, pp. 133-137, Apr. 1999.

[93] V. Hassel, G. E. Oien, and D. Gesbert, "Throughput guarantees for wireless networks with opportunistic scheduling," *Proc. of IEEE GLOBECOM*, pp. 1-6, 2006.

[94] J. A. Stankovic, M. Spuri, K. Ramamrithan, and G. C. Buttazzo, *Deadline Scheduling for Real-Time Systems: EDF and Related Algorithms*. Norwell, MA: Kluwer, 1998.

[95] S. Lu, T. Nandagopal, and V. Bharghavan, "Design and analysis of an algorithm for fair service in error-prone wireless channels," *Wirel. Netw.*, vol. 6, no. 4, pp. 323-343, Jul. 2000.

Chapter 2

A Novel Multiple Access Scheme over Multi-Packet Reception Channels for Wireless Multimedia Networks [1]

A challenge for future wireless multimedia networks is to design efficient multiple access or MAC schemes to support packet-switched multimedia traffic. Recent research work on wireless communication enables the multiple packet reception capability for wireless channels. As a result, conventional MAC scheme that assumes an error-free collision channel cannot be used on the new MPR channels directly. As the MPR channel will be a suitable medium for packet transmissions over many future wireless multimedia networks, in this chapter, we propose a new reservation-based MAC protocol named MRMA for MPR channels. It aims to maximize the channel throughput while satisfying the access delay bounds of different classes of real-time ser-

[1] A version of this chapter has been published. H. Chen, F. Yu, H. C.B. Chan and V. C.M. Leung, "A novel multiple access scheme over multi-packet reception channels for wireless multimedia networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1501-1511, Apr. 2007.

vices such as voice and video. The major contributions of our work are summarized as follows. First, we present the detailed design of MRMA. Since existing reservation protocols cannot be applied directly to MPR channels, we revisit many protocol features and address some new issues, such as adjusting the random access scheme and designing a new slot allocation mechanism to take advantage of the fact that concurrent transmissions can be received successfully. Moreover, as transmission errors can occur over an MPR channel in practice, we need to handle new issues such as lost reservations to ensure the correct operation of the protocol, and account both packet drops and packet errors for realistic packet loss ratio (plr) representation. The second contribution is the development of a Markov model for analyzing the performance of the proposed MRMA protocol. Existing Markov models for conventional packet reservation protocols cannot be used because of the different properties of MPR channels, i.e., the MPR capacity and the error-prone nature. The proposed Markov model is suitable for analysis of access performance for multimedia traffic as it incorporates bi-state traffic sources (e.g., mini-sources for video) in general and on-off voice sources in particular. The model is also employed to validate the simulation model, which is used to carry out detailed performance evaluations. Last but not least, the chapter presents simulation and analytical results to quantify the performance of MRMA for integrated multimedia (voice, video and data) traffic. The results show that the proposed protocol can accommodate a larger number of real-time traffic flows while guaranteeing their plr and delay bounds. Furthermore, the effectiveness of the service differentiation mechanism is demonstrated by the fact that the best effort data traffic has little and insignificant impact on voice and video traffic. The results lead to some interesting observations and provide valuable insights into the design of MAC protocols for MPR channels in future wireless multimedia networks.

The remaining parts of the chapter are organized as follows. Section 2.1 introduces the related work for MPR channels. Section 2.2 introduces the MPR channel model and traffic models used in this chapter. Section 2.3 presents the proposed MRMA protocol, which is modeled analytically in Section 2.4. Simulation and analytical results are presented in Section 2.5 and their implications discussed. Finally, Section 2.6 concludes the chapter.

2.1 Related Work

The MAC schemes should be able to maximize the channel throughput, bear the unreliability of wireless links and at the same time satisfy the QoS requirements, such as constraints on packet delay and *plr*, of different multimedia services.

Traditional MAC schemes for wireless networks such as Aloha, carrier sense multiple access, dynamic TDMA [1] and PRMA [2] have been commonly designed and analyzed assuming an error free collision channel in which a collision destroys all the packets involved whereas a single packet transmission can always be received without error. However, with recent advances in wireless physical layers enabled by sophisticated signal processing techniques, such an assumption is no longer valid in many contemporary and future wireless networks, in which it is possible to correctly receive multiple packets when concurrent transmissions occur. Such an MPR feature exists in CDMA, multi-user detection, multiple-input-multiple-output and spacing-time coding schemes. These schemes may greatly enhance the system performance and are most suitable for supporting multimedia traffic.

Recently, the development of MAC protocols for the MPR channel has attracted considerable interest. The CDMA/PRMA protocol [3] applies packet reservations in CDMA networks, in

which reserved packets and access request packets use the same time slots. Consequently, QoS may not be guaranteed due to possibly unexpected heavy access requests. In the multi-queue service room protocol [4] and dynamic queue protocol [5], the state of the traffic load is estimated and a set of users is scheduled to access the channel to maximize the channel utilization. However, although the characteristics of different traffic types (e.g., voice, data and video) have great impacts on the design of MAC protocols for multimedia traffic (e.g., see [3, 6, 7]), they have not been considered in [4, 5]. Without using realistic traffic characteristics, user state estimations may not be accurate in practice, which may degrade the performance results presented in [4] and [5]. More recently, a QoS-oriented MAC protocol with fair packet loss sharing (FPLS) scheduling [8] has been proposed for wireless CDMA communications. Although BER is controlled by properly arranging simultaneous packet transmissions and packet loss and delay are controlled by proper packet scheduling, the packet loss and delay caused by the random access process have not been considered in [8]. These issues have a significant impact on the overall MAC performance.

2.2 MPR Channel Model and Traffic Models

In this section, we introduce the MPR channel model as well as voice, video and data traffic models that will be used in later sections to evaluate the system performance.

2.2.1 MPR Channel Model

We consider a system with a base station (BS) and many user terminals (UTs) sharing a common wireless channel, which has MPR capability. We focus on the uplink channel that consists of

time slots of fixed size, each accommodating the transmission time of one information packet exactly. The MPR capability of the channel is described by the MPR matrix $\{C_{i,j}\}$, where the element $C_{i,j}$ is the probability that j packets are successfully received given that i packets have been sent simultaneously [4, 5, 9]. For data reception over wireless channels, the impairments caused by noise and interference can only be considered in a statistical sense. The channel statistics, including effects of multi-user interference, mobility and multi-path fading, are used to construct the MPR matrix. For an example of the MPR model, we consider direct sequence (DS) CDMA employing pseudo-noise spreading codes. With the widely used standard Gaussian approximation, the bit error rate B_e of a packet received with $(w-1)$ interfering packets in a CDMA system, where each packet is spread by a randomly generated code with a spreading factor of N and perfect power control is employed, is given by [4, 5]:

$$B_e(w-1) = Q\left(\sqrt{\frac{3N}{(w-1)+3N\sigma^2}}\right) \tag{2.1}$$

where $1/\sigma^2$ gives the signal-to-noise ratio (SNR). Suppose that a packet has L bits and is channel coded to tolerate at most ε bit errors, the success probability P_s of a packet reception is given by:

$$P_s(w-1) = \sum_{i=0}^{\varepsilon} \binom{L}{i} \big(B_e(w-1)\big)^i \big(1-B_e(w-1)\big)^{L-i} \tag{2.2}$$

So we have:

$$C_{i,j} = \binom{i}{j} \big(P_s(i-1)\big)^j \big(1-P_s(i-1)\big)^{i-j} \tag{2.3}$$

Note that the above mapping is an example for the given modulation and coding scheme. Generally, for any system, there is always a mapping from the SNR to the packet error rate. The MPR model serves as a powerful abstraction of the physical layer to facilitate design and analysis of the MAC layer. The above is just an example showing how the MPR matrix can

be obtained in practice. Note that in general the MPR matrix is not constant, but changes as channel statistics vary over time. For example, a channel can alternate between good and bad states, each having a different SNR and hence a different MPR matrix. A time-varying matrix can be employed provided that the channel stays constant over a packet transmission time. For the channel dynamics over one packet transmission time (for example, the fast-fading effects), a stationary MPR can be formulated by accounting for the statistics of the channel dynamics during the packet transmission.

2.2.2 Traffic Models

In this chapter we consider three classes of traffic: voice, data and video traffic. To facilitate the presentation, we assume that each UT hosts exactly one traffic source. For practical cases where one UT can have multiple traffic sources with different classes, the proposed protocol and models should have the same performance. The traffic models for admitted UTs of different traffic classes are given as follows and the traffic parameters used in this chapter are shown in Table 2.1.

For voice traffic, the popular two-state Markov chain model with a constant bit rate (e.g., 8 Kbps) during the talkspurts and the zero bit rate in the silent gaps is employed [3, 6, 7]. The durations of the talkspurts and silent gaps follow the exponential distribution with the expected values of $1/\alpha = 1s$ and $1/\beta = 1.35s$, respectively [3, 6, 7].

For data traffic, the on/off model given in [7] is employed where data packets arrive according to a Poisson process during the on period and no data packet arrives in the off period. The durations of the on/off periods are independent and follow the Weibull probability distribution (see [7] for details and Table 2.1 for the parameters used).

A video UT is always active while connected. Bit rate variations of the source over time is modeled by the autoregressive Markov model in [10] with parameters given in Table 2.1. We assume that the first packet in each video frame carries the most important information.

2.3 Multi-Reservation Multiple Access Scheme

2.3.1 Frame Structure

A TDMA-like frame structure is used to exploit time-division statistical multiplexing. This frame structure is widely used in third generation (3G) wideband CDMA (WCDMA) systems [11], TD-CDMA systems [12] and beyond 3G/ fourth generation (4G) systems [8, 13]. Figure 2.1 shows the frame structure of the uplink channel. A specific number (F) of consecutive slots are grouped into frames. Table 2.1 shows the frame parameters used in this chapter. Note that the frame length ($T_f = 0.02s$) is chosen such that an active (8kbps) voice UT has exactly one packet (160 information bits) to send in each frame during a talkspurt. This is a commonly used approach in packet reservation protocols [6, 14]. There are two types of slots in each frame: reservation slots/mini-slots and information slots. Reservation slots/mini-slots are used by admitted voice and data UTs to send reservation requests to the BS. To improve efficiency, a reservation slot is divided into t mini-slots for carrying reservation requests, and an extended guard space after the last mini-slot allows UTs to receive request acknowledgements from the BS before the next slot. The value of t is chosen based on the tradeoff between the system efficiency and the complexity. As an example, we assume $t = 3$ in this chapter. As explained in Section 2.3.4, the numbers of reservation and information slots are dynamically varied while keeping a fixed frame length. Like other packet reservation protocols, a reliable and

well-organized high-speed downlink broadcast channel from the BS to all the UTs is employed to send downlink packets, control messages, and immediate feedbacks on contention results. The allocation of information slots and announcement of reservation slots are broadcast to all UTs via the downlink channel before the beginning of each uplink frame.

2.3.2 Basic Protocol

When starting a talkspurt, an admitted voice UT first contends in the next frame to send a reservation request via a mini-slot to the BS using the random access scheme described in Section 2.3.3. At the end of the slot, the BS immediately acknowledges the contention result over the downlink channel. An admitted video UT always maintains its connection with the BS through some reserved slots, so that requests via mini-slots are not needed. When a data UT becomes active, it also needs to send a reservation request to the BS via a mini-slot using the random access method in Section 2.3.3. To provide a best effort service to data UTs using residual bandwidth, the data UTs do not contend for reservation mini-slots until they learn from the BS's acknowledgement that the last reservation slot was idle, indicating that voice UTs likely have completed their reservation requests. Having received the requests, subject to availability the BS assigns information slots to UTs according to the slot allocation method described in Section 2.3.4. After connecting to the BS, video and data UTs keep the BS informed of their slot requirements or buffer status by piggybacking this information on the packets transmitted in information slots. This method is commonly used in packet reservation protocols (e.g., see [15]) to enhance the system efficiency.

2.3.3 Random Access Scheme for MPR Channel

The p-persistent random access scheme, whereby a station accesses a reservation slot with permission probability p until successful, is a common approach in reservation protocols [2, 3, 6, 16] for voice and data UTs to convey reservation requests in reservation mini-slots. However, the value p must be reestablished for the MPR channel as its unique property is not considered in existing reservation protocols. In this chapter, we consider three different cases of p-persistent random access: ideal, fixed and adaptive schemes. Denote S as the expected number of successful requests. If the number of requesting UTs is m, it is not difficult to see that:

$$S = \sum_{i=0}^{m} \binom{m}{i} p^i (1-p)^{m-i} \sum_{j=0}^{i} j \times C_{i,j} \tag{2.4}$$

Note that in (2.4), the $C_{i,j}$ values in the MPR matrix are different from those used for information slots, because the length and coding are different between request and information packets. We can choose p such that S in (2.4) is maximized if we know the number of requesting UTs, m. We refer to this as the ideal scheme, which gives the best results for comparison purposes, but is impractical as it requires perfect knowledge of m. We also consider two approaches that are more practical: fixed and adaptive schemes. The fixed scheme assumes a small number of contending UTs and employs a fixed permission probability p to maximize the chance of success. Previous studies of packet reservation protocols [3, 17] have suggested that $p = 0.3$ generally works well. For MPR channels, it is found that $p = 1$ is better since simultaneous transmissions can be received successfully. However, if there are too many contending UTs, using $p = 1$ will give adverse results. Hence, we also consider a simple adaptive scheme, in which $p = 1$ is normally used, but when there are many contending UTs (or collisions), the BS will signal the UTs via the downlink channel to fall back to a lower p in the next frames. The collision effects

and noise effects are inter-related, as reflected in the MPR matrix, and a Kalman filter may be used to detect the collision rate. For simplicity, a fallback is triggered when $\delta\%$ of the requests transmitted in one frame fail. If a completely idle reservation slot is found, the BS signals the UTs to use $p = 1$ again. Generally, the fallback permission probability p and the threshold δ must be determined based on the system parameters including channel statistics and traffic statistics. For the simulation model in this chapter, we choose the fallback $p = 0.3$ and $\delta = 90$ based on the other system parameters. However, these values can be different in other systems.

2.3.4 Slot Allocation

In MRMA, multiple packets can be assigned to a single slot. Hence, to allocate slots to UTs, we must first determine the maximum number of packets that can be accommodated by each slot based on the *plr* requirements. We define K_{vi}, K_{vo} and K_d as the maximum numbers of video, voice and data packets, respectively, that an information slot can support. Let χ_{vi}, χ_{vo}, and χ_d be the *plr* requirements for voice, video and data traffic, respectively. As an example, consider the case of video traffic. To satisfy the *plr* requirement of video, we have:

$$plr = \sum_{i=0}^{K_{vi}} \frac{K_{vi} - i}{K_{vi}} \times C_{K_{vi},i} \leq \chi_{vi} \tag{2.5}$$

Note that if there are K_{vi} packets, with probability $C_{K_{vi},i}$, $plr = (K_{vi} - i)/K_{vi}$, i.e., $(K_{vi} - i)$ packets are not received, according to the MPR channel model. Summing over all the possible values of i, we can obtain the average plr. Hence, given a *plr* requirement of χ_{vi}, we can find K_{vi}, the maximum number of packets that a slot can support based on (2.5). Then, we can determine the required number of slots to send a certain number of packets. Similarly, the required number of slots for other traffic types can also be worked out. As different traffic types

may be able to endure different numbers of bit errors per packet, their MPR matrices C may be different too. Also, different traffic types usually have different *plr* requirements. Therefore the maximum numbers of packets a slot can accommodate may be different for different traffic types. It is better that the same type of traffic should share the same assigned slots whenever possible for ease of management and for maximizing the bandwidth utilization. However, if more than one slots are not "full" (can hold more packets) after all packets are allocated, packets in these slots should be mixed to further save bandwidth. In this situation, packets from different traffic are allocated in one slot and the maximum number of packets this slot can hold must be limited to fulfill the lowest *plr* target among the heterogeneous traffic streams.

Based on the slot capacity for each traffic type determined above and the requests received in the current frame, the BS allocates slots in the next frame to the requesting UTs using the mechanism described below, and sends the allocation results to the UTs via the downlink channel before the start of the new frame. Subject to slot availability, the BS assigns slots in this order: 1) one slot to every admitted video and voice UT with an ongoing reservation, 2) one slot to each new voice request, 3) one slot to each video UT requiring multiple slots, repeated in a round robin fashion until all video requests are fulfilled, 4) one slot to each data UT (new or existing), repeated on a round robin basis, 5) remaining slots assigned as reservation slots. If there are not enough slots, the BS will not be able to allocate slots for the unsatisfied requests and an unsuccessful UT will retry in the next frame if the need for reserved capacity persists. By the first two steps, the number of reservation requests required to set up reservations is minimized so that the channel utilization is enhanced. The best effort data packets are allocated in the step 4), so that they have little and insignificant impact to the voice and video UTs, which are allocated in the first three steps. Only the remaining slots

are assigned as reservation slots, so that the total packet losses are reduced and the channel utilization is further improved.

2.3.5 QoS

Voice and video traffic streams are time sensitive but they can tolerate moderate packet losses. For example, voice UTs can tolerate a *plr* of about 1% [6, 7, 16]. Thus the QoS requirements for voice and video traffic at the MAC layer are expressed in terms of upper bounds for access delay and *plr*. To provide bounded access delay, defined as the time taken for a traffic source to gain access to the channel to send an information packet, we adopt the policy that voice and video packets are discarded from the respective buffers if they cannot be sent within D frames. As an example, we use $D = 1$ in this chapter so that voice and video packets must be sent as soon as possible. This assumption is commonly used in packet reservation protocols [3, 16, 17]. The requirement on *plr* bounds can only be satisfied by MAC in conjunction with a connection admission control (CAC) scheme. Therefore the MAC design objective is to maximize system throughput for specified *plr* values, so as to maximize the number of admitted connections. MRMA has two components contributing to the overall *plr*: plr_c due to information packets corrupted by channel errors, and plr_a due to unsuccessful reservation requests or unsuccessful slot allocations. In contrast with existing packet reservation protocols [2, 3, 6, 16, 17], which generally consider only the second type of packet losses above, the MRMA model is more realistic. Basically, plr_c can be controlled using (2.5) by limiting the number of packets assigned to an information slot. On the other hand, plr_a can be controlled by limiting the number of UTs admitted into the system using a CAC scheme [3, 8, 14]. Normally we keep $plr_a \ll plr_c$, so that when the channel state degrades (causing the MPR matrices to change), the MRMA scheme

can adapt quickly by changing the maximum number of packets assigned to an information slot in the following frames, while the CAC function can respond more slowly by trying not to terminate any existing connection unless plr_a increases so greatly that the overall plr cannot meet the required QoS target. This strategy minimizes the impact on the current traffic when the channel condition is changed. Note that the MPR matrices can be dynamically adjusted based on results of channel state estimations. For real-time traffic, the system throughput can be calculated as the product of the sources' packet generation rates and the corresponding plr since plr includes all packet losses at the MAC layer due to reservation failures or transmission errors.

Data traffic is sensitive to packet losses but not time delay. In MRMA, a best-effort service is provided to data UTs using the residual bandwidth. New data packets are stored in their UTs' buffers until they are transmitted over an allocated information slot. The data traffic has an indirect influence on voice performance in that allocation of information slots to data UTs reduces the number of reservation slots available for voice requests. To address this issue, we design a simple control mechanism as follows. If the BS finds no reservation mini-slot available in q consecutive frames, it will stop assigning slots for data traffic. This ensures that data UTs cannot consume all residual slots for an extended time period. Here, the parameter q is used to balance the QoS of data traffic and voice traffic, and also can be adapted dynamically to realize a simple CAC scheme for data traffic; q can be adjusted gradually to reach a desired operating point. Furthermore, upper/lower bounds of q can also be set based on past statistics. In this chapter, we study the effect of q by simulations.

2.3.6 Preventing Lost and Endless Reservations

As packets can be lost over the MPR channel due to multi-packet interference, MRMA needs to incorporate mechanisms to guarantee proper protocol operations. If the BS does not receive the first voice packet of a UT with a successful reservation, it must distinguish whether it is a packet loss due to transmission errors or the talk-spurt has ended during the reservation process. The BS may wait a little longer, say, 2 or 3 frames, for a voice packet. If the BS cannot receive it during this period, the reservation is cancelled. At the end of the talkspurt, the UT piggybacks on the second last voice packet a request to end the reservation. The BS acknowledges it and wait for the last packet from the UT. If the ending request or the acknowledgement is lost, the UT resends the ending request. If the last packet is lost, the BS terminates the reservation after receiving 2 or 3 blank slots. These mechanisms prevent endless reservations and premature termination of reservations due to lost packets (i.e., reservation losses), and minimize the impact of lost packets on system performance. The losses may also happen to video packets with piggybacked requests for increasing bit rates. In this case, the BS simply continues with the UT's existing reservation (i.e., serve the video UT with the current rate) until the UT reissues the request.

2.3.7 Other Enhancements

The objective of this chapter is to present the core framework of MRMA while giving the flexibility for implementing other advanced services. For example, as mentioned above, while we assume that slots are allocated on a round-robin basis, other allocation mechanisms can also be used without affecting the basic operation of the protocol. Furthermore, although a best-effort service is provided for data traffic, it is also possible to support differentiated data

services by provisioning some dedicated slots and using some appropriate scheduling method (e.g., earliest deadline first policy or weighted fair queuing) under BS control. In addition, it is also of interest to study effective ways to change the system parameters (e.g., q) in an adaptive manner. In summary, the proposed MRMA framework provides the basis for further research work.

2.4 Analytical Model

In this section, we present an analytical model for evaluating the performance of MRMA for bi-state Markov sources in general and on-off voice sources in particular. For example, the bi-state Markov model can be applied to video traffic as mini-sources [10]. Note that the voice and video access processes are in fact quite similar. Like a voice source, a video source also needs to access the BS when connection is first established. Subsequently, the bit rate requirements can be communicated through piggybacking the new bit rate on information packets. Here, we first describe the model for on-off voice sources. Then we will describe how it can be used for video mini-sources. To facilitate the analysis, we neglect lost reservations, and follow the common practice to discretize the traffic models such that the traffic sources only change states at the frame boundaries [18, 19]. A UT can be in one of the following three states: 1) Idle (IDL), i.e., the UT is silent; 2) Connecting (CON), i.e., the UT has started a talk spurt and is trying to connect to the BS; and 3) Reserved (REV), i.e., the UT has successfully reserved a slot.

Consider a group of voice UTs using the MRMA protocol to send packets to a BS over an MPR channel where the MPR matrix remains unchanged. Here, we assume that there are Nv voice UTs. We describe the system state at any moment as a tuple (N_{CON}, N_{REV}), where

N_{CON}, N_{IDL} and N_{REV} are the numbers of CON, IDL and REV UTs, respectively, and $N_{IDL} = Nv - N_{CON} - N_{REV}$. The states ($N_{CON}$, N_{REV}) at the end of the frames form a discrete time Markov chain.

We consider the state transitions in two steps: at the boundary of two frames, and within each frame after each reservation mini-slot. To do this, we define $N_{CON,n}$ as the number of CON UTs at the beginning of frame n (before all slots), and $N_{CON,n,i}$ as the number of CON UTs at the end of the i-th mini-slot in frame n, where the number of mini-slots is h_n. The number of CON UTs at the end of frame n is N_{CON,n,h_n} and the number of CON UTs after all information slots and before all reservation slots is $N_{CON,n,0}$. The notations for REV UTs are defined in a similar manner. The above definitions will be used in the appendices for the calculations. Figure 2.3 shows the state transition graphs.

At the frame boundary, some IDL UTs may become CON. Note that occasionally, an active UT may not be able to secure a reservation successfully before it becomes silent again. In such a situation, the UT will go directly to the IDL state from the CON state. The transition probability T_1 is derived in Appendix A.1. As shown in Figure 2.1, information slots are grouped at the beginning of a frame. At the beginning of a frame, the number of information slots k for the frame can be determined by the system state given by the number of REV UTs and the MPR matrix. Subject to availability, each REV UT is assigned an information slot. At the end of frame n, a REV UT may become IDL (i.e., the UT has sent the last packet in the talkspurt). Also, some CON UTs may become REV after successfully sending a reservation. The state transition probability T_2 is derived in Appendix A.2.

As shown in the appendices, both T_1 and T_2 can be calculated from the system parameters F, T_f, α and β. So, we can get the state transition matrix T from the end of one frame to the

end of the next one as follows:

$$\begin{aligned} &T(N_{CON,n+1,h_{n+1}}, N_{REV,n+1,h_{n+1}} \,|\, N_{CON,n,h_n}, N_{REV,n,h_n}) \\ &= \sum_{N_{CON,n+1}} \sum_{N_{REV,n+1}} T_1(N_{CON,n+1}, N_{REV,n+1} \,|\, N_{CON,n,h_n}, N_{REV,n,h_n}) \times \\ &T_2(N_{CON,n+1,h_{n+1}}, N_{REV,n+1,h_{n+1}} \,|\, N_{CON,n+1}, N_{REV,n+1}) \end{aligned} \tag{2.6}$$

Based on the transition probabilities, we can obtain the stationary state probabilities $\pi(N_{CON}, N_{REV})$ at the end of a frame, and calculate the *plr* and the throughput as follows:

$$plr = \frac{\sum_{N_{CON}} \sum_{N_{REV}} \pi(N_{CON}, N_{REV}) \times (N_{CON} + PLR(N_{REV}))}{Nv \times \frac{\alpha}{\alpha+\beta}} \tag{2.7}$$

$$\eta = \frac{\sum_{N_{CON}} \sum_{N_{REV}} \pi(N_{CON}, N_{REV}) \times (N_{REV} - PLR(N_{REV}))}{T_f} \tag{2.8}$$

where $PLR(N_{REV})$ is defined in (2.9).

$$PLR(N_{REV}) = \begin{cases} \left\lfloor \frac{N_{REV}}{K_{vo}} \right\rfloor \times \sum_i \left(C_{K_{vo},i} \times (K_{vo} - i)\right) + \sum_i \left(C_{N_{REV} \bmod K_{vo},i} \times (N_{REV} \bmod K_{vo} - i)\right) & \\ & N_{REV} < K_{vo} \times F \\ (N_{REV} - K_{vo} \times F) + F \times \sum_i \left(C_{K_{vo},i} \times (K_{vo} - i)\right) & \\ & N_{REV} \geq K_{vo} \times F \end{cases} \tag{2.9}$$

Let us explain (2.7) as follows. Obviously, a CON UT will lose one packet at the end of a frame because it cannot get an information slot in the next frame. Also, if $N_{REV} \geq K_{vo} \times F$, where there are F slots each assigned to K_{vo} UTs, $N_{REV} - K_{vo} \times F$ UTs should lose one packet

each due to no allocation. If $N_{REV} < K_{vo} \times F$, there should be $\lfloor N_{REV}/K_{vo} \rfloor$ slots each with K_{vo} packets and one slot with $(N_{REV} mod K_{vo})$ UTs (x mod y = remainder of x divided by y). Hence, based on the MPR model, $\sum(C_{K_{vo},i} \times (K_{vo} - i))$ and $\sum(C_{N_{REV} mod K_{vo},i} \times (N_{REV} mod K_{vo} - i))$ packet losses are expected for a slot with K_{vo} and $N_{REV} mod K_{vo}$ packets, respectively. Note that in (2.7), the denominator $Nv \times \alpha/(\alpha + \beta)$ is the expected number of active UTs in a frame, and each active UT will send one packet per frame. Similarly, the throughput can be computed by (2.8) where the numerator gives the expected number of packets that can be sent successfully in a frame.

As mentioned, the above model can also be used for video mini-sources, which model the bit rate increase/decrease processes [10]. In this case, each video mini-source alternates between active and idle states, which are equivalent to the talk-spurt and silent states, respectively, in the voice model. Assuming that late packets are discarded in a similar manner, the calculations basically follow the similar approach although the details may be different. The above Markov model cannot be used for data traffic because the data model applies the Weibull distribution, which is not memoryless but gives a more realistic representation of data sources.

Unfortunately, it is not feasible to compute the stationary distribution of the Markov chain derived above for a large number of UTs. It is found that numerical computations using the model can handle up to approximately 50 UTs within a reasonable computation time. Within this limit, the analytical model can be used to calculate the results in minutes while it takes hours to obtain the results by simulations. However, when there are many UTs, it becomes infeasible to use the analytical model because it requires too much memory to store the states or too much time to compute the results. So we apply the analytical model to a small size system to validate the correctness of our simulation model. Then we use the simulation model

to obtain further results under different conditions. This approach is commonly used in the literature [18, 19] for performance evaluations of MAC protocols.

2.5 Results and Discussions

To evaluate the performance of the proposed MRMA protocol, we have implemented a simulator in C++. We consider the simple CDMA system introduced in Section 2.2 with spreading factor $N = 7$. An information packet is 511 bits long and encoded with a (511, 229, 38) BCH code for transmission. A reservation request is 154 bits long and can tolerate 10 bit errors. Unless otherwise specified, Table 2.1 shows the simulation parameters for the system and the traffic. In the simulations, the physical layer is applied statistically and the data link layer is operated step by step. The results are generated from the average of a couple of same simulations, which runs at least 100,000 TDMA frames.

We first justify the use of the adaptive random access scheme. We find by simulations the average number of slots required to allow a certain number of contending UTs to get connected to a BS when different permission probabilities p are used. Figure 2.2 shows that when the number of contending UTs is not large, $p = 1$ has the best performance as it uses the least number of slots. So it is justified to use $p = 1$ when there are not too many contending UTs. We also observe that setting $p \approx 0.3$ can keep the required number of slots at a low level even if the channel is heavily utilized. Hence the above results justify the use of the adaptive scheme, i.e., using $p = 1$ if possible but falling back to $p = 0.3$ when there are many collisions.

Next, we validate the simulation model using the analytical model described in Section 2.4. In this case, a small system with $F = 5$ and $\chi_{vo} = 1\%$ is used. Figure 2.4 compares the plr for

voice sources obtained using the analytical and simulation models when SNR = 10db (baseline or good channel condition) and SNR = 6db (bad channel condition). It can be seen that the analytical and simulation results agree closely with each other. As expected, $p = 0.3$ results in a higher plr than $p = 1$ because the number of contending UTs in this system is very small. Furthermore, $p = 1$ and the adaptive scheme can give a performance that is very close to the ideal scheme. As expected, the results show that at the higher SNR, a better performance can be achieved.

Next, we compare the performance of MRMA and FPLS [8] by simulations at SNR = 6db and SNR = 10db. Note that FPLS also works over an MPR channel but it employs a deadline-based scheduling policy instead of hard reservations. Furthermore, it uses a fixed reservation bandwidth and a simple random access method. To ensure a fair comparison, we use the same frame structure, channel model, and traffic model in the simulations. We assume that FPLS chooses the number of simultaneous transmitted packets in the same way as MRMA, i.e., to meet $\chi_{vo} = \chi_{vi} = 1\%$. For FPLS, we consider that there are 3, 6 or 9 reservation mini-slots per frame and the transmission deadline is one frame time. For video traffic, we assume that FPLS also updates the slot requirements through piggybacking instead of reservation slots. Here, we focus on comparing the plr's of voice and video in an integrated voice/video system. Figure 2.5 and Figure 2.6 show the comparison results for 10 video UTs and variable number of voice UTs. In summary, when there are few UTs and the channel is good (i.e., SNR=10db), both MRMA and FPLS give a similar performance if FPLS uses 3 reservation mini-slots. It is because both FPLS and MRMA keep the plr_c lower than the target plr and the plr_a is close to zero in this situation. However, when the channel condition becomes worse (i.e., SNR=6db) or there are more voice UTs (i.e., the traffic load is higher), MRMA offers lower plr's because the dynamic

reservation bandwidth enables MRMA to utilize the limited bandwidth for both random access and traffic transmission more efficiently than FPLS could, the adaptive access control further improves the random access, and the frame-based reservation mechanism minimizes the required bandwidth for random access. All these factors result in a lower plr_a for MRMA. Note that the performance of FPLS also depends very much on the fixed number of reservation mini-slots, which may be difficult to set to a correct value for dynamic traffic. Essentially, if with a target 1% PLR for voice traffic, FPLS can support 90 to 115 voice UTs or 130 to 160 UTs for SNR = 6db or SNR = 10 db, respectively, while MRMA can support 120 and 170 voice UTs. It is a 4% to 33% or 6% to 30% capacity enhancement. From these two figures, we can also conclude the MRMA provides a higher throughput than FPLS, based on the discussion of the relationship between traffic throughput and plr in Section 2.3.5.

In the following, we further evaluate the performance of MRMA by simulating systems with more UTs. Figure 2.7 shows the plr's of voice and video traffic in a voice/video joint system with $\chi_{vo} = \chi_{vi} = 1\%$, when the number of video UTs is varied while the number of voice UTs is 100. Figure 2.8 shows the plr's versus the number of voice UTs for 15 video UTs. Again, it can be seen that the adaptive scheme can provide a performance close to the ideal scheme. The results are similar in the two figures. Note also that all the curves in Figure 2.7 and 2.8 flatten off at around $plr = 0.0004$ for SNR = 10db. This is because the plr is ultimately limited by plr_c (i.e., channel errors). In the system, based on the target plr of 1%, a slot can support at most 6 packets for SNR = 10db. In fact, with 6 packets assigned to each slot, the achievable plr_c is around 0.0004 according to (2.5), corresponding to the minimum plr shown in the figures. Note that we cannot support 7 packets per slot because in this case, the $plr > 0.01$ according to (5). A similar phenomena can also be observed for SNR = 6db. These two figures also show that

the numbers of voice and video UTs have a similar influence on the target *plr* of both types of traffic, which suggest that it is possible to satisfy the QoS of both types of traffic using a joint CAC scheme to limit the numbers of voice and video UTs.

We have also simulated a multimedia system with voice, video and data traffic, where $\chi_{vo} = \chi_{vi} = \chi_d = 1\%$. Only the results for the adaptive access scheme are shown, as it was found that the results for the adaptive access scheme are very close to those of the ideal scheme. Figure 2.9 shows the result for 90 voice UTs and 10 video UTs, when the number of data UTs varies from 10 to 160. As expected, when the number of data UTs is increased, the *plr*'s of both voice and data traffic increase while the *plr* of video traffic remains almost unchanged. When there are few data UTs, the voice and data *plr*'s are relatively less sensitive to the change of q. However, when there are more data UTs, the data *plr* can be decreased by using a larger q at the expense of increasing the voice *plr*. Figure 2.10 shows the throughput (i.e., bits transmitted per second) for video, voice and data traffic when the number of data UTs is varied. The desirable result that the video and voice throughput is almost unaffected by the data UTs is apparent. When SNR = 10db, the data throughput is less sensitive to the change in the number of data UTs and the q value. Figure 2.11 shows the average data access delay when the number of data UTs is varied. Note that for video and voice traffic, the access delay is bounded within 0.02s (1 frame duration). When SNR = 10db and the number of data UTs is small, the average delay is less affected by the q value. When SNR = 6db, the delay increases more significantly for the same number of data UTs.

Figure 2.12 shows how the parameter q influences the *plr*'s of voice and data traffic when SNR = 10db. When there are 90 data UTs, the traffic load is not heavy, and the parameter q has almost no influence on the voice traffic. So the value of q should be made as large as

possible to enhance the performance of data traffic. When there are 150 data UTs, the traffic load is heavy. Here, the parameter q influences not only the performance of the data traffic, but also the *plr* of the voice traffic. It is obvious that the parameter q here should be set to 1 to keep the voice *plr* within 1%. Based on this observation, we recommend that when the system is underloaded, q should be adjusted to a larger value to increase the system efficiency, and when the system is overloaded, q should be set to a smaller value so that the voice traffic can satisfy its *plr* requirement. The figure also confirms the idea mentioned in Section 2.3.3 that q can be dynamically adjusted to provide basic CAC for data traffic. Essentially, based on the historical statistic of the traffic load, we can estimate an upper bound of q that can guarantee the QoS of voice traffic for a certain traffic load, and also a lower bound of q that can maintain a certain *plr* for data traffic. If the upper bound is lower than the lower bound, the system is overloaded. If not, the parameter q can be adapted gradually between the two bounds.

2.6 Summary

We have proposed a novel MRMA protocol for future wireless multimedia networks with MPR capability. Based on the channel model and the QoS requirements, the proposed scheme controls the number of packets transmitted in each time slot. In addition, service differentiation is provided by assigning different priorities to different traffic types. System performance has been evaluated by simulations and analysis, and compared with FPLS. Results show that the amount of data traffic has only a slight and almost no impact, respectively, on the voice and video performance, verifying the effectiveness of service differentiation. Results also show that, by taking full advantage of the MPR capacity of the channel, MRMA can accommodate a

larger number of real-time traffic streams compared with FPLS while satisfying their access delay bounds and *plr* objectives.

Table 2.1: Simulation Parameters

Definition	**Value**
Channel	
Channel rate after coding	511,000 bps
Channel rate before coding	229,000 bps
Spreading factor	7
Base SNR	10db
Frame structure	
Frame duration	0.02 s
Slot per frame	20
Packet structure	
Packet size before coding	229 bits
Payload size	160 bits
Overhead	69 bits
Packet size after coding	511 bits
Tolerable errors	38 bits
Channel coding	BCH(511,229,38)
Packet size for requests	154 bits
Tolerable errors for requests	10 bits
Voice Model	

continued on next page

Table 2.1: *continued*

Definition	**Value**
Average length of talkspurt	1.0 s
Average length of silent gap	1.35 s
Offered load while in talkspurt	8 kbps
Offered load average	3.4 kbps
Data Model	
Distribution of the on/off duration	Weibull
Average length of on duration	3.3 s
Average length of off duration	22.8 s
Shape parameter	0.88
Number of packets per message	6
Offered load while in on period	15.8 kbps
Offered load average	2.0 kbps
Data buffer size	200
Video Model	
Pixels per video frame	2,400
Video frames per second	30 frame/s
Average bits/pixel	0.421
Variation of bits/pixel	1

continued on next page

Table 2.1: *continued*

Definition	**Value**
Offered load average	30.3 kbps

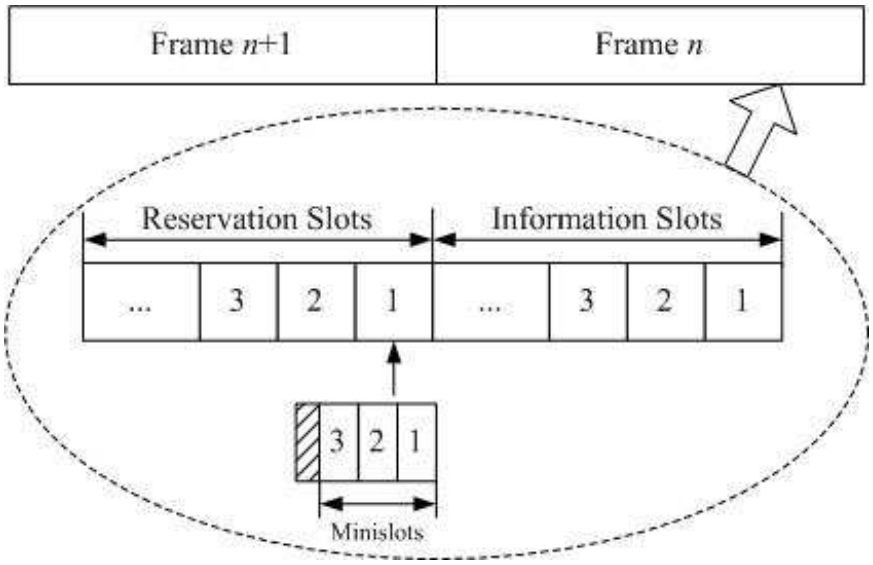

Figure 2.1: Frame structure.

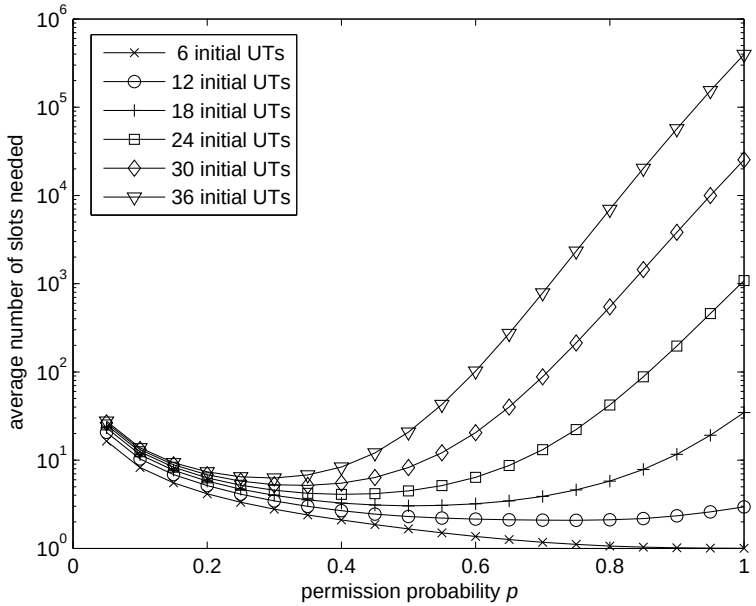

Figure 2.2: Average number of slots needed for different initial numbers of UTs to successfully send requests when the permission probability varies from 0.05 to 1.

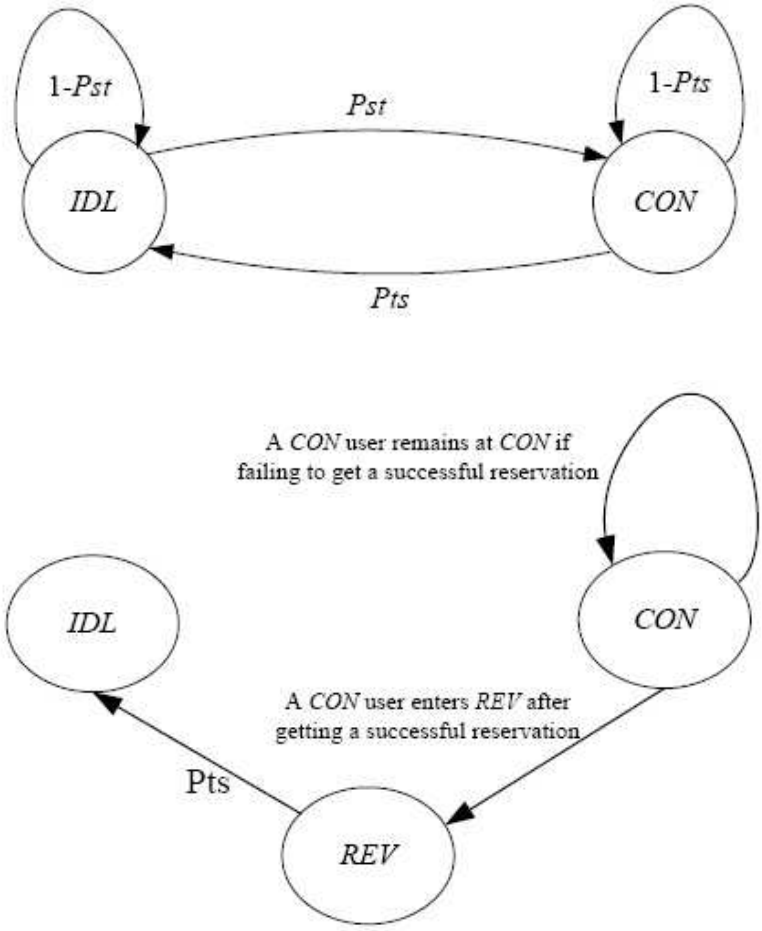

Figure 2.3: State transition graphs at the frame boundary and within a frame.

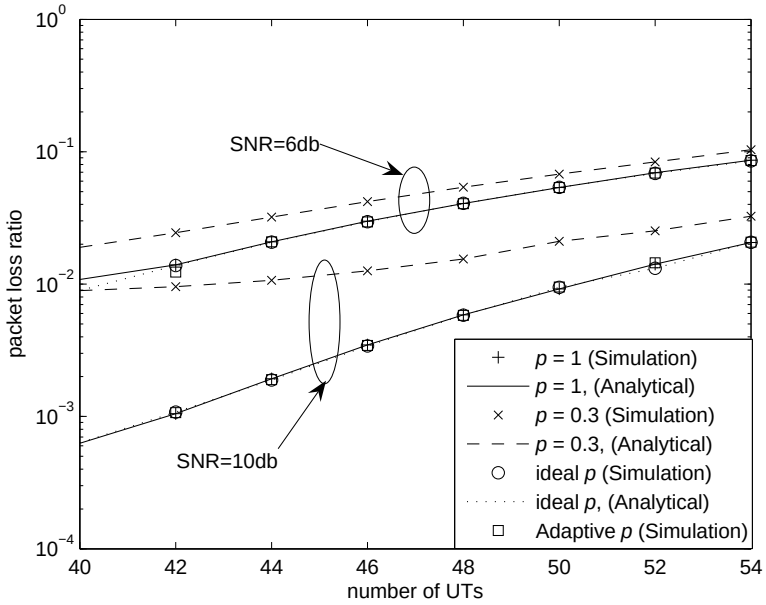

Figure 2.4: Comparison of analytical and simulation results for voice.

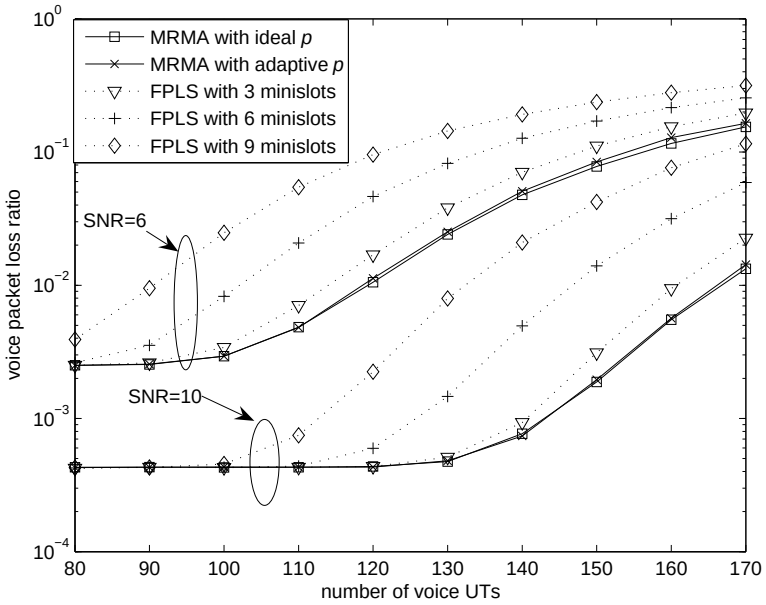

Figure 2.5: Comparison of voice packet loss ratio for MRMA and FPLS.

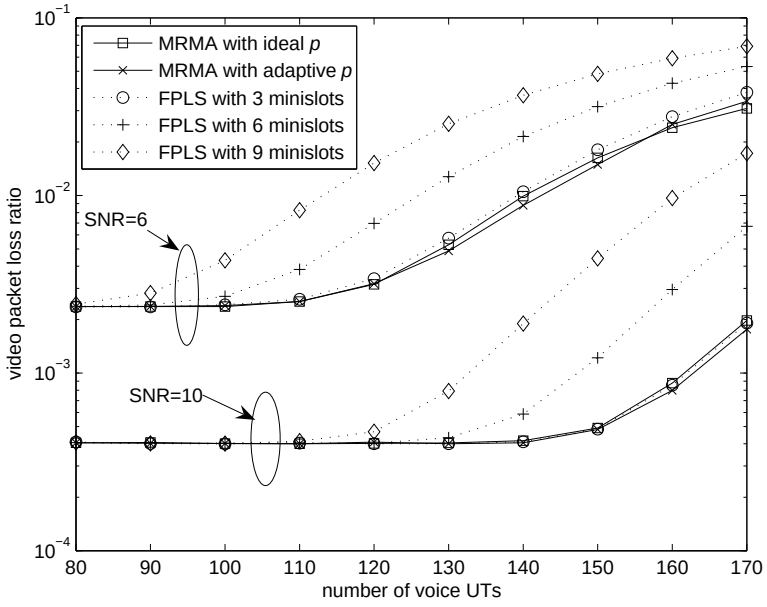

Figure 2.6: Comparison of video packet loss ratio for MRMA and FPLS.

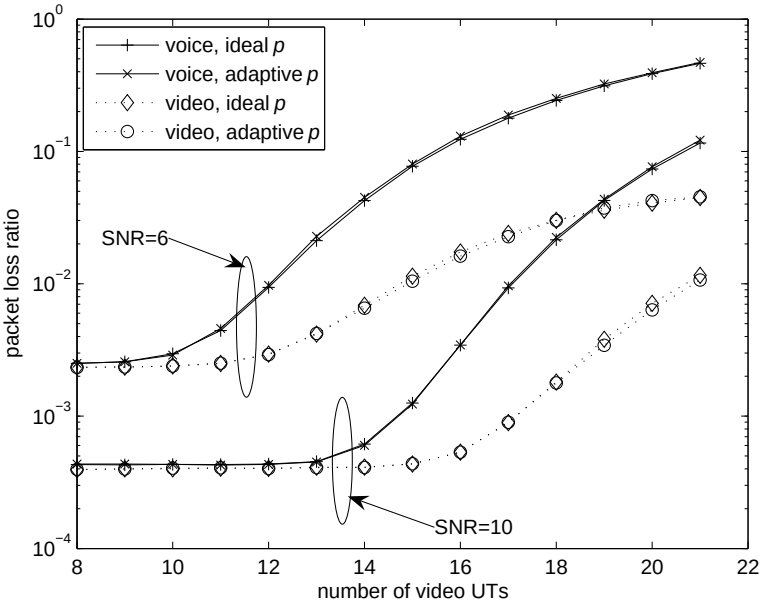

Figure 2.7: Packet loss ratio in a voice/video system with 100 voice UTs and different number of video UTs.

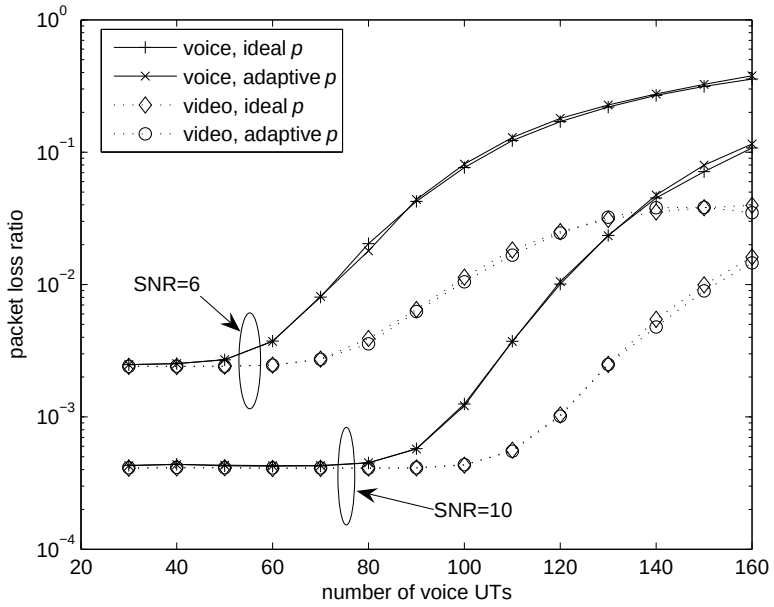

Figure 2.8: Packet loss ratio in a voice/video system with 15 video UTs and different number of voice UTs.

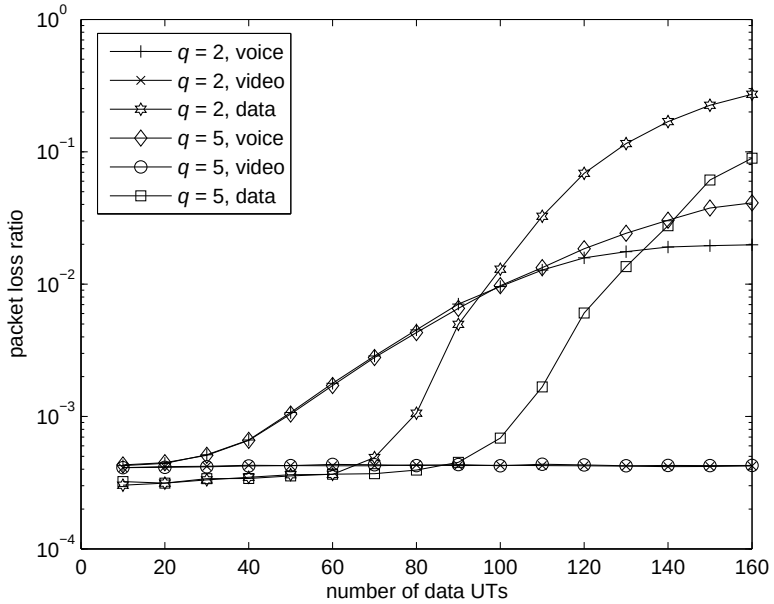

Figure 2.9: Packet loss ratio in a multimedia system with 90 voice UTs and 10 video UTs (SNR=10).

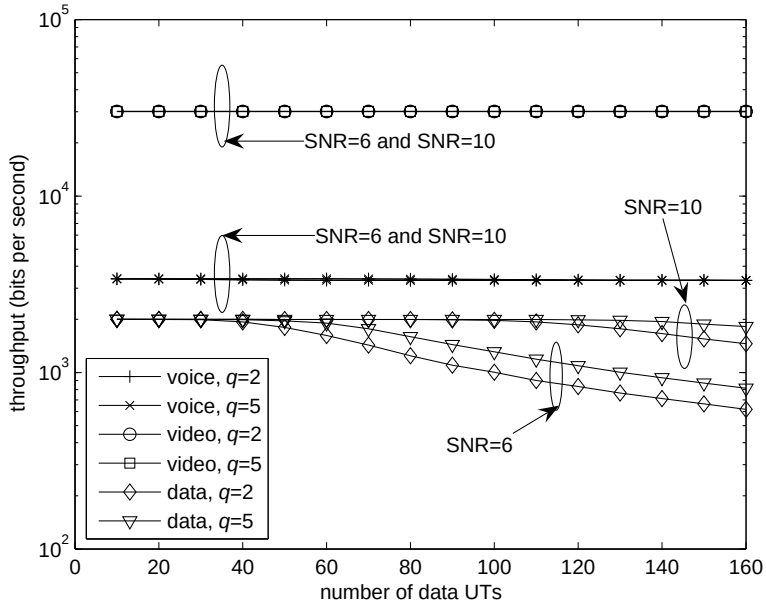

Figure 2.10: Throughput in a multimedia system with 90 voice UTs and 10 video UTs.

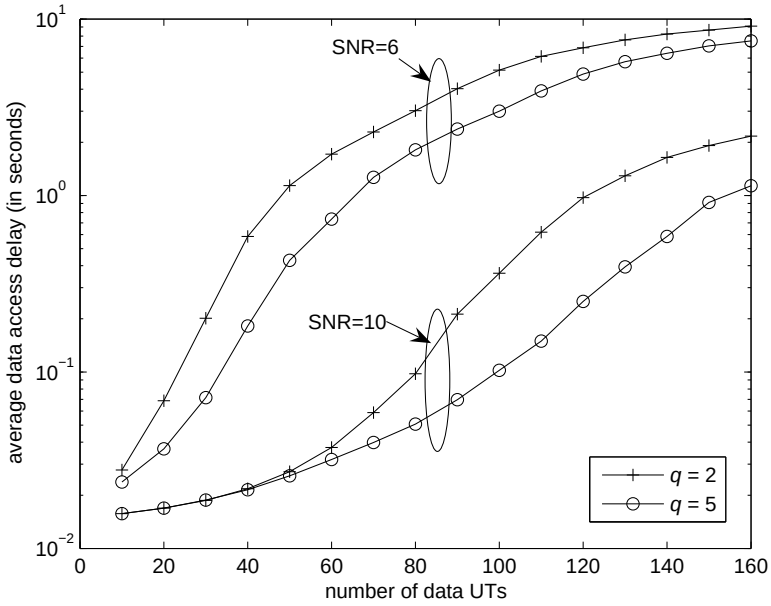

Figure 2.11: Average data access delay in a multimedia system with 90 voice UTs and 10 video UTs.

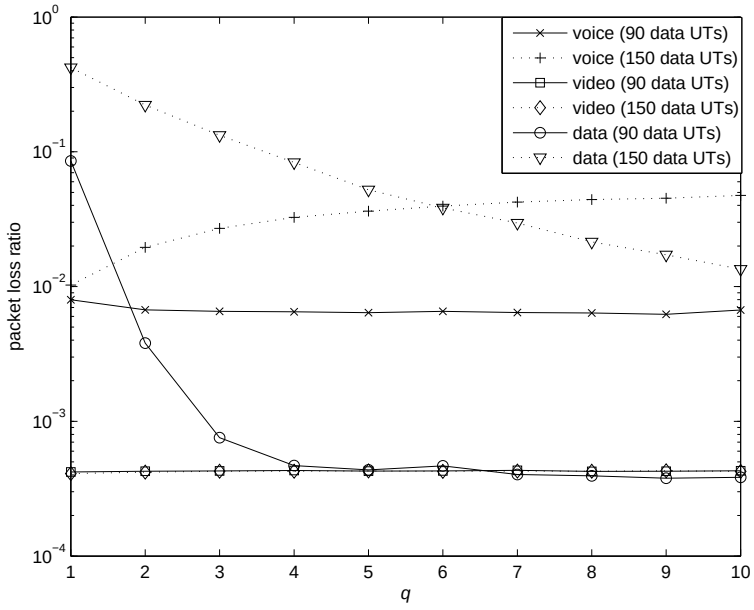

Figure 2.12: Packet loss ratio with 90 voice UTs, 10 video UTs and 90 or 150 data UTs with different q (SNR=10).

Bibliography

[1] N. D. Wilson, R. Ganesh, K. Joseph and D. Raychaudhuri, "Packet CDMA versus Dynamic TDMA for multiple access in an integrated voice/data PCN," *IEEE J. Sel. Areas Commun.*, vol. 11, no. 6, pp. 870-884, Aug. 1993.

[2] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Trans. Commun.*, vol. 37, no. 8, pp. 885-890, Aug. 1989.

[3] A. E. Brand and A. H. Aghvami, "Performance of a joint CDMA/PRMA protocol for mixed voice/data transmission for third generation mobile communication," *IEEE J. Sel. Areas Commun.*, vol. 14, no. 9, pp. 1698-1707, Dec. 1996.

[4] Z. Qing and L. Tong, "A multiqueue service room MAC protocol for wireless networks with multipacket reception," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 125-137, Feb. 2003.

[5] Z. Qing and L. Tong, "A dynamic queue protocol for multiaccess wireless networks with multipacket reception," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2221-2231, Nov. 2004.

[6] S. Elnoubi and A. M. Alsayh, "A packet reservation multiple access (PRMA)-based algorithm for multimedia system," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 215-222, Jan. 2004.

[7] L. Lenzini, M. Luise and R. Reggiannini, "CRDA: A collision resolution and dynamic allocation MAC protocol to integrate data and voice in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 6, pp. 1153-1163, June 2001.

[8] V. Huang and W. Zhuang, "QoS-orientied packet scheduling for wireless multimedia CDMA communications," *IEEE Trans. Mobile Comput.*, vol. 3, pp. 73-85, Jan.-Mar. 2004.

[9] S. Ghez, S. Verdü, and S. C. Schwartz, "Stability properties of slotted ALOHA with multipacket reception capability," *IEEE Trans. Autom. Control*, vol. 33, no. 7, pp. 640-649, July 1988.

[10] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834-844, July 1988.

[11] 3GPP, "Physical layer - general description," *3GPP TS25. 201*, v3.4.0, June 2002.

[12] T. Weber, J. Schlee, S. Bahrenburg, P. W. Baier, J. Mayer and C. Euscher, "A hardward demonstrator for TD-CDMA," *IEEE Trans. Veh. Technol.*, vol. 51, no. 5, pp. 877-892, Sept. 2002.

[13] C. Yeh, "A TCDMA protocol for next generation wireless cellular networks with bursty traffic and diverse QoS requirements," *Proc. PIMRC'02*, pp. 2142-2147, Sept. 2002.

[14] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 146-158, April 1999.

[15] M. J. Karol, Z. Liu and K. Y. Eng, "Distributed queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks," *Proc. of IEEE ICC'95*, pp. 1224-1231, June 1995.

[16] H. C. B. Chan, J. Zhang and H. Chen, "A dynamic reservation protocol for LEO mobile satellite systems," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 3, pp. 559-573, April 2004.

[17] X. Qiu and V. O. K. Li, "Dynamic reservation multiple access (DRMA): A new multiple access scheme for personal communication systems (PCS)," *ACM/Kluwer Wireless Networks*, vol. 2, no. 2, pp. 117-128, June 1996.

[18] S. Nanda, D. J. Goodman and U. Timor, "Performance of PRMA: A packet voice protocol for cellular systems," *IEEE Trans. Veh. Technol.*, vol. 40, no. 3, pp. 544-598, August 1991.

[19] E. D. Re, R. Fantacci, G. Giambene, and W. Sergio, "Performance analysis of an improved PRMA protocol for low Earth orbit-mobile satellite systems," *IEEE Trans. Veh. Technol.*, vol. 48, no. 3, pp. 985-1001, May 1999.

Chapter 3

Cross-Layer Enhanced Uplink Packet Scheduling for Multimedia Traffic over MC-CDMA Networks [1]

We employed an arbitrary slot allocation method in MRMA scheme proposed in Chapter 2. As a result, though the multimedia traffic has higher priority over the best effort data traffic, the further QoS differentiation among different traffic flows is not provided. Such a task can be accomplished by a scheduling algorithm that determines how the channel resource is allocated to different traffic flows. In MC-CDMA, one of the typical MPR enabled system, since scheduling algorithms affect both PER and PDR, it is desirable to consider these physical and MAC layer QoS parameters jointly, i.e., to optimize the overall PLR instead of PER and PDR separately. Designing a scheduling algorithm based on the above optimization approach presents a new set of research challenges. In traditional scheduling algorithms, the PER is typically kept below a threshold. So, given the channel status and the power constraint, it is

[1] A version of this chapter has been published. H. Chen, H. C.B. Chan, V. C.M. Leung, and J. Zhang, "Cross-layer optimization for multimedia transport over multicode CDMA networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no.2, pp. 986-992, Feb. 2010.

not difficult to determine the channel capacity (i.e., the number of packets that can be transmitted in a frame) for scheduling purposes. However, with the above cross-layer optimization approach, the channel capacity is also related to the MAC layer traffic status (i.e., cross-layer information) since the PER is not kept below a threshold anymore. It makes the scheduling task more challenging especially for heterogeneous users where fairness among users has to be addressed as well. A major contribution of this chapter is to tackle these new challenges for cross-layer scheduling. Specifically, we propose a novel scheduling algorithm called cross-layer enhanced packet scheduling (CEPS) for supporting multimedia traffic over MC-CDMA. CEPS focuses on the uplink transmissions because the problem is more challenging than the downlink transmissions. In fact. it is easy to adapt the proposed scheme to schedule the downlink transmissions. CEPS seeks to jointly optimize PER and PDR through a cross-layer optimization framework that takes into account both QoS requirements and fairness. Basically, by considering the MAC layer delay requirements and the queue statuses of different users, both the transmission ordering and the interference among transmissions are controlled by the algorithm to minimize the packet losses (and to maximize the throughput) as seen above the data link layer. Another advantage of CEPS is that it can handle fairness more effectively and flexibly. Simulation results will be presented to illustrate the benefits of the proposed algorithm.

The remainder of this chapter is organized as follows. Section 3.1 introduces the related work. Section 3.2 describes the MC-CDMA system in which we apply our algorithm. Section 3.3 defines the system model and also presents the cross-layer enhancement and fairness provisioning method, which is the central idea of this chapter. Section 3.4 presents the CEPS algorithm for the MC-CDMA system. Section 3.5 presents the simulation results to compare the performance of the proposed algorithm with two other existing algorithms. Section 3.6 concludes the chapter.

3.1 Related Work

One of the challenges in future packet-switched wireless networks is to provision multimedia traffic with the required QoS over unreliable wireless links. This goal is accomplished by scheduling algorithms that fairly and efficiently provide the required QoS to all traffic flows (referred simply as flows in this chapter). Although the cross-layer concept (to optimize the physical layer BER and MAC layer PDR jointly) can be applied to any CDMA networks and other wireless networks with variable rate capability, we study its application in the MC-CDMA [14] networks in this chapter. Due to its ability to support a variety of devices with diverse transmission rates within a single frequency band, MC-CDMA has been studied widly in both academic and industry research [15–17]. It has also been adopted in the WCDMA [18] and HSDPA [19] standards and is expected to be used in many future wireless communication systems.

Many scheduling algorithms [1–5] proposed for CDMA systems attempt to maximize the channel throughput or minimize the PDR while providing fairness among different users. However, they optimize BER at the physical layer and PDR at the link layer separately. Recently, some scheduling algorithms that utilize cross-layer information have also been proposed for wireless networks, especially for systems employing AMC. Opportunistic scheduling algorithms [6–9] can enhance the system throughput. However, they may not be suitable for supporting multimedia traffic, as is our objective in this chapter, since they do not provide hard guarantees on delay and packet loss. In [10], based on the estimation of the effective bandwidth, every QoS-guaranteed flow reserves some specific bandwidth from the system. The reserved bandwidth is used by other flows only when there are no more packets in the queue of the QoS-guaranteed flow. However, the algorithm is based on a very conservative admission control with which the

system can satisfy the throughput requirement of all QoS guaranteed flows when all these users experience the worst channel status simultaneously. So, this approach does not fully utilize the multiplexing gain among QoS guaranteed flows. Also, the estimation of the effective bandwidth greatly depends on having an accurate estimate of the traffic characteristics, which is not always achievable. The protocol in [11] jointly considers channel errors in the physical layer and buffer overflows in the MAC layer. However, it does not take full advantage of cross-layer optimization. In addition, each real-time flow is served at a constant transmission rate based on the estimated effective bandwidth, which may lead to less multiplexing gain. While all the above proposals considered cross-layer information, they did not consider the MAC layer QoS requirements while optimizing the physical layer transmission parameters. Hence the total PLR was not optimized (i.e., the throughput could be further improved). The work in [12] and [13] studied how the MAC layer queue statuses (or retransmission information) and QoS requirements can be used to optimize the physical layer parameters for AMC so that a significant improvement in throughput can be achieved. The parameters were optimized in a statistical manner according to the system dynamics. However, they did not address scheduling issues and the approach cannot be used in CDMA systems.

3.2 Multicode CDMA System

In this chapter, we consider a multi-cell time division MC-CDMA system and focus on the uplink channel. To simplify the presentation, we assume that each user has only one flow. Without loss of generality, consider that the number of time slots in each MAC frame is a fixed number (F), all packets have the same size, and each packet can be transmitted in one time slot using one

code. To achieve different bit rates, a user may transmit multiple packets in each frame using either multiple time slots, or multiple codes over a single time slot, or both. Although in current systems like IS-95, non-orthogonal codes are commonly used [14], we assume that orthogonal codes are used for multiple packet transmissions from the same user over one time slot to achieve a higher bit rate, so that the mutual-inference between the concurrent transmissions of the same user is greatly reduced at the BS due to the same propagation condition on the parallel orthogonal codes [20]. An effective code generation method is introduced in [14]. We also assume that the near-far problem and fading effects are mitigated by perfect power control and can be neglected (an imperfect power control may lead to less effectiveness of the proposed scheme). Also, because there are a large number of interfering streams from all adjacent cells, we can apply the Gaussian approximation, i.e., the inter-cell interference can be considered as a part of the total Gaussian noise with normalized variance σ^2. So, denoting the number of users with i packets transmitted in slot t of frame s as $N_i^{s,t}$, the SINR for a packet in slot t of frame s from a user with i packets in this time slot is [2, 21]:

$$\overline{SINR}_i = \frac{\frac{3}{2}G}{\sum_l l \times N_l^{s,t} - i + 3G\sigma^2 B} \tag{3.1}$$

In (3.1), G is the processing gain, B is the baseband bandwidth for a single code transmission and l is the possible number of packets a user may transmit in a time slot. The BER of such a packet is:

$$\overline{BER}_i = Q(\sqrt{\overline{SINR}_i}) \tag{3.2}$$

where $Q(.)$ is the tail function of the Gaussian distribution. In some practical systems, the SINR and BER models may be more complicated than above. Regardless, the BER and SINR

of a packet in a slot can always be determined given the packets scheduled for transmissions over the same slot and the power constraints on these transmissions.

Furthermore, given the channel coding scheme, the PER can be determined from the BER. Generally, for a coding block or a coding cycle that is w bits long and can tolerate v bit errors, the block error probability is:

$$P_{error} = \sum_{j=v+1}^{w} \binom{w}{j} \overline{BER_i}^{j} (1 - \overline{BER_i})^{w-j} \tag{3.3}$$

The PER is the probability that any block or cycle in a packet is in error. For some systems, the above calculation may not be obtained. But generally, there must be a mapping from the SINR to PER and it can be studied by experiments. We assume that the BS can reliably determine if an uplink packet is correctly received, e.g., by performing a cyclic redundancy check on the received packet. While more sophisticated coding schemes can be applied, a mapping always exists between the channel BER and the PER at the link layer given the coding scheme.

Note that in most of the related previous work, the SINR of each user is guaranteed to be higher than a specified threshold. Based on (3.2) and (3.3), it is equivalent to keeping the PER of each user below a certain threshold.

3.3 System Model and Principle of CEPS

Before explaining the proposed algorithm, we first define the system model, introduce the proposed cross-layer enhancement and study the fairness provision of the proposed scheduling algorithm.

3.3.1 System Model

We consider J classes of multimedia users, where class j (j=1, 2, ..., J) has K_j users. In the subsequent analysis, we consider three traffic classes: conversational, streaming, and interactive [22], which have different QoS requirements. Generally, conversational traffic has a very stringent delay requirement but it can tolerate moderate packet losses. Streaming traffic can also tolerate moderate packet losses but it has less stringent delay requirements, although a delay bound is still desirable. Interactive data traffic requires a very low packet loss rate but it can tolerate much higher delays. Each user station has a buffer to store generated packets. We consider uplink transmissions only, and assume that all users can inform the BS their buffer status via an error-free uplink signaling channel (Note that the modulation and coding scheme for the signaling channel is chosen for high reliability rather than bandwidth efficiency). At the beginning of every MAC frame, the BS applies the scheduling algorithm to decide which packet is to be transmitted in which slot in the frame and informs the user stations accordingly via an error-free downlink signaling channel. This decision is based on the QoS requirements, buffer states, past statistics and channel states of the users.

1. Packet Access Delay and Buffer Size

   *Packet access delay*is one of the most important QoS parameters for multimedia traffic in the MAC layer of a packet switched wireless network, and is defined as the time from the arrival of the packet at the station to the transmission of the packet over the air interface. Generally, there is an expiration time associated with each time-sensitive packet such that the packet is dropped by the sender if it is not sent before it expires [2, 3]. The expiration time is determined by the delay/jitter requirements of the real-time application. In this

chapter, we adopt the common assumption [2, 3] that a packet, which will expire in frame n, must be sent before frame n since the scheduling decisions are made at the beginning of the frames. If the peak packet rate is S_j and the maximum tolerable packet access delay is D_j for class j users, a buffer with size $D_j S_j$ at each sending station will ensure that no buffer overflow happens for these users, and that packets are dropped at the station only due to the packet access delay exceeding the expiration time.

2. Most Urgent Packets and Buffer Status

   As stated above, each packet in the sending buffer has an expiration time (except for background traffic classes). A scheduling algorithm should try its best to transmit each packet before it expires. In each frame, some packets will expire if not sent out in this frame. We name these packets the most urgent packets (MUPs) for the current frame. These packets should be sent out in this frame if possible, or else they would be dropped by the sender. Similarly, we can define the 2nd most urgent packets (2UPs), 3rd most urgent packets (3UPs), etc. To minimize packet drops, 2UPs should be served in each frame after all MUPs have been served, followed by 3UPs, and so on. The BS knows the buffer status of each user, i.e., the numbers of MUPs, 2UPs, 3UPs, ..., from each user in any given frame.

3. Processed Packets, Dropped Packets and PDR

   We know that, in each frame, a MUP is dropped at the sending station if it is not sent. We define the processed packets in a given time period T as the packets sent out or dropped by the sending stations in T. Note that a processed packet may not be a MUP in period T. A packet that is not so urgent may still be sent in T if capacity is available. The PDR

of class j user i in period T is defined as:

$$PDR_{j,i}^{T} = \frac{DP_{j,i}^{T}}{PP_{j,i}^{T}} \tag{3.4}$$

where $DP_{j,i}^{T}$ and $PP_{j,i}^{T}$ are, respectively, the numbers of dropped and processed packets at the user station in the period T. During each scheduling instance, the system should consider not only the packet drops that have happened in previous frames, but also those that will happen in the next frame due to failure to schedule MUPs for transmissions in the next frame.

4. Error Packets and PER

We define the error packets in a given time period T as the packets which are sent in period T but not correctly received. The PER experienced by class j user i in period T is:

$$PER_{j,i}^{T} = \frac{EP_{j,i}^{T}}{PP_{j,i}^{T} - DP_{j,i}^{T}} \tag{3.5}$$

where $EP_{j,i}^{T}$ is the number of error packets in period T. When the system schedules a packet, to fairly distribute the PLR, it must guess whether the packet will be correctly received. The probability that a packet will be correctly received depends on how many packets are allocated in the slots from what stations and the power constraints on these transmissions. For the system introduced in Section 3.3, it can be derived based on (3.3). After a transmission, we assume that the BS knows for sure if the transmitted packet is correctly received or not via some error detection coding such as cyclic redundancy check.

5. Lost Packets and PLR

We define lost packets to include both packets dropped by the transmitter due to delay violations, and packets sent but not received correctly. So, we define the PLR of class j user i in a given period T as:

$$PLR_{j,i}^{T} = \frac{LP_{j,i}^{T}}{PP_{j,i}^{T}} = \frac{DP_{j,i}^{T} + EP_{j,i}^{T}}{PP_{j,i}^{T}} \tag{3.6}$$

where $LP_{j,i}^{T}$ is the number of lost packets from the user station in the period T. It is also given by:

$$PLR_{j,i}^{T} = 1 - (1 - PDR_{j,i}^{T})(1 - PER_{j,i}^{T}) = PDR_{j,i}^{T} + PER_{j,i}^{T} - PDR_{j,i}^{T} \times PER_{j,i}^{T} \tag{3.7}$$

3.3.2 Basic Concept of the Cross-Layer Enhancement

To explain the basic concept of the cross-layer enhancement, we first consider only one user in the system and extend to multiple users later. We consider a scheduling algorithm which process packets one by one like the approaches used in [1–3].

In most existing algorithms, there is a PER or BER threshold for different flows or users. These scheduling algorithms will not admit more packets if adding one more packet into the frame will lead to the violation of the PER requirements of any user in the frame. This kind of mechanism is efficient for transmitting non-urgent packets or non delay sensitive traffic because the throughput is maximized and the PLR of the frame is guaranteed (PER is effectively PLR here since no packet drop occurs). However, this approach is not efficient for transmitting urgent packets or delay sensitive traffic as it may lead to a greater number of packet drops when there are a lot of MUPs or the traffic source rate is very high. Actually, if there are still MUPs in the buffer, by degrading the PER requirements of some packets, more packets can be

transmitted in one frame and thus the overall number of packets losses due to errors and drops in the frame may be reduced. This is the basis of the cross-layer-based scheduling method for delay sensitive multimedia traffic proposed in this chapter. We consider two cases of the buffer status in a frame.

The first case is that there are not too many MUPs and all of them can be transmitted in the frame. We denote this case "all MUPs transmitted" (MAT). Note that, adding lower priority packets (2UP, 3UP, ...) may increase the total packet losses in the frame because the PER may increase while no packet will be dropped. So the algorithm should try to maximize the throughout by keeping the PER reasonably low for each packet. In this case, our algorithm works in the same manner as the existing algorithms, which allows other packets (2UPs, 3UPs, ...) to be scheduled if and only if a predefined PER threshold is not exceeded. Note that, in MAT, PLR for the frame is contributed solely by the PER. We define this threshold as the *PLR threshold* in this chapter since it reflects the capability of a specific flow to tolerate packet losses.

The second case is denoted as $\overline{MAT}$, where there are too many MUPs in a frame so that some packets must be dropped if the PER is kept below the threshold for every transmitted packet. In this case, the PLR of the frame cannot be kept below the threshold. So, the algorithm should try to relax the PER of the frame so that the total PLR for the frame can be minimized. Based on the number of MUPs in the buffer, the scheduling algorithm chooses an appropriate number of MUPs to be transmitted in the current frame to reduce the overall PLR (considering both packet errors and packet drops) of the current frame. Specifically, the scheduling algorithm decides whether the next MUP in the buffer should be added into the frame by checking the expected PLR of the current frame. In $\overline{MAT}$, Packets other than MUPs (i.e., 2UPs, 3UPs, ...)

will not be transmitted.

Since MUPs are always scheduled before other packets, if all MUPs can be scheduled while keeping the PER below the threshold, this belongs in the MAT case; otherwise the $\overline{MAT}$ case.

For a multi-user system, the approach is similar. However, unlike a single user system where we can just minimize the total PLR of a frame, we now need to consider the PLR requirements of different users, whose flows may have different PLR targets. So, in a multi-user system, we optimize the *weighted excess PLR* instead of the total PLR of the frame. In this case, the optimization problem is constrained by the fairness requirements specified below.

3.3.3 Fairness Requirement

In a multi-user system, we define ϕ_j as the PLR threshold of class j users, similar to the PER threshold defined in previous work. We also call these the *target PLRs*. The *target PLR*ϕ_j should be satisfied for all class j users in every frame whenever possible. If there are too many MUPs in one frame (i.e., the $\overline{MAT}$ case), the *target PLR*ϕ_j may not be satisfied in that particular frame for all users, and PLR becomes excessive for some users.

Most previous work on fair scheduling aims to fairly distribute the bandwidth among users or flows. However, this approach may not be suitable for multimedia flows that have variable rates and stringent delay requirements. Instead, in this chapter we propose to provide fair scheduling via fair distribution of excess PLR (FDEL) among all flows, which is similar to the approaches in [2, 3] where different users or flows share the PDR. We show below that FDEL is equivalent to fairly distributing throughput degradations among different flows when there is insufficient link capacity, due to channel or traffic conditions.

In this chapter, we define the *reference period* (RP) as a series of frames on which FDEL

must try to guarantee the fairness among all flows. Its length is decided by the system (or the system operator). If the length of the RP is one frame, FDEL will distribute the current frame's excess PLR for each frame. If the length of RP is n frames, in every frame, FDEL will try to distribute the excess PLR among the previous $n-1$ frames and the current fame. The length of the RP can also be the whole lifetime of the system (from the start of the system to the current frame), which means that FDEL will try to distribute the life time excess PLR (the excess PLR from the beginning of a flow to the current frame) of all active flows. Such a design gives the system the flexibility to fit different practical scenarios. Given a frame t ($t \in \{1, 2, 3, \}$), we denote the RP for the frame t as $T_{RP,t}$, which represent the RP ending on the current frame. Consider its length is L_{RP}. So $T_{RP,t}$ is from the beginning of frame $t - L_{RP} + 1$ to the end of frame t. Then the excess PLR of class j user k during the RP is $(PLR_{j,k}^{T_{RP,t}} - \phi_j)^+$. Here $(.)^+$ means that if the value of (.) is negative it is converted to zero.

When distributing the excessive PLR, FDEL also considers the priorities of different flows. This is accomplished by defining a *fairness weight* (FW) for each flow, which is decided by the system (or the system operator) when the flow is admitted into the system. The FW of flow k of class j is denoted $\gamma_{j,k}$.

We define the *fairness factor* (FF) $FF_{j,k}$ as the weighted excess PLR of user k of class j during the RP.

$$FF_{j,k} = \frac{(PLR_{j,k}^{T_{RP,t}} - \phi_j)^+}{\gamma_{j,k}}, j \in 1...J, k \in 1, ..., K_j \tag{3.8}$$

In FEDL, fairness is satisfied if:

$$FF_{j,k} = FF_{i,l}, \forall i, j \in 1...J, k \in 1...K_j, l \in 1...K_i \tag{3.9}$$

Though we define FF as the weighted excess PLR, it also reflects the throughput degradation

of each flow as follows.

$$\begin{aligned} FF_{j,k} &= \frac{(PLR_{j,k}^{T_{RP,t}} - \phi_j)^+}{\gamma_{j,k}} \\ &= \frac{1}{\gamma_j}((1-\phi_j) - (1 - PLR_{j,k}^{T_{RP,t}}))^+ \\ &= \frac{1}{\gamma_j PP_{j,k}^{T_{RP,t}}}((1-\phi_j)PP_{j,k}^{T_{RP,j}} - (LP_{j,k}^{T_{RP,j}}))^+ \end{aligned} \tag{3.10}$$

Here $(1-\phi_j)PP_{j,k}^{T_{RP,j}}$ is the number of packets expected to be transmitted during $T_{RP,t}$ since there are $PP_{j,k}^{T_{RP,j}}$ packets processed and the flow can tolerate PLR up to ϕ_j. Also, $PP_{j,k}^{T_{RP,j}} - LP_{j,k}^{T_{RP,j}}$ is the number of packets successfully transmitted among $PP_{j,k}^{T_{RP,j}}$ packets. Essentially, FF is equivalent to the weighted throughput degradation ratio. Note that FW here determines how traffic flows can share the excess PLR. So it should not be designed based on the PLR requirements directly. Instead, it should be determined based on how important the PLR requirements should be satisfied for the traffic flows.

Equation (3.9) gives the fairness constraint. It can be directly applied to the scheduling of MUPs-to calculate of FF over the RP ending at the current frame. However, in the case of MAT, to schedule 2UPs, 3UPs, etc., which do not contribute to the PLRs of the RP ending at the current frame, we must predict the FF of future frames. For example, 2UPs are scheduled based on the estimated FF for the RP ending at the next frame, since 2UPs become MUPs in the next frame.

Scheduling with FDEL has some advantages. Because of the rate variations of multimedia traffic, the number of MUPs may vary from frame to frame. So a good scheduling algorithm must have the ability to adapt to the variations. FDEL, which is equivalent to fairly sharing the throughput degradation ratio, can automatically provide more bandwidth in one frame to users with more MUPs to balance the excessive PLR. The fairness constraint (3.9) also provides

the flexibility of distributing excess PLR or throughout degradation among different classes of users with different QoS requirements. Note that effectively, the FWs γ represent the priorities of the users. This is because a flow with a lower weight will have a better chance to transmit the MUPs and therefore will result in a lower excess PLR and lower throughput degradation ratio.

3.3.4 Objective Functions

In this subsection we formally define the objective function of the cross-layer enhancement method proposed in this chapter. Since scheduling is performed frame by frame, the objective function is updated every frame. As discussed before, we consider scheduling for two different cases of the buffer states in a frame, namely MAT and $\overline{MAT}$.

The scheduling task is straightforward in the MAT case. The algorithm should maximize the throughput of the frame while meeting the PLR threshold for each flow in the current frame (note that it may not be possible to guaranteed the PLR threshold in the whole RP) and satisfying the FDEL constraint. Assuming that the current frame is frame t, the objective function in this case is:

$$\max_{PP_{j,k}^t} \{\sum_{j=1}^{J} \sum_{k=1}^{K_j} PP_{j,k}^t\} \tag{3.11}$$

$$s.t. \quad PLR_{j,k}^t \leq \phi_j,$$

$$\frac{(PLR_{j,k}^{T_{RP,t+x}} - \phi_j)^+}{\gamma_{j,k}} = \frac{(PLR_{i,l}^{T_{RP,t+x}} - \phi_i)^+}{\gamma_{i,l}},$$

$$\forall i, j \in 1...J, k \in 1...K_j, l \in 1...K_i, x \in \{0, 1, 2, 3...\}$$

The first constraint makes sure that the PLR thresholds of all flows are satisfied. Note that $PLR_{j,k}^t$ is the PLR of the flow k of class j for the current frame. It is also effectively the same as

$PER_{j,k}^{t}$ since no packet is dropped in frame t. The second constraint is the fairness constraint. Here $T_{RP,t+x}$ is the RP until the end of frame $t+x$. Note that as described in the last subsection we need to predict FF of the RP ending at frame $t+x$, in order to fairly schedule 2UPs, 3UPs and so on. CEPS chooses the number of packets from the current flows as the ones scheduled for the current frame. It maximizes the throughput by choosing as many packets as possible, while satisfying the above constraints.

The other case $\overline{MAT}$ is that there are too many MUPs in a frame so that the scheduler must determine how many and which of them should be transmitted in the frame. In a single user system, we seek to minimize the PLR of the current frame if there are too many MUPs in a frame. Similarly, in a multi-user system, the algorithm aims to ensure that the sum of FFs (weighted excess PLRs) of all flows can be minimized for the *RP*ending at the current frame. It is equivalent to minimizing the sum of all weighted throughout degradation ratios. The objective function in this case is:

$$\min_{PP_{j,k}^{t}} \{\sum_{j=1}^{J}\sum_{k=1}^{K_j} \frac{(PLR_{j,k}^{T_{RP,t}} - \phi_j)^+}{\gamma_{j,k}}\} \tag{3.12}$$

$$s.t. \quad \frac{(PLR_{j,k}^{T_{RP,t}} - \phi_j)^+}{\gamma_{j,k}} = \frac{(PLR_{i,l}^{T_{RP,t}} - \phi_i)^+}{\gamma_{i,l}},$$

$$\forall i,j \in 1...J, k \in 1...K_j, l \in 1...K_i$$

The constraint in (3.12) ensures fairness among all flows. CEPS also chooses the number of packets from current flows as the ones scheduled for the current frame. The numbers are decided to minimize the sum of FFs (weighted excess PLRs) for the RP. Note that 2UPs, 3UPs and so on are not scheduled in this case. So we only need to consider the fairness constraints with respect to the current frame.

Using the above objective functions, the overall PLR can be optimized (i.e., instead of optimizing PER and PDR separately) subject to the FDEL constraint. Effectively, the algorithm considers both the physical layer situation and status of the MAC layer queues.

3.4 The Proposed CEPS Algorithm

3.4.1 Basic Principles

With the fairness constraint and objective functions provided by (3.11) and (3.12) in the previous section, we discuss how they can be realized in a practical system in this section.

Because of the discrete nature of packet-switched systems, it is obvious that the fairness constraints in (3.11) and (3.12) cannot hold perfectly. In practical systems, the proposed CEPS algorithm schedules packets one by one like other algorithms for packet-switched networks [1–3]. Specifically, it selects a packet from the flow with the largest FF (weighted excess PLR) so that the fairness constraints can be met as much as possible. Note that, to transmit one more MUP, the FF (weighted excess PLR) will decrease a little bit. In this way, it can try the best to ensure the FFs of all flows are as close as possible.

At the beginning of every frame, the algorithm does not know whether there are too many MUPs. While FFs (or predicted FFs) of all flows determine which packets will be served, the algorithm simply adds packets as long as the PER of each user is below its PLR threshold. Note that for both MAT and $\overline{MAT}$ cases, the scheduling does the same thing as above. If all packets can be scheduled in this way, the work is done. If the PER of any user reaches to its PLR threshold during the process, the algorithm looks at the buffer to see whether there are still MUPs not scheduled. If there are no MUPs in the buffer, it can be concluded that the

case is MAT, and the scheduling should stop here as the objective for MAT is already achieved. If there are still MUPs in the buffer, it is $\overline{MAT}$ and further work need to be done for the cross-layer optimization. For case $\overline{MAT}$, the algorithm then schedules packets one by one and see whether the sum of all weighted excess PLR is decreasing. Scheduling stops when the sum of all FFs (weighted excess PLRs) stops decreasing (and start increasing). In this way, the minimum sum of FFs (weighted excess PLRs) is found.

Based on the basic principles given above, we describe the CEPS algorithm in details as follows. Our focus is to support three types of traffic: conversational/voice, streaming video and interactive data. Note that background data traffic can also be easily supported using the residual capacity. In the proposed CEPS algorithm for MC-CDMA networks, scheduling a packet consists of three stages. In the first stage, which user should be served next is decided according to the fairness constraint in the objective functions in Section 3.3.4, and one of its packet is selected out. In the second stage, the selected packet is put into a specific slot in the frame, which aims to maximize the residual bandwidth and maximize the multi-code benefit. In the third stage, the algorithm determines whether to end the scheduling by checking the sum of weighted excess PLR in the case of $\overline{MAT}$, which is based on the discussion in the optimization objective in (3.12) in Section 3.3.4. Figure 3.1 shows the flow chart of the proposed CEPS algorithm. As shown by simulation results presented later, the system performance can be greatly enhanced by CEPS.

To facilitate the description, we make some more definitions of the system states at the beginning of a frame and the QoS requirements of different traffic classes here. For user k of traffic class i, we assume that the numbers of processed and lost packets in the previous s frames are $\rho_{i,k,s}$ and $\zeta_{i,k,s}$, respectively. The numbers of MUPs, 2UPs, 3UPs, , for this user in

the current frame are $m_{i,k}^{(1)}$, $m_{i,k}^{(2)}$, $m_{i,k}^{(3)}$, and so on. The PLR requirement and the FW of a class i user k are ϕ_i and $\gamma_{i,k}$, respectively.

3.4.2 Stage 1

In stage 1, an appropriate user is chosen to be served next and one of its packet is selected out. It is obvious that MUPs must be served before other packets, as they will be dropped if not transmitted in the current frame while other packets can wait to be sent out in the next frame. Similarly, 2UPs must be served next, followed by 3UPs, and so on.

As we discussed in Section 3.4.1, the algorithm will choose the user with the largest FF (weighted excess PLR) to serve next, among all users with MUPs (or 2UPs if there are no MUPs, 3UPs if there are no 2UPs, and so on) in the buffer. This ensures that the fairness constraint is satisfied as much as possible. According to the definitions of the FF in (3.8) and (3.10), the algorithm calculates the FFs as follows.

1. While scheduling MUPs, the FF of user k in class i is:

$$_{i,k} = \frac{1}{\gamma_{i,k}} \left(\frac{\zeta_{i,k,L_{RP}-1} + m_{i,k}^{(1)} - z_{i,k} + \eta_{i,k}(z_{i,k})}{\rho_{i,k,L_{RP}-1} + m_{i,k}^{(1)}} - \phi_i \right)^+ \tag{3.13}$$

   Here $z_{i,k}$ is the number of scheduled MUPs of the user and $\eta_{i,k}(_{zi,k})$ is the expected number of packet errors for the $z_{i,k}$ scheduled packets. The term $\rho_{i,k,L_{RP}-1} + m_{i,k}^{(1)}$ is the expected number of processed packets in $T_{RP,t}$, and $\zeta_{i,k,L_{RP}-1} + m_{i,k}^{(1)} - z_{i,k} + \eta_{i,k}(z_{i,k})$ is the expected number of lost packets in $T_{RP,t}$. Essentially, (3.13) gives the FFs at the end of this frame if the scheduling is ended at this point. Note that here $\eta_{i,k}(_{zi,k})$ depends on

all packets which have already been scheduled in this frame, how the packets are put into slots and also the physical layer channel status.

2. While scheduling 2UPs, the predicted FF is:

$$_{i,k} = \frac{1}{\gamma_{i,k}} \left(\frac{\zeta_{i,k,L_{RP}-1} + m_{i,k}^{(2)} - z_{i,k}^{2} + \eta_{i,k}(m_{i,k}^{1} + z_{i,k}^{2})}{\rho_{i,k,L_{RP}-2} + m_{i,k}^{(1)} + m_{i,k}^{(2)}} - \phi_i \right)^{+} \tag{3.14}$$

   Here $z_{i,k}^{(2)}$ is the number of 2UPs already scheduled for the user in the current frame and $\eta_{i,k}(m_{i,k}^{(1)} + z_{i,k}^{(2)})$ is the corresponding expected number of packet errors. Note that, $m_{i,k}^{(1)}$ MUPs must have been scheduled. The processed and lost packets in $T_{RP,t+1}$ are $\rho_{i,k,L_{RP}-2} + m_{i,k}^{(1)} + m_{i,k}^{(2)}$ and $\zeta_{i,k,L_{RP}-2} + m_{i,k}^{(2)} - z_{i,k}^{(2)} + \eta_{i,k}(m_{i,k}^{(1)} + z_{i,k}^{(2)})$, respectively. Although there are no packet drops in the frame (since all MUPs have been scheduled already), we consider the packets that will be dropped in the next frame if the scheduling stop at this point and no further packet scheduling occurs in the next frame.

3. The FF is similarly defined by modifying (3.14) for scheduling 3UPs, 4UPs, ...

In stage 1, FFs are used to determine the service sequences. However, they are calculated in the stage 2 since some information needed for calculation can only be available in stage 2.

3.4.3 Stage 2

In stage 2, CEPS algorithm puts packet into slots wisely maximize the residual bandwidth after the target PLRs for all packets scheduled are satisfied so that as many packets with lower urgency can be transmitted as possible. Of course, a random assignment that does not optimize anything like in some previous works [2, 3, 21], can also be used here.

Since packets from the same user uses orthogonal codes, the algorithm always tries its best to assign multiple packets from the same user in the same time slot to maximize the system

performance. CEPS maintains a counter and an array for each user. The counter stores the number of packets being selected from the buffer of the user's terminal and the array records the slot positions of the packets as scheduled for transmissions in the frame. Every time a packet is selected in stage 1, the counter for the corresponding user increases by one. CEPS then tries to put the packet into some slot with packets from the same user.

If after the packet added to the frame, all packets scheduled can still keep their PLR threshold, the scheduling process returns to the stage 1 directly. However, if all slots are "filled up" in the sense that adding one more packet to the slot will lead to the PERs of some packets to become higher than their corresponding target PLR, the scheduling process goes to stage 3 if there are still MUPs in the buffer (case $\overline{MAT}$), or the scheduling process ends directly (case MAT). In stage 3, whether or not this new packet is added to the frame is determined by whether the sum of FFs (weighted excess PLRs) decreases after the addition.

Optimal packet coordination is an NP-hard problem. The above process can find a relatively good solution. Note that FFs are needed to assist the operation. Given the transmission parameters of the packets, the expected PERs can be easily obtained from a table calculated from (3.1), (3.2) and (3.3) (or similar physical layer channel model). The FFs can be determined from (3.13) or (3.14), and then the sum of the FFs (weighted excess PLR) can be calculated. These metrics are also used by stage 1 and stage 3 to make decisions.

3.4.4 Stage 3

While the capacity of a frame is easy to calculate for traditional scheduling algorithm, it becomes an undetermined one for our cross-layer enhancement. Actually, the capacity depends on which packets we choose to send in the frame. So, according to the discussion in Section 3.3.4, CEPS

tries to minimize the sum of FFs (weighted excess PLR) for all users.

In stage 3, CEPS determines when to stop the scheduling more packets for transmissions in the current frame in case $\overline{MAT}$. Generally, one more packet can always be scheduled if the frame has not been "filled up" (as defined in the previous subsection). If the frame has been "filled up", a packet can be scheduled only if it is a MUP (in the $\overline{MAT}$ case) and the sum of FFs (weighted excess PLR) is decreased by scheduling the packet in some slot in the current frame. So, stage 3 is only needed when the frame is "filled up", i.e., only when scheduling a new packet violates the target PLR of some packets that have already been scheduled. Essentially, CEPS delays the packet scheduling to the next frame if the new packet is not a MUP. Otherwise it will attempt to schedule the packet and compare the resulting sum of FFs (weighted excess PLRs) with the previous value to see whether the system performance is enhanced. If not, the MUP is not scheduled but dropped instead. The dropping of a MUP indicates that the frame is truly full and then the scheduling algorithm ends.

3.5 Performance Evaluation

A simulation model was implemented in C++ to evaluate the performance of the proposed CEPS algorithm and to compare it with FPLS [3] and PSDRDG [2]. Note that FPLS was not originally designed for MC-CDMA systems. To help it take the advantage of multi-codes, the slot assignment mechanism proposed in Section 3.4.3 is used also with FPLS. Some common parameters for all simulations are listed in Table 3.1. In the simulations, the physical layer is applied statistically and the data link layer is operated step by step. The results are generated from the average of a couple of same simulations, which runs at least 100,000 TDMA frames.

As mentioned before, three classes of traffic are considered: conversational/voice traffic, streaming video traffic, and interactive data traffic. For the voice traffic, each traffic source is represented by the commonly used two-state (on/off) Markov model [23]. During the "on" or "talk-spurt" state, packets are generated at a constant rate of one packet per frame. During the "off" or "silence gap" state, no packet is generated. The durations of the "on" and "off" states are exponentially distributed with mean values of 1.0 s and 1.35 s, respectively. For the video traffic, the traffic source is modeled by a Markov-Modulated Poisson Process [24] with the stationary distribution of (0.2, 0.4, 0.3, 0.1). Packets are generated based on a Poisson process with the rates of (2, 4, 6, 8) packets per frame in the respective states. The data traffic source is modeled by the on-off model proposed in [23]. Each data traffic source alternates between "on" and "off" states. During the "on" state, messages are generated based on a Poisson distribution with the rate 0.5 messages per frame. Each message consists of 6 packets. No message is generated during the "off" state. The durations of the "on" and "off" states are varied based on the Weibull distribution. The shape and scale parameters are 0.88 and 3.10, respectively for the "on" state, and 0.88 and 21.4, respectively for the "off" state. The typical QoS parameters used in this chapter are based partly on [2] and are listed in Table 3.2. Also the length of the RP is set to 100 frames.

We first study the cross-layer performance enhancement achieved by CEPS. To fairly compare CEPS with FPLS and PSDRDG, we set the FWs according to the target PLRs. Figures 3.2-3.5 show the simulation results with 30 video users, 50 data users and the number of voice users varied from 70 to 145. Figure 3.2 shows the PLRs of different traffic classes. When the number of voice users is lower than 90, all three algorithms achieve the same PLR for each traffic class, because the PLR is dominated by the PER for all three algorithms (i.e., $PDR \approx 0$). How-

ever, when the number of voice users increases, Figure 3.2 shows that CEPS achieves a much lower PLR for each traffic class, and it can support more than 125 voice users with $PLR < 0.01$, whereas the other two algorithms can only support about 90 voice users. The results clearly demonstrate the effectiveness of the cross-layer enhancement of CEPS in system capacity (about 39%). Figure 3.3 compares the mean packet delay for all traffic classes. Note that the mean delay > 0.01 (half a frame) because a packet must wait on average at least half a frame before the start of the frame in which it can be transmitted. Also, the mean delay is always lower than the delay bound by at least one frame because a packet must be sent out before it expires (i.e., a packet expiring in frame x must be sent in or before frame $x - 1$). It can be seen that for video and data traffic, which has a delay bound of more than one frames, CEPS can achieve a lower mean delay when the number of voice users is moderate. In the case of voice traffic, the mean delays for the three algorithms are similar because all packets have to be transmitted within one frame or else they will be dropped. Figure 3.4 confirms that the maximum delay experienced by each user is lower than the required delay bound. The results also reveal that for video traffic and data traffic, CEPS can achieve a lower maximum delay when the number of voice users is small. Figure 3.5 shows the throughput of different traffic classes. It can be seen that when the number of voice users increases, the voice throughput increases while the video throughput decreases because some resources are allocated to the additional voice users. When the number of voice users is larger than 85 (i.e., the system is heavily loaded and there are some packets dropped, see Figure 3.2), CEPS yields a higher throughput for voice and video traffic than the other two algorithms. The data traffic throughput for all three algorithms are similar. Because of its very low target PLR and smaller FW, data traffic packets are rarely dropped. Thus the throughput variation (caused by the variation of the PLR) on data traffic

is too small to be observed. Figures 3.6-3.8, respectively, show the PLR, average access delay and throughput performance of the system with 100 voice users, 50 data users and the number of video users varying from 21 to 36. In general, the performance is similar to that presented above (the capacity of the system is increased by 22%). Similar to the previous scenario, these figures show that CEPS can achieve a better performance than the other two algorithms.

Next, we show the benefits of using cross-layer optimization when the channel condition or $\frac{1}{\sigma^2}$ varies. Note that $\frac{1}{\sigma^2}$ gives the reciprocal of the normalized noise variance (including the inter-cell interference). That means, the larger the value of $\frac{1}{\sigma^2}$, the better the channel condition is and the larger the SINR is as determined by (3.1). Figures 3.9-3.11 show the performance of the system with 100 voice users, 30 video users and 50 data users, while $\frac{1}{\sigma^2}$ varies from 9dB to 10.4dB. Figure 3.9 shows that as $\frac{1}{\sigma^2}$ increases, the PLRs of all traffic classes decrease; however, CEPS clearly achieves better PLR performance than the other algorithms for each traffic class. Note that there is a sharp decrease of PLR between $\frac{1}{\sigma^2} = 9.8dB$ and $\frac{1}{\sigma^2} = 9.9dB$. With $\frac{1}{\sigma^2} \leq 9.8dB$, a video or voice packet can tolerate at most 8 interfering packets but with $\frac{1}{\sigma^2} \geq 9.9dB$, this number is increased to 9. Thus, the system capacity is increased significantly and the PLRs of all traffic classes decrease dramatically. There is another sharp decrease for the PSDRDG and the FPLS between $\frac{1}{\sigma^2} = 10.2dB$ and $\frac{1}{\sigma^2} = 10.3dB$. However, it has almost no impact to CEPS. With $\frac{1}{\sigma^2} = 10.2dB$, CEPS can already serve all traffic classes with almost no packet drop. Thus a further increase in system capacity at that point does not benefit CEPS much. Figure 3.10 shows that as $\frac{1}{\sigma^2}$ decreases, the mean packet delay of each traffic class decreases also, as expected. As explained previously, the mean delay of voice traffic for the three algorithms are similar. For other traffic classes, for $\frac{1}{\sigma^2}$ from 9dB to 9.8dB, all three algorithms have almost the same mean packet delay because the system is a bit overloaded such that almost

all packets must wait until the deadline to be sent out. From 9.9dB to 10.2dB, CEPS exhibits a lower mean packet delay than the other two algorithms because of cross-layer enhancement. Between 10.3dB and 10.4dB, the mean packet delays for the three algorithms are again similar because the system capacity is large enough so that almost all packets can be served within one frame. Figure 3.11 shows the throughput of each traffic class under different channel conditions. It can be seen that CEPS can achieve a higher throughput for video traffic and voice traffic than the other two algorithms especially when $\frac{1}{\sigma^2}$ is small. When $\frac{1}{\sigma^2}$ increases to 10.3dB, the throughput enhancement disappears because nearly all packets can be transmitted due to very good channel condition. As explained earlier, the throughput for the data traffic are almost the same for the three algorithms because of the very low PLR target.

In summary, the above simulation results show the great benefits of the cross-layer enhancement in CEPS. The PLRs of different traffic classes are reduced so that the system can support more users. Furthermore, the mean packet delay is reduced and the throughput is increased under most conditions. The simulation results also show that CEPS has a better ability to cope with the variations of noise and inter-cell interference. It can meet the target PLR over a wider range of $\frac{1}{\sigma^2}$ than the other two algorithms.

Next, we show the effectiveness of CEPS in providing fairness. We first compare all three algorithms with the target PLR set according to (i.e., proportional to) the FW. Figure 3.12 shows the weighted excess PLR of different traffic classes for the three algorithms when there are 100 voice users, 50 data users and different number of video users. The weighted excess PLRs are all zero when the number of video users is small. However, they go up when there are more video users. It can be seen that only CEPS can achieve the desirable result of having approximately the same weighted excess PLR for all traffic classes. The other two algorithms

have different weighted excess PLRs for different traffic classes when the number of video users is not so large, because FPLS or PSDRDG only proportionally distribute the PDR to the different traffic classes without considerations on the different PERs. Figure 3.13 shows the weighted excess PLR when there are 30 video users and 50 data users, and the number of voice users is varied between 70 and 160. It shows a similar result to that presented in Figure 3.12. Hence, with the proposed cross-layer optimization framework, CEPS can achieve better fairness. Furthermore, from these two figures, we can also observe that the weighted excess PLR of CEPS is much lower than those of FPLS and PSDRDG.

To further illustrate the flexibility of CEPS, we use a different set of QoS parameters for the three traffic classes as listed in Table 3.3. In this scenario, the FWs are not set according to the target PLRs. Figure 3.14 shows the weighted excess PLR of all traffic classes when there are 100 voice users, 50 data users and different number of video users. Note that FPLS always proportionally distributes the PDR to different users with respect to their target PLRs and PSDRDG always proportionally distributes the PDR to different users with respect to their FWs. It can be seen that for FPLS, the weighted excess PLRs for different traffic classes diverge substantially, while PSDRDG can provide the same weighted excess PLRs for different traffic classes when the number of video users is large. However, when the number of video users is not so large, say 28, the target PLR of voice and video traffic is satisfied while that of data traffic not; so it weakens the fairness. Compared with these two algorithms, CEPS meets the target PLR of different traffic classes while keeping the weighted excess PLRs the same. Figure 3.15 shows a similar result with 30 video users, 50 data users and different number of voice users. It can also be seen that FPLS has diverged weighted excess PLR and PSDRDG cannot meet the target PLR of all traffic classes at the same time. Again, CEPS outperform the other

algorithms by meeting the two aforementioned requirements.

3.6 Summary

In this chapter, we have presented a cross-layer optimization for scheduling multimedia traffic in MC-CDMA networks. Based on it, we proposed a novel uplink scheduling algorithm called CEPS. CEPS can also be applied to downlink transmission with a few revisions. The algorithm has two new contributions. First, it jointly optimizes link layer PDR and physical layer PER to improve the overall PLR and hence the system performance. Second, it distributes the excess packet losses according to predefined weights so as to maintain fairness more effectively and flexibly. We have performed extensive simulations to compare the performance of the proposed CEPS algorithm with two previously proposed scheduling algorithms that proportionally distribute excess PDR. The simulation results show that the CEPS algorithm can greatly decrease the PLR experienced by users especially when the traffic demand is heavy, and support more users in the system while meeting a specified PLR objective. The simulation results also show that CEPS is also effective and more flexible in maintaining fairness among different users. Note that the superior performance of CEPS is because of its cross-layer design and it does not relax the fairness target to achieve better system performance.

Table 3.1: Parameters for The System

Parameters	Values
Processing Gain	32
Normalized noise variance	0.1
Packet size	511 bits
Payload size	229 bits
Tolerable bit errors	38 bits
Time slot per frame	20

Table 3.2: Typical QoS Parameters Setting (Setting 1) for Different Traffic Classes in This Chapter.

Parameters	Values
Voice target PLR	10^{-2}
Video target PLR	10^{-2}
Data target PLR	10^{-5}
Voice delay bound in frames	2
Video delay bound in frames	20
Data delay bound in frames	20
Voice FW	1
Video FW	1
Data FW	0.001

Table 3.3: Another QoS Parameters Setting (Setting 2) for Different Traffic Classes in This Chapter.

Parameters	Values
Voice target PLR	2×10^{-2}
Video target PLR	10^{-2}
Data target PLR	10^{-5}
Voice delay bound in frames	2
Video delay bound in frames	20
Data delay bound in frames	20
Voice FW	1
Video FW	2
Data FW	0.1

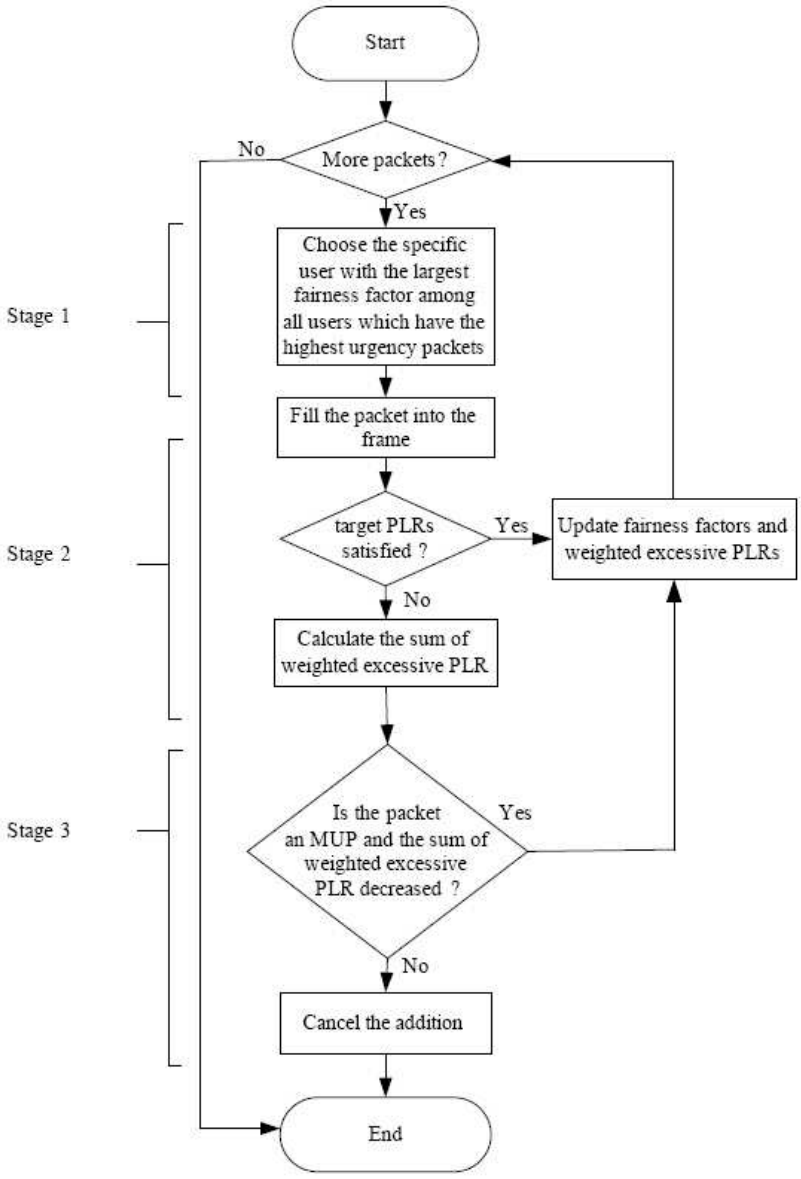

Figure 3.1: Flow chart of the proposed CEPS algorithm.

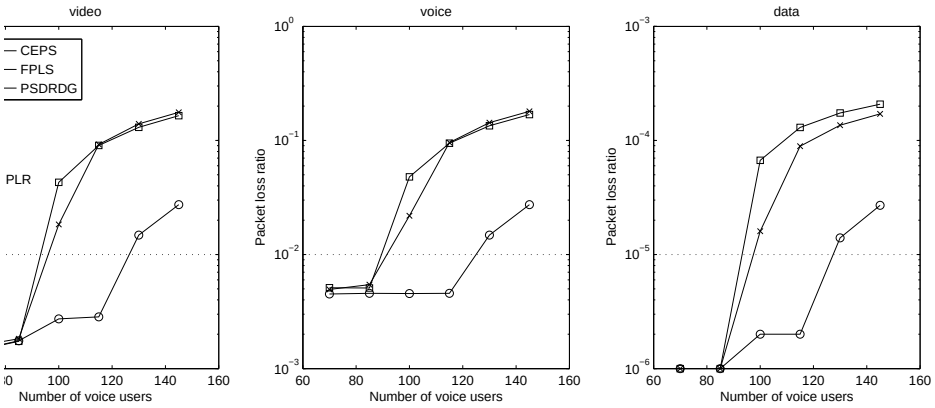

Figure 3.2: Packet loss ratio for different traffic classes when there are 30 video users, 50 data users and variable number of voice users.

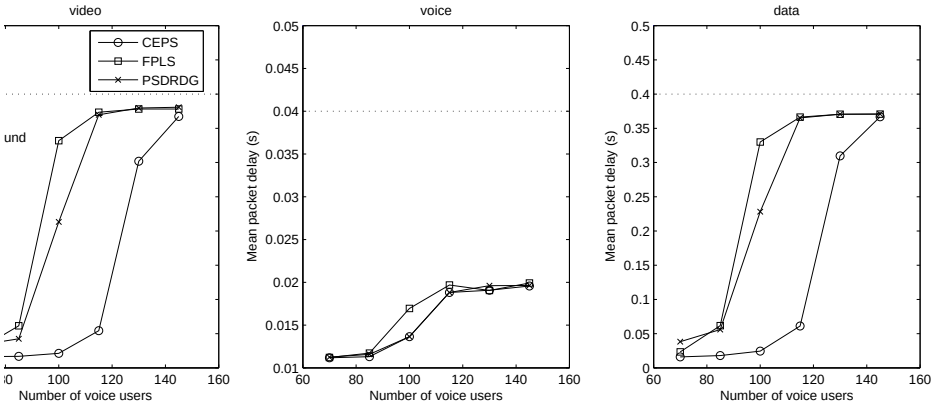

Figure 3.3: Mean packet delay for different traffic classes when there are 30 video users, 50 data users and variable number of voice users.

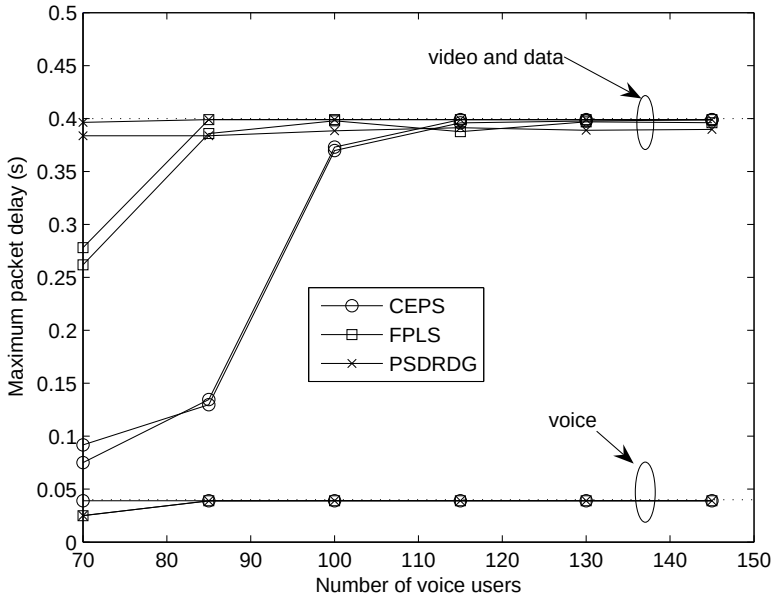

Figure 3.4: Maximum packet delay for different traffic classes when there are 30 video users, 50 data users and variable number of voice users.

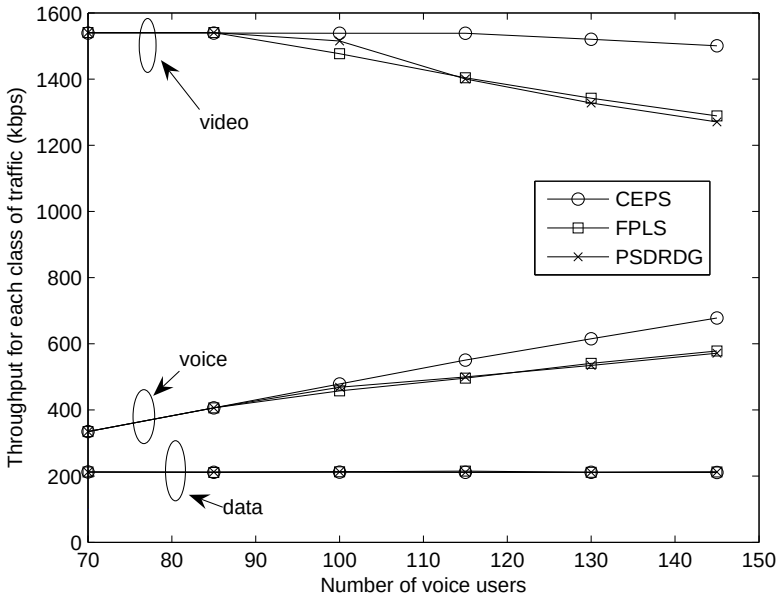

Figure 3.5: Throughput for different traffic classes when there are 30 video users, 50 data users and variable number of voice users.

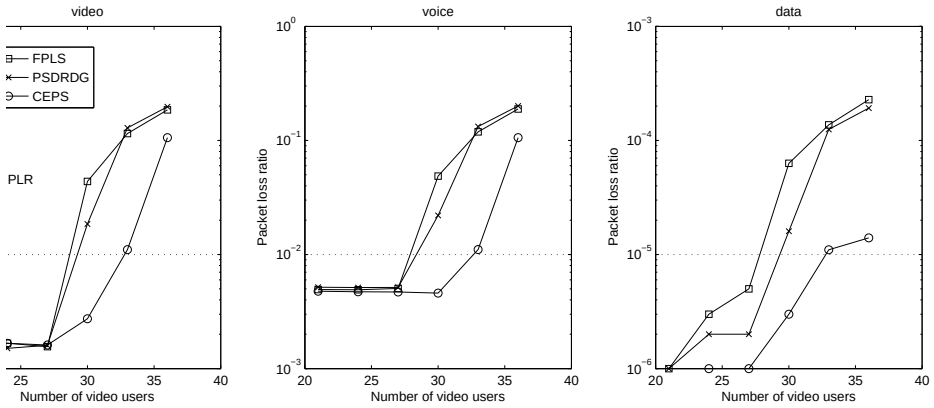

Figure 3.6: Packet loss ratio for different traffic classes when there are 100 voice users, 50 data users and variable number of video users.

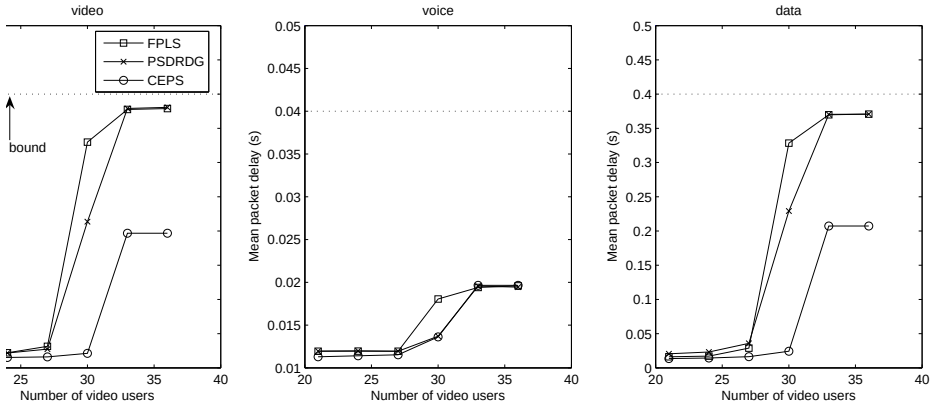

Figure 3.7: Mean packet delay for different traffic classes when there are 100 voice users, 50 data users and variable number of video users.

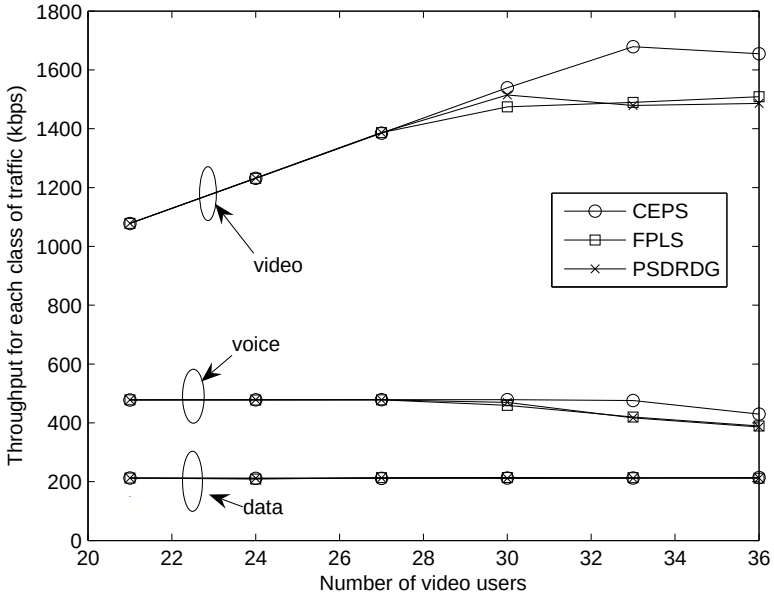

Figure 3.8: Throughput for different traffic classes when there are 100 voice users, 50 data users and variable number of video users.

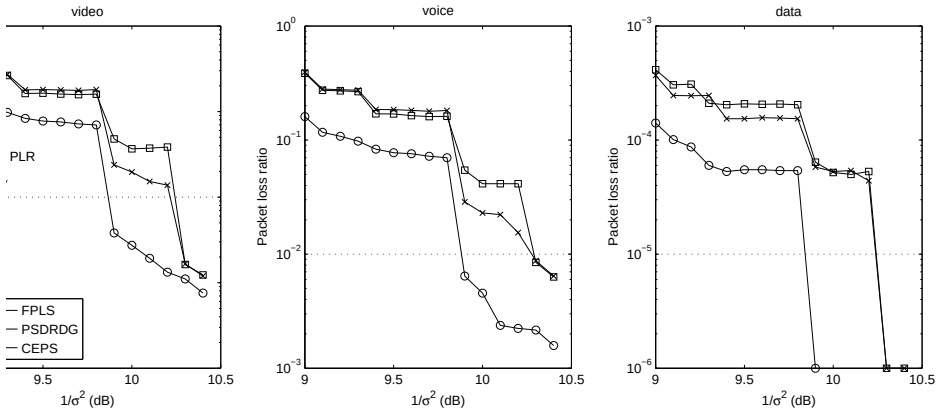

Figure 3.9: Packet loss ratio for different traffic classes with 100 voice users, 30 video users and 50 data users.

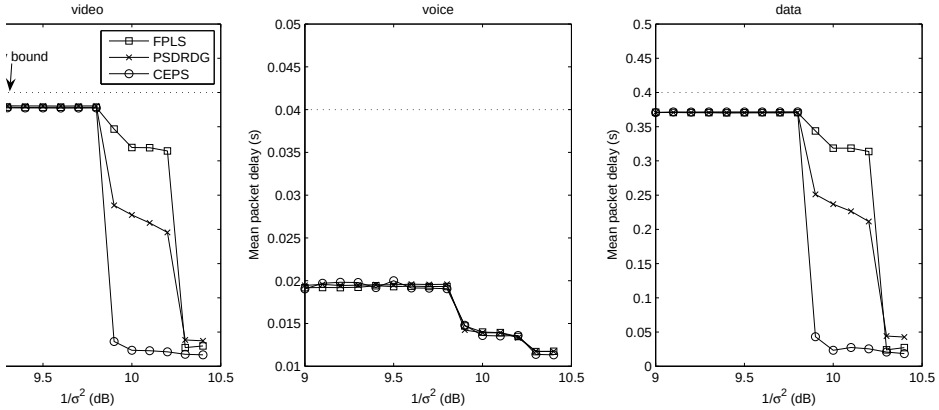

Figure 3.10: Mean packet delay for different traffic classes with 100 voice users, 30 video users and 50 data users.

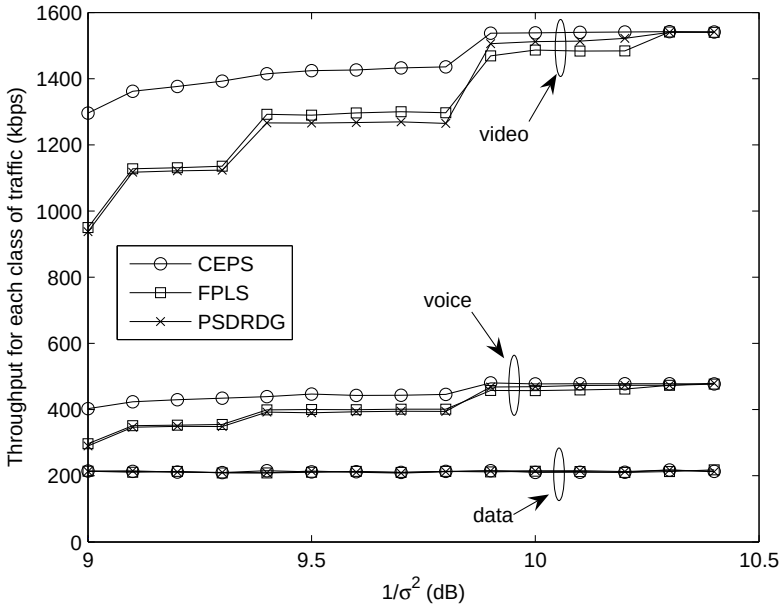

Figure 3.11: Throughput for different traffic classes with 100 voice users, 30 video users and 50 data users.

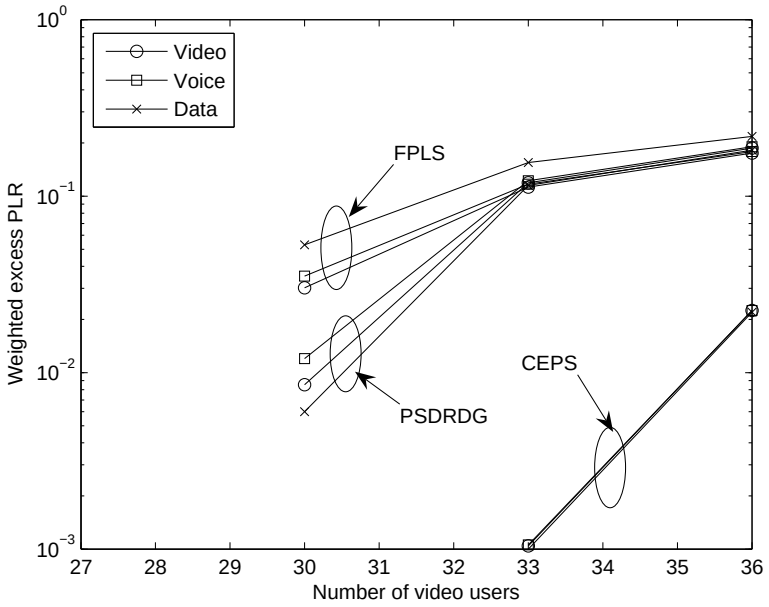

Figure 3.12: Weighted excess PLR of different traffic classes with 100 voice users, 50 data users and variable number of video users.

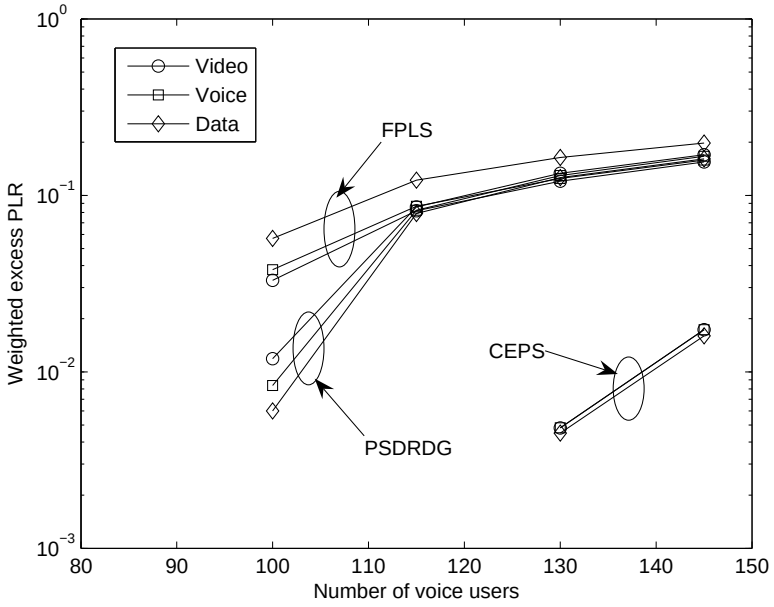

Figure 3.13: Weighted excess PLR of different traffic classes with 30 video users, 50 data users and variable number of voice users.

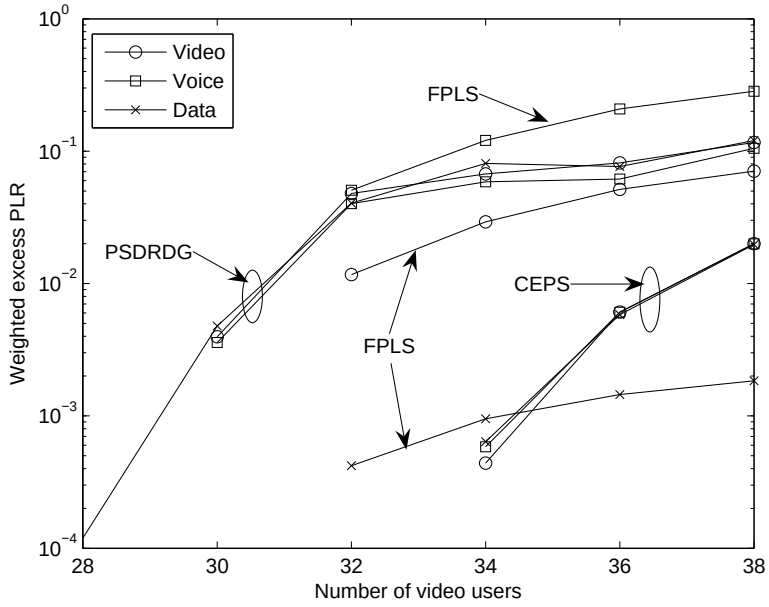

Figure 3.14: Weighted excess PLR of different traffic classes with 100 voice users, 50 data users, variable number of video users and QoS parameters in Table 3.3.

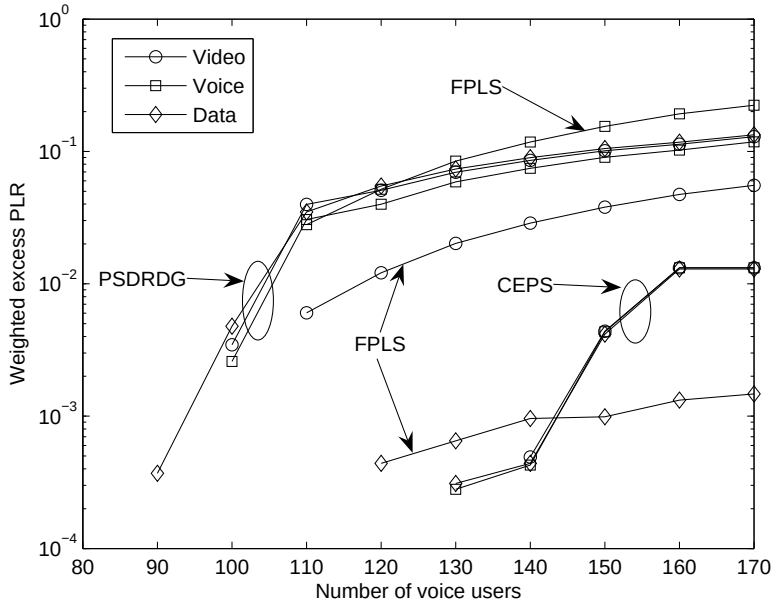

Figure 3.15: Weighted excess PLR of different traffic classes with 30 video users, 50 data users, variable number of voice users and QoS parameters in Table 3.3.

Bibliography

[1] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 146-158, April 1999.

[2] P. Y. Kong, K. C. Chua, and B. Bensaou, "A novel scheduling scheme to share dropping ratio while guaranteeing a delay bound in a multicode-CDMA network," *IEEE/ACM, Trans. Networking*, vol. 11, no. 6, pp. 994-1006, Dec. 2003.

[3] V. Huang and W. Zhuang, "QoS-orientied packet scheduling for wireless multimedia CDMA communications," *IEEE Trans. Mobile Comput.*, vol. 3, pp. 73-85, Jan.-Mar. 2004.

[4] M. A. Arad and A. Leon-Garcia, "Scheduled CDMA: A hybrid multiple access for wireless ATM networks," *Proc. of IEEE Pers., Indoor & mobile Radio Commun. 1996*, pp. 913-917, Oct. 1996.

[5] O. Gurbuz and H. Owen, "Dynamic resource scheduling strategies for QoS in W-CDMA," *Proc. of IEEE GLOBECOM 1999*, pp. 183-187, Dec. 1999.

[6] X. Liu, E. K. P. Chong and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Selected Areas Commun.*, vol. 19, no. 10, pp. 2053-2064, Oct. 2001.

[7] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling for wireless systems with multiple interfaces and multiple constraints," *Proc. of the 6th ACM/SIGCOM MSWiM*, pp. 11-19, Sep. 2003.

[8] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," *Proc. of IEEE GLOBECOM 2003*, pp. 533-537, Dec. 2003.

[9] S. H. Ali and V. C. M. Leung, "Mobility assisted opportunistic scheduling for downlink transmissions in cellular data networks," *Proc. of IEEE WCNC2005*, pp. 1213-1218, Mar. 2005.

[10] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantee in adaptive wireless networks," *IEEE J. Selected Areas Commun.*, vol. 23, no. 5, pp. 1056-1066, May 2005.

[11] D. Zhao, X. Shen and J. W. Mark, "Radio resource management for cellular CDMA systems supporting heterogeneous service," *IEEE Trans. Mobile Computing*, vol. 2, no. 2, pp. 147-160, Apr.-Jun. 2003.

[12] Q. Liu, S. Zhou and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142-1153, May 2005.

[13] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, May 2005.

[14] C. Lin and R. D. Gitlin, "Multi-code CDMA wireless personal communication networks," *Proc. of IEEE Commun.*, pp. 1060-1064, Jun. 1995.

[15] A. Hamid, R. Hoshyar, and R. Tafazolli, "Joint rate and power adaptation for MC-CDMA over tempo-spectral domain," *Proc. IEEE WCNC2008*, pp. 969-973, Mar. 2008.

[16] Z. Han, G. Su, A. Kwasinski, M. Wu, and K. J. R. Liu, "Multiuser distortion management of layered video over resource limited downlink multicode-CDMA," *IEEE Trans. Wireless Communi.*, vol. 5, no. 11, pp. 3056-3067, Nov. 2006.

[17] B. S. Thian, Y. Wang, T. T. Tjhung, and L. W. C. Wong, "A hybrid receiver scheme for multiuser multicode CDMA systems in multipath fading channels," *IEEE Trans on Vehicular Tech.*, vol. 56, no. 5, pp. 3014-3023, Sept. 2007.

[18] 3GPP, "Spreading and modulation (FDD)," *3GPP TS25.213*, v3.4.0, Dec. 2000.

[19] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming multimedia users in high-speed downlink packet access (HSDPA)," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 844-855, August 2006.

[20] C. S. Chang and K. C. Chen, "Medium access protocol design for delay-guaranteed multicode CDMA mutimedia networks," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1159-1167, Nov. 2003.

[21] P. Y. Kong, K. C. Chua and B. Bensaou, "Multicode-DRR: A packet-scheduling algorithm for delay guarantee in a multicode-CDMA network," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2694-2704, Nov. 2005.

[22] 3GPP, "Quality of service (QoS) concept and architecture," *3GPP TS23.107*, v6.4.0, Mar. 2006.

[23] L. Lenzini, M. Luise and R. Reggiannini, "CRDA: A collision resolution and dynamic allocation MAC protocol to integrate data and voice in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 6, pp. 1153-1163, June 2001.

[24] F. Yu, V. Krishnamurthy and V. C. M. Leung, "Cross-layer optimal connection admission control for variable bit rate multimedia traffic in packet wireless CDMA networks," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 542-555, Feb. 2006.

Chapter 4

Cross-Layer Optimization for Multimedia Transport over Multicode CDMA Networks [1]

An important task for contemporary packet-switched networks is to provision the QoS required to transport multimedia traffic streams while efficiently utilizing the system capacity. As wireless spectrum can be expensive, it is particularly important to optimize multimedia transport over wireless networks so that more multimedia traffic can be admitted into the system to enhance the bandwidth utilization in the presence of a heavy traffic load. We have proposed a CEPS algorithm in Chapter 3 to optimize the FLR of a MC-CDMA system. The objective of CEPS is to schedule packets from different traffic flows by considering a FLR utility function instead of handling BER and FDR independently, so that the FLR utility is minimized and different traffic flows can share the excess FLR (defined as the FLR excess to the pre-designed target) fairly. CEPS performs the optimization frame by frame by meeting the optimization

[1] A version of this chapter has been published. H. Chen, H. C.B. Chan and V. C.M. Leung, "Two cross-Layer optimization methods for transporting multimedia traffic over multicode CDMA networks," in *Proc. of IEEE WCNC'07*, March 2007

objective while scheduling every single data frame. It does not take the advantage of the traffic model or statistics of multimedia traffic flows. As a further enhancement of CEPS, in this chapter, we focus on optimizing the resource allocation in a MC-CDMA system by considering the traffic model/statistics and traffic dynamics of the system. We propose a cross-layer method named traffic adaptive scheme to optimize the maximum number of simultaneous data frame transmissions in the system, taking into account the packet arrival rate and the queue status. The problem is formalized as an MDP and solved by linear programming. Also, we provide the analysis of the system performance for the proposed optimization scheme in terms of the FLR, system throughput and packet access delay. To alleviate the computational complexity of the optimized scheme, we also propose an approximate method named rate adaptive scheme. By modeling multimedia traffic using the Markov-modulated Poisson process (MMPP) traffic model [22], the performance of these two methods is evaluated. We also implement a simulation model to validate the theoretical analysis and to compare the proposed schemes with other schemes. The results show that the system performance can be significantly improved by the cross-layer optimization methods in terms of FLR, throughput and the packet access delay, especially when the traffic load of the system is heavy. Also, the results show that the rate adaptive scheme can produce a performance close to the traffic adaptive scheme, when the system traffic load is heavy, although it is not as good as the latter when the system is lightly loaded.

The remainder of this chapter is organized as follows. Section 4.1 introduces the related work. We describe the MC-CDMA system and the MMPP traffic model in Section 4.2. In Section 4.3, we present an overview of the system operations on determining the maximum number of simultaneous transmissions, scheduling, and slot allocation. Then, we present the

queue analysis, traffic adaptive optimization scheme, QoS analysis and rate adaptive scheme in Section 4.4. Section 4.5 gives the results and discussion. Section 4.6 summarizes the chapter.

4.1 Related Work

Previous work on CDMA networks usually aims to separately meet the physical layer QoS requirement in terms of SINR or BER, and MAC layer QoS requirement in terms of FDR [1–4]. Recently, many cross-layer methods have been proposed to improve the system performance of wireless networks. For example, in [11, 12], based on the channel state awareness and an AMC method, the system can choose to serve a user with a better channel condition more often so that the throughput of the system can be improved. In [13], using AMC, a fixed bandwidth is reserved for every QoS-guaranteed flow based on the estimated effective bandwidth and the channel status, and then the traffic is scheduled based on both the buffer status and the channel condition for the specific user. In [14], the queue status and the channel status are used to determine the spreading factor and the optimal number of simultaneous transmissions. In [15], a cross-layer design is proposed for IEEE 802.16/WiMAX networks. Specifically, the physical layer information are used to determine link layer scheduling parameters. In [16], a cross-layer scheduling algorithm is proposed for a radio access network using a multicarrier air interface in a multicell multiuser context. It considers channel, physical layer and application-related information to make the scheduling decisions.

In general, all of the above-mentioned work [11–16] still deals with QoS at the physical and MAC layers separately. However, these two layers are in fact closely related since: 1) a data frame is only correctly received when it is neither dropped at the MAC layer (due to deadline

violation or queue overflow) nor corrupted by uncorrectable errors at the physical layer, and 2) allowing more simultaneous data frame transmissions in the CDMA channel reduces frame drops at a given packet arrival rate but increases BER and hence frame corruptions due to heavier multiple-access-interference (MAI). This means that there is a trade-off between minimizing the MAC layer's FDR and the physical layer's BER. The optimal solution should therefore minimize the total FLR at the MAC layer by jointly optimizing FDR and BER across the MAC and physical layers. Essentially, this aims to minimize the overall PLR experienced by higher layers. Note that this cross-layer optimization method is particularly important for handling time-varying physical channels and variable bit rate multimedia traffic in future wireless communication systems. This research issue has gained much interest in recent years. For example, in [17], the queue status is taken into account in determining the SINR bounds for AMC so that the overall FLR can be optimized. In [18], the upper and lower SINR bounds for each AMC mode is designed based on the maximum retransmission time of the system. However, both designs are not applicable for CDMA systems and it assumes a Poisson traffic model which is not suitable for multimedia traffic. [19] proposes a two-dimensional markov channel mode and an AMC threshold searching algorithm for a system employing AMC and automatic repeat request. And with the queuing analysis of the system, the link layer packet arrival rate and the physical layer tolerable packet error rate are determined to provide specific FLR and delay. However, it requires exhaustive search for the optimization and it does not count the dynamic of the traffic flows.

4.2 System Model

4.2.1 Multicode CDMA System

In this chapter, we consider a multi-cell system employing TDMA and MC-CDMA in each TDMA time slot. We focus on the uplink common traffic channel in one radio cell. Without loss of generality, the uplink common traffic channel is divided into TDMA frames of duration T_f and each TDMA frame consists of a fixed number (F) of time slots, each with a duration of T_s. And also we assume that perfect signaling channels exist for the mobile stations to convey their states (e.g., data rates, queue lengths) at the end of each TDMA frame to the base station, and for the base station to inform the mobile stations the transmission schedule for the next TDMA frame before it starts. Also, the channel status is also exchanged via the signaling channel for power control purpose.

We consider that randomly chosen pseudo-noise codes with the same spreading gain G are used in the system. For every mobile station, packets from the higher layer are fragmented and encapsulated into data frames with the same specific size so that each data frame can be exactly transmitted in one time slot using one spreading code. To achieve different data rate, a mobile station may transmit more than one data frames 1) in more than one time slot with one code, 2) in one time slot with different codes, or 3) in more than one time slot with more than one codes. Note that, the number of data frames transmitted by one mobile station in one time slot is limited by the power constraint on the mobile station. The maximum number of codes to be used in one time slot is determined by the system to provide adequate physical layer QoS BER. Essentially, it is the parameter that we hope to adapt to optimize the system performance in this chapter.

We know that with certain resource allocation schemes, the total traffic load in a TDMA frame in each cell can be estimated and approximately distributed to each time slot. So the inter-cell interference can be estimated. Following a common practice, we assume that noise and inter-cell interference follow a Gaussian distribution with the normalized variance (σ^2). To facilitate the development of the analytical model, we also consider that all data frames use non-orthogonal spreading codes even if they are from the same mobile station, like in many current systems (for example, IS95). Furthermore, we assume that the near-far problem and the fading effects are mitigated by perfect power control and can be neglected (an imperfect power control may lead to inaccurate calculations and less effectiveness of the proposed optimization). So, $ber(u)$, the BER for a data frame if there are totally u data frames transmitted in the time slot can be determined as in [1–3]. And given the channel coding scheme, the expected frame error probability for such a data frame, $fer(u)$, can be calculated based on the SINR too.

4.2.2 MMPP Traffic

The MMPP model has been shown to be effective in representing many types of multimedia traffic, including voice [23], MPEG video [24] and general data [25]. We apply the MMPP model for all the traffic flows in the system (note that the Poisson model is a special case of the MMPP model).

For simplicity, we consider that each mobile station has only one traffic flow. We assume that there are C classes of MMPP traffic flows and K_c traffic flows for each class c in the system. A class c ($c \in \{1, \ldots, C\}$) traffic flow has S_c different traffic states and packets arrive following a Poisson process with a specific rate $R_{c,s}$ in each traffic state s ($s \in \{1, \ldots, S_c\}$). Transitions between traffic states are governed by a continuous-time Markov chain (or more generally a

Markov renewal process). The infinitesimal generating matrix for the Markov chain is defined as

$$\Gamma_c = \begin{bmatrix} -\nu_{c,(1,1)}, \gamma_{c,(1,2)}, \ldots, \gamma_{c,(1,S_c)} \\ \gamma_{c,(2,1)}, -\nu_{c,(2,2)}, \ldots, \gamma_{c,(2,S_c)} \\ \ldots \\ \gamma_{c,(S_c,1)}, \gamma_{c,(S_c,2)}, \ldots, -\nu_{c,(S_c,S_c)} \end{bmatrix} \tag{4.1}$$

where $\gamma_{c,(i,j)}$ $(i, j \in 1 \ldots S_c, i \neq j)$ is the transition rate from state i to state j for class c traffic flows and $\nu_{c,(i,i)} = \sum_{j \neq i} \gamma_{c,(i,j)}$ is the rate that a traffic flow leaves the state i. We can get the stationary distribution of the traffic states of class c traffic flows as a row vector $\chi_c = [\chi_{c,1}, \chi_{c,2}, ..., \chi_{c,S_c}]$, which must satisfy $\chi_c \Gamma_c = 0$ [26].

As in many related studies, we apply an approximate discrete MMPP traffic model in our analysis and optimization, instead of using the above continuous model directly. Note that, the average duration of one traffic state is very long compared to a TDMA frame duration. So we can safely assume that the traffic state changes only occur at the boundaries of TDMA frames. In another word, we assume that the traffic state remains unchanged during a TDMA frame.

4.3 System Operation

At the beginning of a TDMA frame, the system first makes some decisions. Normally, it decides how many simultaneous data frame transmissions at most can be accommodated in one time slot (i.e., an upper bound for the simultaneous data frame transmissions), how many data frames from each traffic flow should be scheduled, and how the time slots are allocated to transmissions of all scheduled data frames. After making the decisions, all the mobile stations get informed by the base station via the reliable signaling channel. Then they get the scheduled

data frames out of the corresponding queue buffers for the subsequent transmissions. During the TDMA frame, all the scheduled data frames are transmitted via time slots based on the slot allocation decision. At the same time, newly arrived packets enter the respective queues after fragmentation and encapsulation. We assume that there is a queue buffer which can hold at most V_c data frames for every class c traffic flow. If the newly arrived data frames find a full buffer, it is simply dropped. We assume that data frames dropped and transmitted with uncorrectable errors can be handled by the upper layers, i.e., the MAC layer does not perform data frame retransmissions.

In general, all the decision factors mentioned above can be optimized to enhance the system performance and QoS. In this chapter, as an example, we focus on deciding the maximum number of data frames (an upper bound) that can be simultaneously transmitted in each time slot of a TDMA frame. The other two decision factors are also utilized in the optimization. In the following, we briefly discuss how the decisions are made and give examples on the scheduling and slot allocation methods.

4.3.1 Maximum Number of Simultaneous Transmissions

Generally, most CDMA systems aim to keep the physical layer BER below a threshold value ber_0. It results a specific upper bound $u = u_0$ of the number of data frames simultaneously transmitted in one time slot. This upper bound is governed by the levels of noise and inter-cell interference. For the system introduced here, it is:

$$u_0 = \arg\max_u \{ber(u) \leq ber_0\} \tag{4.2}$$

This function is commonly used in existing CDMA systems. In this chapter, we propose am optimization for the upper bound u of the number of simultaneous data frame transmissions in one time slot. Note that if the aggregative packet arrival rate of all traffic flows are very high or the queue buffers of most users are almost full, it is better to slightly increase u so that more data frames can be sent out in the current TDMA frame and the FDR can decrease dramatically. Of course the BER as well as the FER for the data frames transmitted in the TDMA frame may be a little higher. However, the proposed optimization may find the optimal u to minimize the total FLR (counting both FDR and FER) over time. The details of the optimization are presented in Section 4.4.

4.3.2 Scheduling

The scheduling algorithm is an important element of the system. It decides how to schedule data frames from each traffic flow in a TDMA frame so that the system can provide fairness to different traffic flows. Here, we describe a scheduling decision as a scheduling decision vector,

$$M = (m_{1,1}, m_{1,2}, \ldots, m_{1,K_1}, m_{2,1}, \ldots, m_{C,K_C}). \tag{4.3}$$

In the vector, $m_{c,k}$ is the number of data frames to be transmitted for class c traffic flow k in the TDMA frame according to the scheduling decision.

One major factor for making the scheduling decision is the buffer states of all traffic flows at the beginning of the TDMA frame. It can be described as a buffer state vector,

$$N = (n_{1,1}, n_{1,2}, \ldots, n_{1,K_1}, n_{2,1}, \ldots, n_{C,K_C}) \tag{4.4}$$

and $n_{c,k}$ is the number of data frames in the buffer for class c traffic flow k. Note that the number of data frames scheduled for any traffic flow cannot be greater than the number of data

frames in its queue, i.e., $m_{c,k} \leq n_{c,k}$. Also, it must be aware that there is a buffer size V_c for any class c traffic flow, i.e, $n_{c,e} \leq V_c$.

Another important factor is the traffic states of all traffic flows in the TDMA frame. It can be described as a traffic state vector,

$$D = (d_{1,1}, d_{1,2}, \ldots, d_{1,K_1}, d_{2,1}, \ldots, d_{C,K_C}) \tag{4.5}$$

and $d_{c,k}$ is the traffic state of class c traffic flow k in the TDMA frame. With the traffic state, the system knows the packet arrival rate of all the traffic flows.

Also, the decision is limited by the maximum number (the upper bound u) of simultaneous data frame transmissions. Specifically, the total number of data frame transmissions in the TDMA frame cannot be greater than uF and we have $\sum_{c=1}^{C} \sum_{k=1}^{K_c} m_{c,k} \leq uF$.

Some other factors such as historical bandwidth allocation and historical FLR may also be considered for making a scheduling decision.

In this chapter, the scheduling function is defined as a mapping which determines the scheduling decision vector M by the maximum number (upper bound u)of simultaneous data frames transmissions, the buffer state vector N and the traffic state vector V. Such a function is defined as:

$$\Omega : \{u, N, D\} \rightarrow M \quad (u \geq 1, N \in \Upsilon, D \in \Xi, M \in \Psi) \tag{4.6}$$

Here Υ, Ξ and Ψ are the spaces of the buffer state vector N, the traffic state vector D and the scheduling decision vector M respectively and defined as: $\Upsilon = \{N | n_{c,k} \in \{0 \ldots K_c\}\}$, $\Xi = \{D | d_{c,k} \in \{1 \ldots S_c\}\}$, $\Psi = \{M | m_{c,k} \in \{0 \ldots K_c\}\}$. Note that the above definition of scheduling function is valid even if some static parameters such as priorities and QoS requirements for

different traffic flows are taken into account. Essentially, if any static parameters are applied in the scheduling, the scheduling definition can be presented as:

$$\Omega : \{u, N, D, Params\} \rightarrow M \quad (u \geq 1, N \in \Upsilon, D \in \Xi, M \in \Psi) \tag{4.7}$$

Here $Params$ represents all static parameters applied. Since all these parameters are static and do not vary among different TDMA frames, the scheduling decision will be determined for any TDMA frame given the buffer states and the traffic states of all traffic flows in the TDMA frame. As a result, $Params$ can be neglected from (4.7) and the definition of (4.6) is still valid. For example, if a specific traffic flow i of class j has a higher priority than all other traffic flows, it needs to occupy any available slot as long as it still has data frames to transmit. In this situation, the scheduling decision is partially determined by whether $n_{j,i} > 0$. In another words, it is still based on the traffic states and the buffer states of all traffic flows.

4.3.3 Slot Allocation

Given u the maximum number of the simultaneous data frame transmissions allowed in one time slot and M the scheduling decision vector, the system also need to inform the mobile stations in which time slot their scheduled data frames should be transmitted. If every time slot holds exactly u transmissions of data frames, the BER of all data frames can be easily determined. However, if there are not so many data frames transmitted in one TDMA frame, the BER of a data frame is affected by the number of data frames transmitted in the same time slot. In other words, it depends on the slot allocation scheme. Here we consider a simple slot allocation scheme, which assigns data frames to every time slot evenly. In this case, any time slot may have $\lfloor \frac{z}{F} \rfloor$ or $\lfloor \frac{z}{F} \rfloor + 1$ data frames with probability $1 - \frac{z \bmod F}{F}$ and $\frac{z \bmod F}{F}$, respectively, if there

are totally z ($0 \leq z \leq uF$) data frames scheduled in the TDMA frame, i.e., $z = \sum_{c,k} m_{c,k}$. So $E_{fer}(z, u)$, the expected FER of data frames in the TDMA frame can be easily calculated. Note that, other slot allocation methods can also be applied in practice and a similar function E_{fer} can also be found. In this chapter, we assume that a slot allocation method is provided and $E_{fer}(z, u)$, can be determined. This function is important for evaluating and optimizing the system performance.

4.4 Traffic Adaptive Optimization

In this chapter, we focus on the cross-layer optimization of u, the upper bound of the number of simultaneous data frame transmissions in each time slot of a TDMA frame. The objective is to minimize the FLR (i.e., taking into consideration both FER and FDR) of the system so that the channel throughput is maximized (Note that the throughput is the aggregate packet arrival rate factored by (1-FLR)). To make such an optimization, the system needs to know the scheduling function and slot allocation when every specific u is given for a TDMA frame. As discussed in Section 4.3.3, any slot allocation scheme can be used as long as the function $E_{fer}(z, u)$, can be acquired. The optimization of u concerns not the actual slot allocation but the function $E_{fer}(z, u)$. On the other hand, any scheduling function defined as (4.6) can be used for the optimization. We will also show a specific example in Section 4.4.1 which minimizes the FLR of every TDMA frame.

4.4.1 Queue Analysis

Before presenting the optimization of u, we first need to analyze the number of data frame losses of a TDMA frame if the upper bound u of the number of simultaneous data frame transmission is decided, the buffer state vector N and the traffic state vector D is known, and the slot allocation and the scheduling decision vector M is given.

To facilitate the analysis, we define the buffer state vector at the beginning of the TDMA frame t as $N(t)$ and the corresponding number of data frames in the buffer of class c traffic flow k as $n_{c,k}(t)$. Similarly, we define $D(t)$ and $M(t)$ as the traffic state vector and the scheduling decision vector for TDMA frame t, and $d_{c,k}(t)$ and $m_{c,k}(t)$ as the corresponding traffic state and the number of scheduled data frames of class c traffic flow k. The maximum number of simultaneous data frame transmissions (the upper bound) for the TDMA frame is defined as $u(t)$. Also, we assume that a specific slot allocation method is always used and $E_{fer}(z, u)$ is the expected frame error probability for each data frame transmitted in a TDMA frame with totally z data frames in the TDMA frame and with u applied as the maximum simultaneous data frame transmissions in one slot.

We denote the number of arriving data frames for traffic flow k of class c in TDMA frame t as $A_{c,k}(t)$. Here to simplify the presentation, we assume that one packet is always converted to exactly one date frame after the fragmentation and encapsulation. Note that the following model can also be extended to handle other situations (i.e., the assumption is not mandatory). Since the traffic state of class c traffic flow k is $d_{c,k}(t)$, the packets should arrive according to a Poisson process with rate $R_{c,d_{c,k}(t)}(t)$ according to the MMPP model introduced in Section 4.2.2. It is not difficult to compute the probability that there are i data frames arriving in

TDMA frame t for class c traffic flow k as:

$$P\{A_{c,k}(t) = i\} = \frac{(R_{c,d_{c,k}(t)})^i \times e^{-R_{c,d_{c,k}(t)}}}{i!} \tag{4.8}$$

So we can get the conditional probability of the buffer state vector at the beginning of the TDMA frame $(t+1)$ as

$$P\{N(t+1)|N(t), D(t), M(t)\} = \prod_{c=1}^{C}\prod_{k=1}^{K_c} P\{n_{c,k}(t+1)|n_{c,k}(t), d_{c,k}(t), m_{c,k}(t)\} \tag{4.9}$$

where

$$\begin{aligned} &P\{n_{c,k}(t+1)|n_{c,k}(t), d_{c,k}(t), m_{c,k}(t)\} \\ &= \begin{cases} P\{A_{c,k}(t) = n_{c,k}(t+1) - (n_{c,k}(t) - m_{c,k}(t))^+\} & \text{if } n_{c,k}(t+1) < V_c \\ \sum_{i=V_c-(n_{c,k}(t)-m_{c,k}(t))^+}^{\infty} P\{A_{c,k}(t) = i\} & \text{if } n_{c,k}(t+1) = V_c \end{cases} \end{aligned} \tag{4.10}$$

Here the value of $(x)^+$ is x or 0 if x is less than 0. Normally, $m_{c,k}(t)$ should be less than $n_{c,k}(t)$ for a good scheduling decision, which means that $(n_{c,k}(t) - m_{c,k}(t))^+$ should always equal to $n_{c,k}(t) - m_{c,k}(t)$.

Note that data frames may be dropped if the queue buffer is full when they arrive. If $n_{c,k}(t+1) < V_c$, no data frames will be dropped and the number of arrived data frames in the TDMA frame t is $n_{c,k}(t+1) - (n_{c,k}(t) - m_{c,k}(t))^+$. So the probability of this case is exactly the probability that there is such data frames arriving. And if $n_{c,k}(t+1) = V_c$, it indicates that arrived data frames may be dropped and the number of arrived data frames in the TDMA frame can be any number greater than $V_c - (n_{c,k}(t) - m_{c,k}(t))^+$. The probability of this case is that there are over $V_c - (n_{c,k}(t) - m_{c,k}(t))^+$ data frames arrived in the TDMA frame.

With the above analysis, we can get the expected number of data frames dropped for class

c traffic flow k for TDMA frame t as follows:

$$DR_{c,k}(N(t), D(t), M(t)) = \sum_{i=0}^{\infty} (i + (n_{c,k}(t) - m_{c,k}(t))^+ - V_c)^+ P\{A_{c,k}(t) = i\} \tag{4.11}$$

Here $P(A_{c,k}(t) = i)$ is the probability that there are i data frames arrived to the buffer for class c traffic flow k in TDMA frame t. And $(i + (n_{c,k}(t) - m_{c,k}(t))^+ - V_c)^+$ data frames are dropped because of buffer overflow.

On the other hand, if we know how these data frames are allocated into the time slots, we can also calculate the expected data frame errors for each traffic flow. The expected data frame errors for traffic flow k of class c in TDMA frame t is:

$$ER_{c,k}(M(t), u(t)) = m_{c,k}(t) E_{fer}(\sum_{i,j} m_{i,j}(t), u(t)) \tag{4.12}$$

So the expected number of data frame losses for class c traffic flow k in TDMA frame t is

$$LO_{c,k}(N(t), D(t), M(t), u(t)) = ER_{c,k}(M(t), u(t)) + DR_{c,k}(N(t), D(t), M(t)). \tag{4.13}$$

And the expected number of total data frame losses is

$$\sum_c \sum_k LO_{c,k}(N(t), D(t), M(t), u(t)).$$

Note that this number also depends on the function $E_{fer}(z, u)$, which is a static parameter (i.e., it does not change over time).

As we discussed before, to perform the optimization of u, we need to determine the scheduling function based on u and the slot allocation method. Here we consider the following scheduling function which minimizes the total data frame losses in a TDMA frame:

$$M(t) = arg \min_{M(t)^*} \sum_c \sum_k LO_{c,k}(N(t), D(t), M(t)^*, u(t)). \tag{4.14}$$

Note that, this scheduling function is of the same form as (4.6). Other scheduling functions can also be applied as long as they are of the form (4.6). For example, we can also consider the priority control or QoS requirement for some traffic flows when determining the scheduling function.

4.4.2 The Optimization

Based on the above analysis, in this section, we present a cross-layer-based optimization mechanism to determine the maximum number (or the upper bound of the number) of data frames transmitted in one time slot of a TDMA frame to minimize the total FLR of all traffic flows. This optimization method is based on such a principle. The more data frame transmissions are allowed in every time slot in a TDMA frame, the higher chance the data frames are transmitted with errors and the less chances the data frames expire. Especially, a larger upper bound of the number of simultaneous transmissions in one time slot for a TDMA frame may decrease the large FDR experienced when the traffic flows are in a "burst" (i.e., there happen to be a lot of data frames arrived in a short period). Since the FLR accounts both frame errors and the frame drops, the minimal FLR can be acquired by choosing an appropriate upper bound of the number of data frame transmissions allowed in every time slot.

As stated in previous sections, to perform the optimization, there must be a scheduling function as (4.6) and $E_{fer}(z, u)$ must be given. Then, the decision of u in every TDMA frame is a function of the buffer state vector N and the traffic state vector D. Here, as presented in Section 4.3, we assume that the vectors $N(t)$ and $D(t)$ are known by the system at the beginning of TDMA frame t via the signaling channels. We perform the cross-layer optimization using a MDP [27] as explained below.

State space

We define the system state of the MDP at the beginning of the TDMA frame t, $G(t)$, as $G(t) = [N(t), D(t)]$. So the state space here of the system state is:

$$\Phi = \{G(t) = [N(t), D(t)] : N(t) \in \Upsilon, D(t) \in \Xi\} \tag{4.15}$$

Actions

At the beginning of TDMA frame t, the maximum number of data frame transmissions allowed in one time slot, $u(t)$, is determined. Then, the specific scheduling function and slot allocation method are applied to determine the packet transmissions in the TDMA frame. This number $u(t)$ is the action to be determined by the MDP. Its state space is $\Delta = \{u(t) : u(t) \geq 1\}$.

State transition

It is easy to get the state transition probability $P(G(t+1) \mid G(t), u(t))$ since the state transition of $D(t)$ depends only on the MMPP traffic model for each traffic flow and the transition of $N(t)$ depends only on $D(t)$ and the scheduling decision $M(t)$ which is determined by $N(t)$, $D(t)$ and $u(t)$ given a scheduling function defined as in (4.6). We have:

$$P\{G(t+1) \mid G(t), u(t)\} = P\{D(t+1) \mid D(t)\} \times P\{N(t+1) \mid N(t), D(t), u(t)\} \tag{4.16}$$

Here according to (4.1) we have $P\{D(t+1) \mid D(t)\} = \prod_{c=1}^{C} \prod_{k=1}^{K_c} \gamma_{c,(d_{c,k}(t),d_{c,k}(t+1))}$. Since the scheduling function is provided, the scheduling decision $M(t)$ is determined given $N(t)$, $D(t)$ and $u(t)$. We can have

$$\begin{aligned} P\{N(t+1) \mid N(t), D(t), u(t)\} =& P\{N(t+1) \mid N(t), D(t), M(t), u(t)\} \\ =& P\{N(t+1) \mid N(t), D(t), M(t)\} \end{aligned} \tag{4.17}$$

Note that $P\{N(t+1) \mid N(t), D(t), M(t)\}$ is already calculated according to (4.9) and (4.10). It is not difficult to see that the constructed markov decision process is a unichain model.

Cost

Depending on the optimization objective, the cost can be defined in different ways. In this chapter, it is defined as the total data frame losses over time. Given the system state $G(t)$ and the action $u(t)$, it is not difficult to get the cost for a TDMA frame as:

$$CT(t) = \sum_{c} \sum_{k} LO_{c,k}(N(t), D(t), M(t), u(t)) \tag{4.18}$$

Then the average cost over time is then:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \sum_{c} \sum_{k} LO_{c,k}(N(t), D(t), M(t), u(t)). \tag{4.19}$$

The linear programming solution

The above sections actually defined different components of an MDP. To acquire the minimum average cost over time, i.e., to minimize the number of data frame losses over time, we need to choose an appropriate decision $u(t)$ for each system state $G(t)$. The decision can be stochastic. In other words, for a specific system state $G(t)$, the optimal decision may be to choose $u(t) = 1$ with probability p_1, $u(t) = 2$ with probability p_2 and so on. Then, to solve the MDP, we get the following mapping:

$$G(t) \to \{p_1, p_2, \dots\} \tag{4.20}$$

Note that, a deterministic decision (to choose an explicit $u(t)$ for a specific $G(t)$) is a special case of the stochastic decision.

The MDP described above can be transformed to a linear programming problem as follows:

$$\begin{aligned} & \min_{x(.)} \sum_{G\in\Phi}\sum_{u>0} x(G,u)c(G,u) \\ s.t. \quad & \sum_{u>0} x(G,u) - \sum_{G^*\in\Phi}\sum_{u>0} P\{G \mid G^*,u\}x(G^*,u) = 0 \\ & \sum_{G\in\Phi}\sum_{u>0} x(G,u) = 1 \end{aligned} \tag{4.21}$$

Here $c(G,u)$ is the cost (expected number of data frame losses) for one TDMA frame when the system state for the TDMA frame is G and the decision is u, i.e., $c(G,u) = \sum_{c,k} LO_{c,k}(N,D,M,u)$ where $G = (N,D)$ and $M = \Omega(N,D,u)$.

Also, $P\{G \mid G^*,u\}$ is the state transition function defined in Section 4.4.2, and $x(G,u)$ is the probability of the case that the system state is G and the maximum number of simultaneous transmissions is decided as u. Then we can get the decision on $u(t)$ for TDMA frame t as:

$$p\{u(t) = i \mid G(t)\} = \frac{x(G(t),i)}{\sum_{j=0}^{\infty} x(G(t),j)} \tag{4.22}$$

4.4.3 QoS Analysis

The above optimization minimizes the number of data frame losses over time by choosing an appropriate value of u for a TDMA frame. By doing so, the FLR of the system can be minimized since the expected number of data frames arrived to the system over time is a constant. Note that it is determined by the traffic models of the traffic flows. Based on the above, we can perform further analysis of the system QoS as follows.

From the above optimization, we can get the stationary distribution of the system state G as:

$$B(g) = P\{G = g\} = \sum_{j=0}^{\infty} x(g,j) \tag{4.23}$$

Assume that the buffer state vector and the traffic state vector corresponding to g is n and d, and the corresponding scheduling decision vector given n, d, u is m according to the scheduling function. Then we can determine the expected number of data frame losses of class c traffic k in one TDMA frame to be $\sum_g B(g)LO_{c,k}(n,d,m,u)$. So, it is not difficult to get the corresponding FLR $\delta_{c,k}$ and the average TDMA frame throughput $\theta_{c,k}$ for class c traffic flow k as follows:

$$\delta_{c,k} = \frac{\sum_g B(g)LO_{c,k}(n,d,m,u)}{R_c T_f} \tag{4.24}$$

$$\theta_{c,k} = R_c T_f - \sum_g B(g)LO_{c,k}(n,d,m,u) \tag{4.25}$$

Here R_c is the average data frame arrival rate of class c traffic flows. And $R_c T_f$ is the expected number of newly arrived data frames to the buffer queue in one TDMA frame.

If a data frame arrives to the buffer in a TDMA frame, there is an expected delay of half of a TDMA frame from the arrival to the beginning of the next TDMA frame. If it is sent out in the next TDMA frame, there is expectedly another half TDMA frame delay if the aforementioned even allocation method is used. Actually, every data frames transmitted has these two parts of delays totally one TDMA frame as long as it is not dropped for buffer overflow (no matter which TDMA it is sent out). Additionally, if such a data frame is not transmitted in a TDMA frame, it will experience one extra TDMA frame delay. So the expected delay for a data frame of class c traffic flow k (if the data frame is not dropped for buffer overflow) can be calculated as follows.

$$\begin{aligned}
\eta_{c,k} &= T_f + \lim_{T\to\infty} \frac{\sum_{t=0}^{T} \left(n_{c,k}(t) - m_{c,k}(t)\right) \times T_f}{\xi_{c,k} R_c T T_f} \\
&= \frac{\xi_{c,k} R_c T_f}{\xi_{c,k} R_c} + \frac{\sum_g B(g) n_{c,k}}{\xi_{c,k} R_c} - \frac{\sum_g B(g) m_{c,k}}{\xi_{c,k} R_c} \\
&= \frac{\sum_g B(g) n_{c,k}}{\xi_{c,k} R_c}
\end{aligned} \tag{4.26}$$

Here $n_{c,k}$ and $m_{c,k}$ are the components of n and m respectively. $\xi_{c,k}$ equals to $1 - \delta_{c,k}$. It represents the percentage of the data frames of class c traffic flow k which are not dropped for buffer overflow. In the first line, T_f is the first two parts delay discussed above for a data frame and the limitation shows the expected extra delay for one data frame of class c traffic flow k. Note that, during a period $t = 0 \rightarrow T$, the number of class c traffic k data frames arrived in the buffer is $\xi_{c,k} R_c T T_f$. For a TDMA frame t, $n_{c,k}(t) - m_{c,k}(t)$ buffered data frames are not transmitted in the TDMA frame and experience one extra TDMA frame delay. So the first line gives the average delay for data frames not dropped of class c traffic flow k. In the last step, $\xi_{c,k} R_c T_f$ is the average data frames arrived in one TDMA frame and $\sum_g B(g) m_{c,k}$ is the expected data frames transmitted in one TDMA frame. They must be the same.

4.4.4 Limitation and Approximation

The above optimization provides a set of decisions on u for different system states G. The decision can be pre-calculated and stored in the system so that the system can apply it at the beginning of every TDMA frame. However, the linear programming solution of the optimization may become computationally infeasible if the state space of the system is very large (e.g., there are too many traffic flows in the system and the buffer size of each traffic flow is very large). Some simplification and approximation techniques can be utilized to alleviate the problem. For example, for computation purposes, we can assume that all the traffic flows share a single buffer. It will produce better performance since less data frames will be dropped due to shared buffer resources. Also, we provide another scheme which apply the same principle as the optimization above, as u should increase a little bit during "burst" period. We call it the rate adaptive optimization here.

As discussed in Section 4.2.2, the duration between updates of the traffic state for a MMPP traffic flow is normally much longer than a TDMA frame. So we assume that the traffic states of all traffic flows should remain unchanged for a certain period of time. Within that period, the system tries to find the best u. based on the traffic state vector D. Note that if the traffic state vector and the value u are known, the scheduling decision vector $M(t)$ for TDMA frame t depends on the buffer state vector $N(t)$ only. So the transition of the buffer state $N(t)$ which should depend on $M(t)$, $D(t)$ and u becomes deterministic. We have $P\{N(t+1) \mid N(t)\} = P\{N(t+1) \mid N(t), D, M\}$ which can be calculated according to (4.9) and (4.10). The buffer state vectors $N(t)$ at the beginning of TDMA frame t can form a Markov chain. By solving the Markov chain, we can easily obtain the stationary distribution of $N(t)$ as $\pi(n) = P\{N(t) = n\}$. The best u for the period with the traffic state vector D can then be found as follows:

$$u^* = \arg\min_u \sum_n \quad \pi(n) \sum_c \sum_k LO_{c,k}(n, D, M, u) \Bigg). \tag{4.27}$$

When the traffic state vector changes, the optimal u^* should be updated accordingly. The aforementioned method can be used to reduce computation cost significantly. Compared to the traffic adaptive scheme, the rate adaptive scheme does not need to formulate the MDP. Although the computation complexity cannot be compared directly, the dimension of the variables is reduced dramatically (from the product of dimensions of the decision state space, traffic state space and buffer state space to a single decision variable).

4.5 Results and Discussions

We implemented a simulation model to validate the analysis and to study the performance enhancements of the proposed optimization methods. We also compared our work with the transmission scheme proposed in [1–4, 28], where packets are scheduled according to their BER requirements. Furthermore, we compared the two adaptive scheme with the transmission scheme used in CEPS [20, 21]. We have studied a multicode CDMA system (as described in Section 4.2) with a spreading gain of 7 and a noise level of $\sigma^2 = 10db$. Each TDMA frame is 0.02 second long and is divided into two time slots. Each data frame is 511 bits long. It carries a 160 bits payload and a 69 bits header. The channel code is a (511,229,38) BCH code, which can tolerate 38 bit errors. We assume that a packet always has the same length and can be exactly transmitted by one data frame, so that no fragmentation is needed. We also assume that a buffer queue of size 30 is shared by all traffic flows. Note that we use a very small system here for studying the performance of the proposed scheme. A bigger system may lead to larger computation complexity but the proposed scheme should have a similar performance. In the simulations, the physical layer is applied statistically and the data link layer is operated step by step. The results are generated from the average of a couple of same simulations, which runs at least 100,000 TDMA frames.

We first study the basic idea of dynamically adjusting the maximum number of simultaneous data frame transmissions in a time slot. In this study, we simply assume that packets arrive to a mobile station according to a Poisson process. The BER threshold is equivalent to 1% FER, which leads to an upper bound $u = 6$ of the number of simultaneous data frame transmissions in one time slot for the traditional threshold-based scheme. Figure 4.1 shows the simulation

and analytical results of FLR when the packet arrival rate varies. Furthermore, we try different u for the same system and compare their performances. It is clear that the curves for different u have different shapes because the systems with different u have different capacities in one TDMA frame. When the packet arrival rate increases, all these curves are supposed to achieve the limit 1. As we can observe, $u = 6$ used by the traditional scheme (which conforms to (4.2)) achieves the best performance only when the packet arrival rate is between 420 packets/second and 560 packets/second. When the packet arrival rate is lower than 420 packet/second, $u = 5$ achieves a better performance. When the packet arrival rate is greater than 560 packets/second, $u = 7$ gives a much better performance than others. Note that, when the packet arrival rate is greater than 560 packets/second, the system is a little bit overloaded because the FLR of all cases of u are higher than the desired 1% FER. These results support the motivation for the optimization methods proposed in this chapter. In other words, the maximum number of simultaneous data frame transmissions should be changed adaptively based on the traffic flow status.

Figure 4.2 shows the corresponding normalized throughput. Here the normalized throughput is the percentage of successfully transmitted packets from all packets arrived. It is obvious that $u = 6$ gives the best result when the packets arrival rate is lower than 560 packets/second but its performance degrade greatly when the packet arrival rate is higher. And $u = 7$ gives the best performance among all four cases when the packet arrival rate is higher than 560 packets/second. And also, when the packet arrival rate is lower than 420 packets/second, the FLR for every different u is far below 1 so that the normalized throughput is very close to 1.

Figure 4.3 shows the corresponding packet access delay. It is interesting to see that while $u = 6$ shows the best performance in terms of FLR when the packet arrival rate is between 420

packets/second and 560 packets/second arrival rate, the corresponding delay is not the best. Note that, actually this makes sense since the higher service rate, the less delay is experienced. It also reminds us that the performance in terms of FDR should also follow the commonsense that the higher service rate, the less drops happen. The reason of the lower FLR of $u = 6$ compared with $u = 7$ and $u = 8$ during this range of packet arrival rate is that the FLR is dominated by the FER and the FDR is much smaller than FER here. So the effect on FDR is overridden by the dominating FER performance.

All the above three figures show that the simulation results match closely with the analytical results based on the analytical model in Section 4.4.3. It verifies the correctness of the analytical model as well as validates the simulation programs.

Next, we study the proposed optimization methods (i.e., the traffic adaptive scheme and the approximated version, the rate adaptive scheme which updates u according to the system traffic state vector described in Section 4.4.4) for a system with different configurations. We compare them with the scheme proposed in [1–4, 28] (specified as "BERG" in the figures. BERG stands for BER guaranteed), which stops scheduling packets in a slot if the BER will exceed the pre-designed target. We assume that the pre-designed target BER here is equivalent to a 1% FER, and the transmission scheme used in CEPS [20, 21] (specified as "CEPS" in the figures). Note that in CEPS, the data frames are all associated with an expiration time and the data frame drops happens and only happens when data frames expire. So we modify the analytical model of our scheme a little bit (the buffer transition and the FLR calculation) to realize the same buffering mechanism.

We first investigate the performance of the three schemes in the system described above. The aggregative packet arrival rate varies also from 250 packets/second to 750 packets/second.

We assume the expiration time of arrived packets is two TDMA frames, which means a data frame will be dropped if it is not transmitted in the next two TDMA frames upon its arrival.

Figure 4.4 shows the performance in terms of the FLR. It can be seen clearly that both the rate adaptive and traffic adaptive schemes can provide a better performance than the BERG scheme. When the traffic load is light, CEPS scheme provides a better performance than the rate adaptive scheme because it concerns the buffer status while the rate adaptive scheme concerns the traffic flow rate. Note that, the traffic status does not change during each simulation in this figure. So the buffer status is more important than the traffic status for the optimization. And the traffic adaptive can achieve the best performance since it counts both the traffic status and the buffer status. When the traffic load is heavy, all these schemes provides similar performance. It is because when the packet arrival rate is very high, the decisions made by the traffic adaptive scheme and CEPS are almost fixed at $u = 7$ for each frame, which is the same value used in the rate adaptive scheme. Another interesting observation is that the FLR of the rate adaptive scheme does not vary smoothly with the packet arrival rate. Actually, since the packet arrival rate is fixed in each independent simulation, the rate adaptive scheme only applies the best fixed u for each simulation. For the packet arrival rate from 250 packets/second to 425 packets/second, the best u for the rate adaptive scheme is 5. For the packet arrival rate from 450 packets/second to 575 packets/second, the best u for the rate adaptive scheme is 6. When the packet arrival rate is higher than 575, the best u for the rate adaptive scheme is 7. So, as a result, the curve for the rate adaptive scheme has two sharp turns at the packet arrival rate of 425 packets/second and 575 packets/second. The results also show that the two schemes can outperform the traditional scheme. Figure 4.5 shows the corresponding system throughput. It is easy to see that the two adaptive schemes and the CEPS scheme have similar performance,

which is much better than that of the BERG scheme when the packet arrival rate is higher than 600 packets/second. When the packet arrival rate is lower than 600 packets/second, all schemes have the similar performance because all of them have very little FLR. Although the differences in FLR are quite apparent, the differences in throughput are very little (all of them can achieve a throughput very close to the packet arrival rate).

Next, we study the performance of the above schemes with variable number of multimedia traffic flows. We consider a typical VOIP traffic source, which can be represented by a two-state MMPP. At the "talk" state, its packet arrival rate is 25 packets/second. At the "silence" state, its packet arrival rate is 0. The average durations of the "talk" and "silence" states are 1 and 1.35 seconds respectively.

Figure 4.6 shows the FLR of the four schemes when the number of traffic flows in the system varies from 32 to 64. It can be seen that the rate adaptive and traffic adaptive schemes perform much better than the traditional scheme. Specifically, if the target FLR is 1%, the traditional scheme can support about 45 traffic flows while the rate adaptive and traffic adaptive scheme can support about 50 and 55 traffic flows respectively (CEPS can support 49 traffic flows). Note that, when the number of traffic flows increases in the system, the performance of the two adaptive schemes becomes much closer. Also, CEPS scheme could provide a better FLR when the traffic load is light, because the packets arrival rate is relatively low so that it is more important to adapt to the buffer status than to the traffic status. However, when the traffic load is high, the traffic adaptive and rate adaptive schemes perform better for it is more important to concerns the packets coming the next TDMA frame. Actually, for target FLR 1%, the traffic adaptive scheme can obtain a 10% capacity improvement compared to CEPS. Figures 4.7 shows the corresponding simulation results on throughput.

4.6 Summary

In this chapter, we have proposed a traffic adaptive optimization scheme for multi-code CDMA operating over a TDMA framework. Using a Markov decision process model, it seeks to determine the maximum number of simultaneous data frame transmissions that can be supported in a time slot of a TDMA frame. To facilitate implementation, we also propose an approximation scheme called the rate adaptive scheme. Both schemes aim to jointly optimize the physical layer's BER and the MAC layer's FDR to minimize the overall FLR. Simulation results show that these two schemes can improve the FLR and hence the system capacity substantially as compared to the traditional scheme and a previous work CEPS. While the traffic adaptive optimization scheme can give a better performance in general, the computation cost is higher. In most cases, the rate adaptive scheme can give a performance close to that of the traffic adaptive optimization scheme, especially when the system traffic load is heavy.

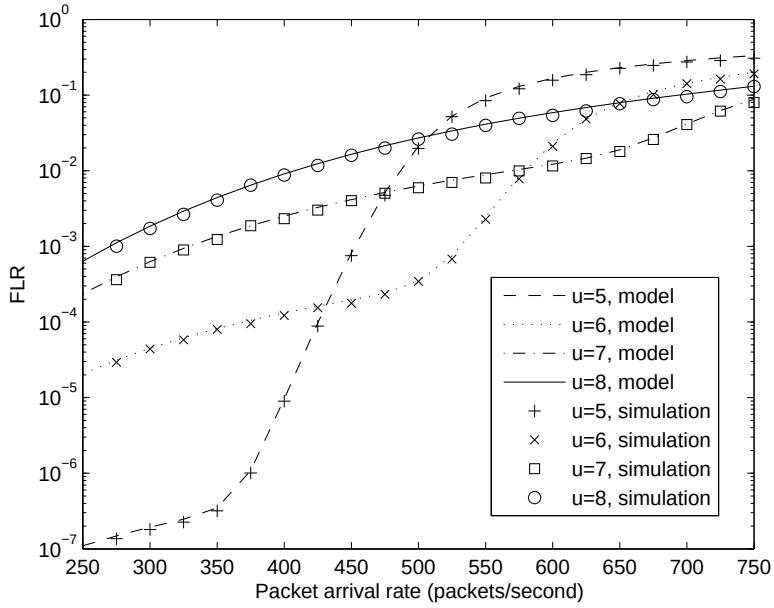

Figure 4.1: Data frame loss ratios vs. data frame generation rates for different fixed u_0.

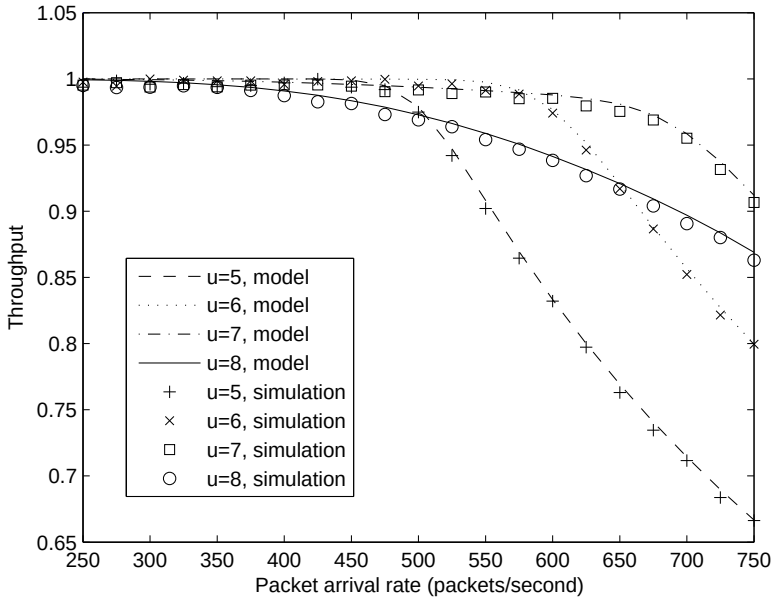

Figure 4.2: Throughput vs. data frame generation rates for different fixed u_0.

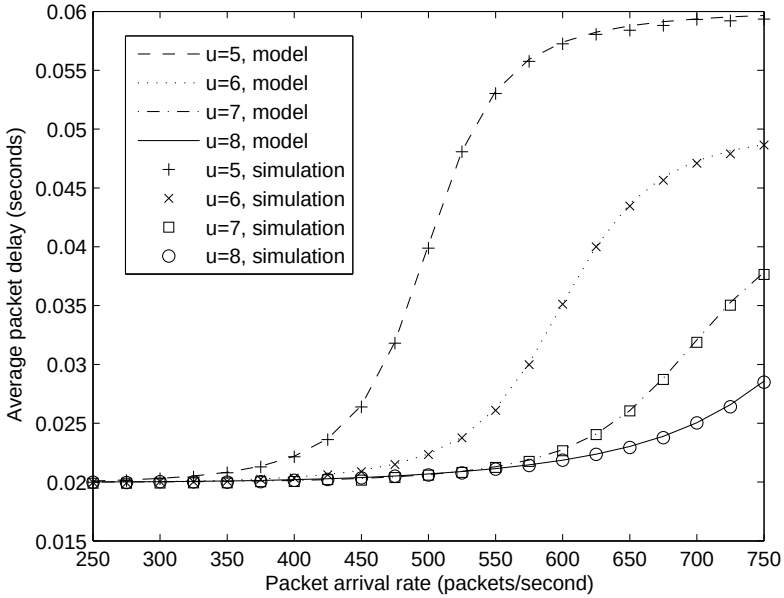

Figure 4.3: Packet access delay vs. data frame generation rates for different fixed u_0.

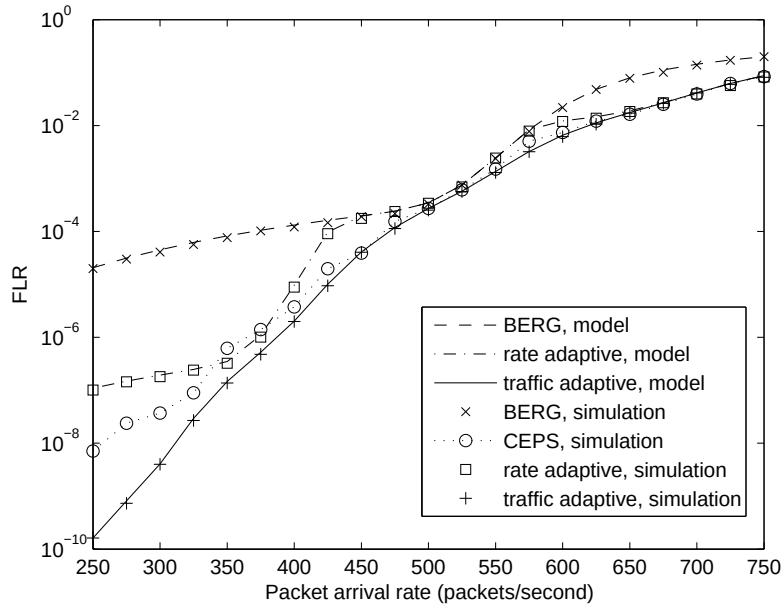

Figure 4.4: Data frame loss ratios vs. data frame generation rates for different schemes on u_0.

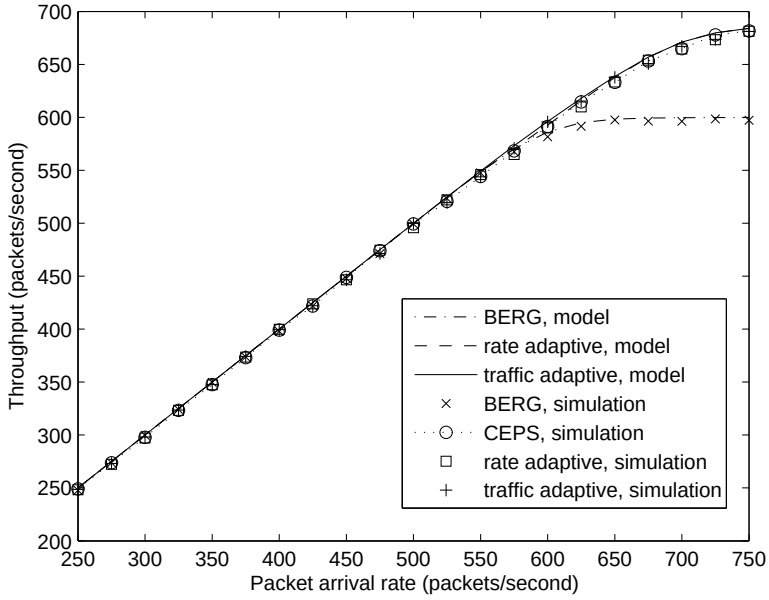

Figure 4.5: Throughput vs. data frame generation rates for different schemes on u_0.

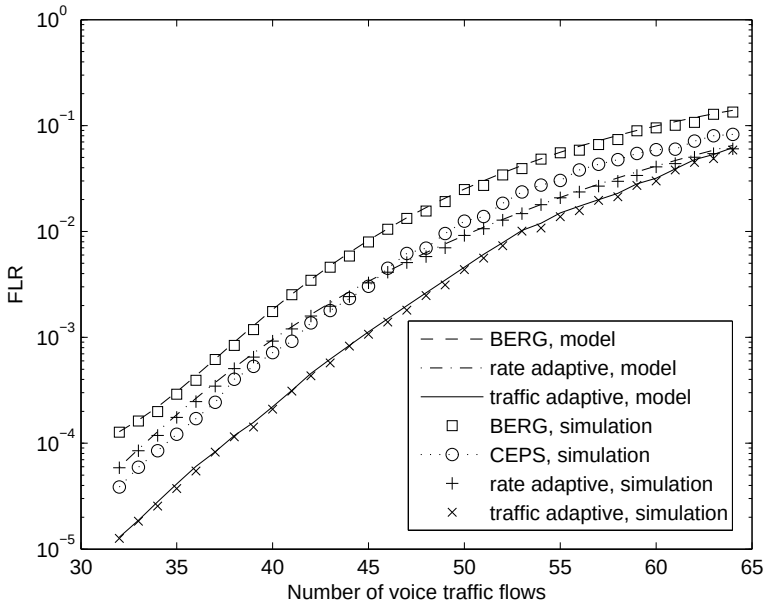

Figure 4.6: Data frame loss ratios vs. number of traffic flows for different schemes on u_0.

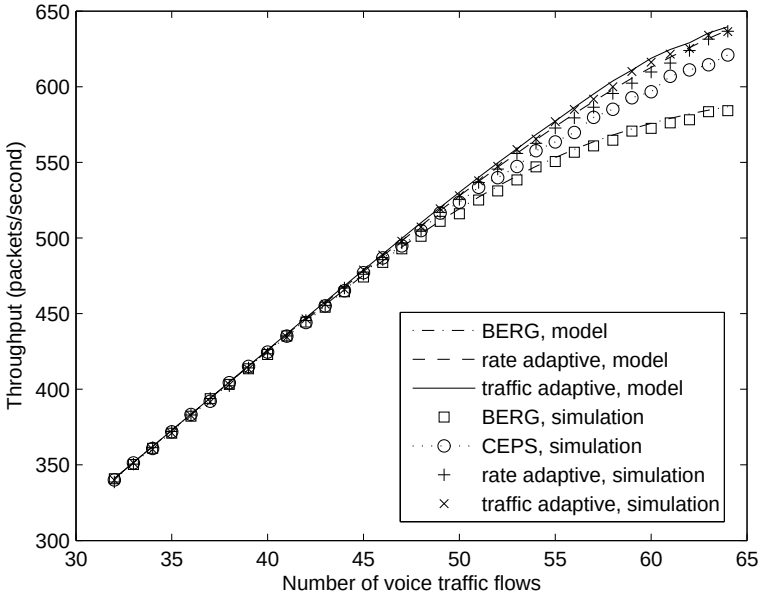

Figure 4.7: Throughput vs. number of traffic flows for different schemes on u_0.

Bibliography

[1] P. Y. Kong, K. C. Chua, and B. Bensaou, "A novel scheduling scheme to share dropping ratio while guaranteeing a delay bound in a multicode-CDMA network," *IEEE/ACM, Trans. Networking*, vol. 11, no. 6, pp. 994-1006, Dec. 2003.

[2] C. S. Chang and K. C. Chen, "Medium access protocol design for delay-guaranteed multicode CDMA mutimedia networks," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1159-1167, Nov. 2003.

[3] P. Y. Kong, K. C. Chua and B. Bensaou, "Multicode-DRR: A packet-scheduling algorithm for delay guarantee in a multicode-CDMA network," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2694-2704, Nov. 2005.

[4] V. Huang and W. Zhuang, "QoS-orientied packet scheduling for wireless multimedia CDMA communications," *IEEE Trans. Mobile Comput.*, vol. 3, pp. 73-85, Jan.-Mar. 2004.

[5] J. Zhou and M. Gurcan, "An improved multicode CDMA transmission method for Ad Hoc networks," *Proc. of IEEE WCNC 2009*, pp. 1-6, April 2009.

[6] L. Luo, J. Zhang and Z. Shi, "Novel block-interleaved multi-code CDMA system for UWB communications," *Proc. of IEEE ICUWB 2007*, pp. 648-652, Sept. 2007.

[7] Y. Ma, J. Jin and D. Zhang, "Throughput and channel access statistics of generalized selection multiuser scheduling," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 2975-2987, August 2008.

[8] C. D. Iskander, "Performance of multicode DS/CDMA with noncoherent M-ary orthogonal modulation in the presence of timing errors," *IEEE Trans. Vehicular Techno.*, vol. 57, no. 6, pp. 3867-3874, Nov. 2008.

[9] 3GPP, "Spreading and modulation (FDD)," *3GPP TS25.213*, v3.4.0, Dec. 2000.

[10] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming multimedia users in high-speed downlink packet access (HSDPA)," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 844-855, August 2006.

[11] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling for wireless systems with multiple interfaces and multiple constraints," *Proc. of the 6th ACM/SIGCOM MSWiM*, pp. 11-19, Sep. 2003.

[12] S. S. Kulkarni and C. Rosenberg, "Opportunistic scheduling policies for wireless systems with short term fairness constraints," *Proc. of IEEE GLOBECOM 2003*, pp. 533-537, Dec. 2003.

[13] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantee in adaptive wireless networks," *IEEE J. Selected Areas Commun.*, vol. 23, no. 5, pp. 1056-1066, May 2005.

[14] L. Alonso, and R. Aqusti, "Automatic rate adaptation and energy-saving mechanisms based on cross-layer information for packet-switched data networks," *IEEE Commun. Magaz.*, vol. 42, no. 3, pp. S15-S20, Mar. 2004.

[15] Y. Li and G. Zhu, "M-gated scheduling and cross-layer design for heterogeneous services over wireless networks," *IEEE Trans. Vehicular Technol.*, vol. 58, no. 4, pp. 1983-1997, May 2009.

[16] V. Cirvino, V. Tralli and R. Verdone, "Cross-layer radio resource allocation for multicarrier air interfaces in multicell multiuser environments," *IEEE Trans. Vehicular Technol.*, vol. 58, no. 4, pp. 1864-1875, May 2009.

[17] Q. Liu, S. Zhou and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142-1153, May 2005.

[18] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, May 2005.

[19] J. Ramis, L. Carrasco, and G. Femenias, "A two-dimensional markov model for cross-layer design in AMC/ARQ-based wireless networks," *Proc. of IEEE GLOBECOM 2008*, pp. 1-6, Nov. 2008.

[20] H. Chen, H. C. B. Chan and V. C. M. Leung, "Cross-layer enhanced real-time packet scheduling over CDMA networks," *Proc. of IEEE ICON2006*, pp. 1-6, Sept. 2006.

[21] H. Chen, H. C. B. Chan, V. C. M. Leung and J. Zhang, "Cross-layer enhanced uplink packet scheduling for multimedia traffic over MC-CDMA networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 986-992, Feb. 2010.

[22] M. Schwartz, *Broadband Integrated Networks*: Prentice Hall, 1996.

[23] J. N. Daigle and J. D. Langford, "Models for analysis of packet-voice communication systems," *IEEE J. Sel. Areas Commun.*, vol. 4, no. 6, pp. 847-855, Sep. 1986.

[24] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 446-459, Aug. 1993.

[25] A. T. Anderson and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 719-732, Jun. 1998.

[26] E. P. C. Kao, *An introduction to stochastic processes*, Duxbury Press, 1996.

[27] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley & Sons, New York, 1994.

[28] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless multimedia networks," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 146-158, April 1999.

Chapter 5

A Cross-layer Scheduling Scheme for Wireless Multimedia Transmissions with Adaptive Modulation and Coding [1]

The previous two chapters applied the cross-layer consideration of the QoS in multimedia transmissions over wireless links in MC-CDMA systems. Note that, though some other systems do not have the multiple packet reception capability, they can also utilize the cross-layer QoS consideration to improve the system performance in terms of QoS provisioning, channel throughput and utilization. In this chapter, we propose a novel scheduling algorithm called QoS-CLS for serving real-time multimedia traffic flows in wireless networks employing AMC. Unlike other algorithms, QoS-CLS is based on the cross-layer consideration of QoS. The modulation and coding pair (we also call it the AMC mode in this chapter) is not selected based on the physical

[1] A version of this chapter has been published as a conference paper and awarded the best paper of the conference. H. Chen, H. C.B. Chan and V. C.M. Leung, "A cross-layer scheduling scheme for wireless multimedia transmissions with adaptive modulation and coding," *Proc. of IEEE Broadnets*, Sep. 2008.

layer FER requirements only. Instead, it is selected using cross-layer information to achieve the best channel throughput while satisfying the FLR requirement, which is a key QoS parameter for the user application. As a result, the overall channel throughput or system performance can be greatly enhanced. QoS-CLS also takes into account the dynamics of the channel state, traffic state and buffer state of the system in order to implement an optimal decision policy. We present results to show that this cross layer consideration of QoS can greatly improve the channel throughput. Maintaining fairness is another important issue in a multi-user system. QoS-CLS can incorporate different definitions of fairness. First, for some multimedia traffic which is not only delay sensitive but also FLR sensitive, QoS-CLS can guarantee the FLR of such traffic flows by keeping it below some preset thresholds. Second, for traffic flows without hard FLR guarantee requirements, QoS-CLS adopts the proportional distribution of FLR, which was also used in previous work [24] (Note that [24] did not consider possible cross-layer optimization between FER and FDR). By analyzing the dynamics of the system, we formulate the scheduling problem and AMC mode selection as an MDP with the aim of maximizing the channel throughput. Then, by transforming the MDP to a linear programming optimization problem, we can add the required QoS constraints to guarantee the FLR of some specific traffic flows, and/or to proportionally distribute the FLR among other traffic flows according to fairness requirements. We also provide two approximation methods to reduce the computational complexity of the optimization. To facilitate implementation, the scheduling policy can be pre-calculated and applied later in real-time.

The remainder of the chapter is organized as follows. Section 5.1 introduces the related work. Section 5.2 presents the system models used in this chapter. Section 5.3 studies the basic dynamics of the system. Section 5.4 presents the MDP model and the linear programming model

for QoS-CLS. Section 5.5 proposes two approximation methods for the algorithm to address the computational complexity. Section 5.6 presents the numerical results and discussions. Section 5.7 concludes the chapter.

5.1 Related Work

AMC is widely used in today's wireless communication systems. Because of its capability to maximize system efficiency over the varying mobile channel, AMC has been adopted in the physical layer of several important standards, including 3GPP [1, 2], 3GPP2 [3, 4], HIPERLAN/2 [5], IEEE 802.11 [6, 7], IEEE 802.15.3 [8], and IEEE 802.16 [9]. AMC is often applied in combination with different advanced transmission methods such as multi-carrier code division multiple access, multiple-input multiple-output, and cooperative transmissions. When scheduling user transmissions in AMC-based systems, the overall system throughput can be improved by selecting the user with the best channel status to transmit; however, such a strategy may result in unfairness and QoS degradation problems, particularly for real-time multimedia traffic [10]. Hence there is a great need to design a fair and efficient scheduling algorithm to support multimedia communications over AMC-based systems.

Note that real-time multimedia traffic is usually delay sensitive [20]. In this chapter, we assume that each multimedia data frame has a certain delivery deadline, and the frame is dropped from the buffer if it cannot be sent when the specified deadline expires. For other non delay sensitive traffic, although the frames do not have any deadline, they may also be dropped due to buffer overflow. We define the fraction of frames dropped at the data link layer as the FDR, which is a key data link layer QoS measure. In this chapter, we consider a system with

data link retransmission capability. In this case, an error frame will always be returned to the buffer for retransmission, if possible. That means, a frame is only dropped due to deadline expiration or buffer overflow. In such a system, the total frame error or drop ratio experienced by the upper layer, which is defined as FLR, is effectively equivalent to FDR. We also define the FER as the probability that a data frame is transmitted but received with uncorrectable errors. Note that when the FER of the physical channel becomes worse due to poor channel quality, the FDR may also increase because more retransmissions will effectively reduce the available bandwidth for normal transmissions thus increasing the chance of buffer overflow or deadline expiration. In this chapter the QoS requirement of a multimedia traffic stream is specified by the maximum FLR and maximum delay (i.e., the deadline). Our objective is to design an efficient scheduling algorithm that satisfies QoS requirements of traffic flows while optimizing the channel throughput.

In recent years, many scheduling algorithms have been proposed for AMC-based systems. In the proportional fairness [11] and score-based fairness [12], a weight is calculated for each user based on present or historical statistics of the channel quality and average throughput. By doing so, these algorithms seek to enhance the overall channel utilization while preventing any user from hogging the channel. However, these algorithms cannot provide a good service to real-time multimedia traffic because they do not consider the delay requirements of traffic flows. In the modified proportional fairness [13, 14], maximum carrier-to-interference ratio with early delay notification [15] and channel quality-based minimum throughput assurance [16] algorithms, the delay requirements or current delays experienced by users are used to prioritize the traffic flows for scheduling purposes. However, they cannot precisely control the delay experienced by each packet. Moreover, they operate at the expense of sacrificing system efficiency. In [17], the

fundamental scheduling algorithm and the advanced scheduling algorithm (ASA) are proposed for non-real-time and real-time traffic, respectively. However, ASA cannot effectively handle a situation that more than one real-time traffic flows have urgent packets to be sent out. In [18], a weighted round-robin scheduling algorithm is proposed to distribute the average delay among different traffic flows by studying the queue dynamics. Both [17] and [18] cannot provide precise QoS guarantee for real-time traffic flows. In [19], the required QoS can be guaranteed, but the proposed scheduling method must be used in conjunction with a very conservative admission control that ensures the system can satisfy the bandwidth requirement of all real-time traffic flows even when all these users experience the worst channel status simultaneously. Also, it requires a large scheduling cycle which can be precisely shared proportionally among the traffic flows.

In addition to deciding which traffic flow to serve next, the scheduling algorithm in an AMC-based system also needs to determine the AMC mode to be used for transmission of the traffic flow. In most systems, an AMC mode is chosen to keep the physical layer FER below the required threshold of the traffic flow without considering the FDR. However, this approach may not work effectively if a large number of frames are queued at the link layer, in which case it may be better to select an AMC mode with a higher transmission rate to clear the outstanding frames as quickly as possible. Although this selection may cause the FER to increase, the FLR may be greatly reduced to give a better overall system performance. Some other work also investigated ways to improve system performance by jointly considering FER and FDR. In [21] and [22], the upper and lower SINR bounds for each AMC mode are determined by studying the maximum retransmission time or the queue and channel dynamics of the system, respectively. However, neither of these schemes makes decision based on the current buffer

state. Also, they do not consider the essential scheduling problem—to decide which traffic flow to serve. In [23], a scheduling algorithm is proposed for multi-code code division multiple access systems. The current buffer status is utilized to determine the number of packets from each user to be transmitted in one TDMA frame, so as to minimize the overall FLR. However, the decision depends only on the current buffer state and does not consider the traffic models or statistics.

5.2 System Description

We consider downlink scheduling in a wireless network employing AMC. The system consists of a base station at the centralized cell site broadcasting to a number of mobile terminals in the cell. As in HSDPA and code division multiple access — high data rate (CDMA-HDR), we assume that there is a downlink data channel that is organized in fixed size transmission time intervals (TTI) with length W_0. At the beginning of each TTI, the system determines the traffic flow to be served and the AMC mode to be used during the TTI. We also assume that an error-free feedback channel exists from each mobile terminal to the base station to convey the channel state information (CSI) and acknowledge successful transmissions. To ensure the effective operation of the system, we also consider that the system is provided with an appropriate admission control algorithm for admitting traffic flows and an effective traffic policing mechanism to shape the traffic flows. Note that the FLR cannot be guaranteed for any traffic flow if the system admits too many traffic flows with FLR guarantee requirements. Also, without an effective traffic policing mechanism, the system cannot provide effective fairness among different traffic flows as one traffic flow may generate too much traffic to seriously affect the QoS of other traffic flows.

We assume that there are J traffic flows in the system. We also define the traffic flow scheduled for transmission in TTI t as $d(t) \in \{1, \ldots, J\}$.

We assume that there are K possible modulation and coding pairs (AMC modes) in the system. For AMC mode $i \in \{1, \ldots, K\}$, the transmission rate is R_i and at most α_i frames can be transmitted in one TTI. Without loss of generality, we assume that $R_i < R_j \; \forall i < j$. So we also have $\alpha_i < \alpha_j$. We define the AMC mode selected for TTI t as $k(t) \in \{1, \ldots, K\}$.

For contemporary multimedia communications, the MMPP [25] is widely used to model different kinds of traffic, such as video, voice and best effort data. In this chapter, we assume that traffic flow i has M_i different states and data packets arrive according to a Poisson process with a distinct rate $D_{i,j}$ per TTI at each state j. To simplify the analysis, we assume in this chapter that the size of each packet can be exactly encapsulated into 1 data frame. By discretizing the model with the TTI, we can get the transition matrix $\Phi_i = \{\phi_{i,j,j'}\}$, where $\phi_{i,j,j'}$ is the probability that the traffic state of traffic flow i turns to j' in the next TTI given that the traffic state is j in the current TTI. Also, from the $D_{i,j}$ and Φ_i, it is easy to get the average data frame arrival rate D_i for traffic flow i. We denote the traffic state of traffic flow i at TTI t as $g_i(t) \in \{D_{i,1} \ldots D_{i,M_i}\}$.

We consider two different buffering mechanisms in this chapter. Class one (C1) buffering mechanism is suitable for real-time traffic flows. The delay of each frame must be strictly bounded. A fixed initial expiration counter EP_i is associated with each frame of traffic flow i upon arriving and it will decrease by 1 after each TTI. Traffic flow $d(t)$ is selected to transmit its frames by AMC mode $k(t)$ in TTI t. It will transmit $\alpha_{k(t)}$ frames (or all the frames in the buffer, whichever is less). Among them, those frames received with uncorrectable errors are returned to the buffer for retransmissions. If a frame's expiration counter is 1 and it is not

sent out or received correctly in the current TTI, it is dropped from the buffer. The expiration counters of all frames left in the buffers are decreased by 1 after the TTI. Also, some new frames may arrive in the buffers during the TTI. In practice, the buffer size is finite. So, we assume that all traffic flows are buffered separately and for traffic flow i, at most EB_i frames which have the same expiration time can be stored in the buffer. There is no specific delay guarantee for the class two (C2) buffering mechanism. Instead, there is a buffer of size EN_i for traffic flow i. During TTI t, some frames sent out may be returned to the buffer for retransmission because of uncorrectable errors. If the buffer is full, new frames will be dropped directly. This class of buffering mechanism is suitable for non real-time traffic or real-time traffic without a strict delay requirement. We assume that there are J_1 and J_2 ($J_1 + J_2 = J$) traffic flows employing C1 and C2 buffering mechanisms, respectively. Generally, the buffer state of traffic flow $i \in \{1 \dots J_1\}$ at the beginning of TTI t is defined as $b_i(t) = [b_{i,1}(t), b_{i,2}(t), \dots, b_{i,EP_i}(t)]$, where $b_{i,j}(t)$ is the number of frames with expiration counter j in the buffer. And the buffer state of traffic flow $i' \in \{J_1 + 1 \dots J\}$ at the beginning of TTI t is described by $n_{i'}(t)$, the number of frames in the buffer. We assume that the buffer states of all traffic flows are known by the system.

In this chapter, we assume that the channel of each user is a frequency flat fading channel, which remains unchanged in a TTI. It corresponds to a block fading mode, which is suitable for slowly varying wireless channels [26]. For such channels, the channel quality can be described by a single parameter of received SINR ϵ for each TTI. So, the general Nakagami-m model, which represents a large class of fading channels, is used to model the channel [27]. The received SINR ϵ of a traffic flow is a random variable with the density function

$$p_\epsilon(\epsilon) = \frac{m^m \epsilon^{m-1}}{\bar{\epsilon}^m \Gamma(m)} \exp(-\frac{m\epsilon}{\bar{\epsilon}}) \tag{5.1}$$

Here $\bar{\epsilon} = E\{\epsilon\}$ is the average received SINR, $\Gamma(m) = \int_0^\infty t^{m-1}dt$ is the Gamma function and m is the Nakagami fading parameter ($m \geq 1/2$). Since each AMC mode has different capability to resist noise and interference, we define the function f_i mapping the SINR ϵ to the FER for AMC mode i, i.e., if the SINR is ϵ, the FER under mode i is $f_i(\epsilon)$. Generally, the function $f_i(\epsilon)$ can be found by experiments. In [21], an approximation of the FER in the presence of additive white Gaussian noise (AWGN) is proposed as

$$f_i(\epsilon) = \begin{cases} 1, & \text{if } 0 < \epsilon < \iota_i, \\ \rho_i \exp(-\chi_i \epsilon), & \text{if } \epsilon \geq \iota_i, \end{cases} \tag{5.2}$$

where ρ_i, χ_i and ι_i are mode-dependent parameters. Table 5.1 provides these parameters for some popular AMC modes for 1080 bit data frames. Their accuracy has been verified in [21]. To analyze the channel dynamics and apply it in QoS-CLS, the Finite-State Markov Channel (FSMC) [28] model, which has been widely used in recent research on wireless channels, is employed in this chapter. Generally, the channel state is modeled in L levels. There are a upper bound ϵ_i and a lower bound ϵ_{i-1} of SINR for level i. So, for each TTI, the channel state is one of the L levels and it may transit to an adjacent level or remain in the same level in the next TTI [28]. Defining the transition probability that the state changes from i to j after a TTI as $\theta_{i,j}$, we can get the transition matrix $\Theta = \{\theta_{i,j}\}$. Also, we define the average FER for level i in AMC mode j as $F_{j,i}$. The determination of Θ and $F_{j,i}$ from the Nakagami-m model can be found in [22]. We denote the channel state of traffic flow i at TTI t as $c_i(t)$. Note that, in this chapter, we assume a frequency flat slow fading channel. However, for a frequency selective fast fading channel, as long as the channel statistics can be obtained and a FSMC model can still be formulated with the statistics, the algorithm proposed in this paper should still work. In this chapter, we assume that perfect CSI is available at the mobile terminal

receiver using a training-based estimation method and is conveyed to the base station through the feedback channel. The performance of the proposed scheme is related to the accuracy of the CSI estimation. An imperfect CSI estimation may lead to the degradation of the performance of the proposed scheme.

5.3 Basic System Dynamics

Based on the description above, we can get the dynamics of the channel state and the traffic state from Θ and Φ_i ($i \in \{1, \ldots, J\}$) easily. The buffer state dynamics is more complicated. It depends on not only the buffering mechanism but also the current channel state, traffic state, scheduled traffic flow and selected AMC mode.

We first study the buffer dynamics for the C1 buffering mechanism, i.e., the transition probability from $b_i(t)$ to $b_i(t+1)$ for traffic flow i. We assume that the corresponding channel state $c_i(t)$ and the traffic state $g_i(t)$ is known. The system determine that the traffic flow $d(t)$ sends its frames and the AMC mode $k(t)$ is applied in TTI t.

If $d(t) = i$, the number of frames transmitted in the TTI from traffic flow i is given as follows:

$$h_i(t) = \min\left\{\alpha_{k(t)}, \sum_{j=1}^{EP_i} b_{i,j}(t)\right\}$$

We consider that $u_{i,j}(t)$ of $b_{i,j}(t)$ frames with expiration counter j from traffic flow i are transmitted in TTI t. Therefore, we have

$$u_{i,j}(t) = \begin{cases} 0, & \text{if } 0 \le h_i(t) \le \sum_{j'=1}^{j-1} b_{i,j'}(t) \\ h_i(t) - \sum_{j'=1}^{j-1} b_{i,j'}(t), & \text{if } \sum_{j'=1}^{j-1} b_{i,j'}(t) < h_i(t) \le \sum_{j'=1}^{j} b_{i,j'}(t) \\ b_{i,j}(t), & \text{if } \sum_{j'=1}^{j} b_{i,j'}(t) < h_i(t) \end{cases}$$

Note that the transmitted frames with uncorrectable errors will be returned to the buffer. We consider that $z_{i,j}(t)$ of the $u_{i,j}(t)$ transmitted frames are returned to the buffer. Then we must have

$$z_{i,j}(t) = b_{i,j-1}(t+1) - (b_{i,j}(t) - u_{i,j}(t)) \quad \forall j \geq 2$$

So given $b_{i,j}(t)$, $c_i(t)$, $d(t)$ and $k(t)$, the probability that there are $b_{i,j-1}(t+1)$ frames with expiration counter $j-1$ in the buffer of traffic flow i in TTI $t+1$ is the probability that there are $z_{i,j}(t)$ frames errors from the $u_{i,j}(t)$ frame transmissions. So, we have the probability

$$\begin{aligned} w_i(t) &= P\{b_{i,1}(t+1), \ldots, b_{i,EP_i-1}(t+1) | b_i(t)\} \\ &= P\{z_{i,2}(t), \ldots, z_{i,EP_i}(t) | b_i(t)\} \\ &= \prod_{j=2}^{EP_i} BN(u_{i,j}(t), z_{i,j}(t), F_{k(t),c_i(t)}) \end{aligned} \tag{5.3}$$

where

$$BN(i,j,p) = \begin{cases} \binom{i}{j} p^j (1-p)^{i-j}, & 0 \leq j \leq i \text{ and } 0 \leq p \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

is the binomial probability function.

Besides the transmitted frames, some new frames will arrive in the buffer of traffic flow i with an initial expiration counter of EP_i. So the probability that there are $b_{i,EP_i}(t+1)$ frames with expiration counter EP_i in the buffer of traffic flow i is

$$\begin{aligned} q_i(t) &= P\{b_{i,EP_i}(t+1) | g_i(t)\} \\ &= \begin{cases} AR(b_{i,EP_i}(t+1), D_{i,g_i(t)}), & b_{i,EP_i}(t+1) \neq EB_i \\ \sum_{j=EB_i}^{\infty} AR(j, D_{i,g_i(t)}), & b_{i,EP_i}(t+1) = EB_i \end{cases} \end{aligned} \tag{5.4}$$

where

$$AR(i,j) = \frac{j^i \, e^{-j}}{i!}$$

is the Poisson probability function.

Then, we get the buffer dynamics as follows:

$$\begin{aligned} p^{buf1}(b_i(t+1), b_i(t)) \doteq & P\{b_i(t+1)|b_i(t)\} \\ = & w_i(t) \times q_i(t) \end{aligned} \tag{5.5}$$

We also analyze the frame losses in TTI t for traffic flow i given $b_i(t)$, $c_i(t)$, $g_i(t)$, $d(t)$ and $k(t)$. First, $F_{k(t),c_i(t)} \times u_{i,1}(t)$ data frames are expected to be dropped because they are transmitted with errors and do not have a chance to be retransmitted. Second, $b_{i,1}(t) - h_i(t)$ data frames will be dropped too after the TTI because of the deadline expiration if $b_{i,1}(t) > h_i(t)$. Third, j data frames may be dropped with probability $AR(j + EB_i, D_{i,g_i(t)})$ because of too many new frames arriving during the TTI. Taking into account the three cases, the expected number of data frame losses for traffic flow i in TTI t is

$$\begin{aligned} \eta_i(b_i(t)) = & F_{k(t),c_i(t)} \times u_{i,1}(t) \\ & + (b_{i,1}(t) - h_i(t))^+ + \sum_{j=1}^{\infty} j \times AR(j + EB_i, D_{i,g_i(t)}) \end{aligned} \tag{5.6}$$

On the other hand, if $d(t) \neq i$, we have a similar analysis and $h_i(t) = 0$, $u_{i,j}(t) = 0$. We have:

$$w_i^*(t) = \begin{cases} 1, & z_{i,j}(t) = b_{i,j-1}(t+1) - (b_{i,j}(t) - u_{i,j}(t)) = 0 \ \ \forall j \\ 0, & \text{otherwise} \end{cases} \tag{5.7}$$

and the buffer dynamics $p^{buf1*}(b_i(t+1), b_i(t)) = w_i^*(t) \times q_i(t)$ does not depend on $k(t)$. The expected frame drops

$$\eta_i^*(b_i(t)) = (b_{i,1}(t) - h_i(t))^+ + \sum_{j=1}^{\infty} j \times AR(j + EB_i, D_{i,g_i(t)}) \tag{5.8}$$

does not depend on $k(t)$ either.

Next we consider the buffer dynamics for C2 buffering mechanism. We want to work out the transition probability from $n_i(t)$ to $n_i(t+1)$ for each traffic flow i given $c_i(t)$, $g_i(t)$ $d(t)$ and $k(t)$.

If $d(t) = i$, the number of transmitted data frames of traffic flow i in TTI t is

$$l_i(t) = \min\{\alpha_{k_{d(t)}(t)}, n_i(t)\}$$

We assume that j of the $l_i(t)$ data frames are transmitted with uncorrectable errors and p new data frames arrive at the buffer in TTI t. It is clear that if $n_i(t+1) < EN_i$

$$j + p = n_i(t+1) - (n_i(t) - l_i(t))$$

and if $n_i(t+1) = EN_i$

$$j + p \geq n_i(t+1) - (n_i(t) - l_i(t))$$

Note that the probability of having j data frame errors from $l_i(t)$ transmissions is $BN(l_i(t), j, F_{k(t),c_i(t)})$. Furthermore, the probability that p data frames arrive during the TTI is $AR(p, D_{i,g_i(t)})$. So, by considering every possible j and p, we can obtain

$$p^{buf2}(n_i(t+1), n_i(t)) \doteq P\{n_i(t+1)|n_i(t)\}$$

$$= \begin{cases} \sum_{j=0}^{\min\{l_i(t),v_i(t)\}} BN(l_i(t), j, F_{k_{d(t)}(t),c_i(t)}) AR(v_i(t) - j, D_{i,g_i(t)}), & n_i(t) - l_i(t) \leq n_i(t+1) < EN_i \\ \sum_{j=0}^{l_i(t)} \sum_{p=v_i(t)-j}^{\infty} BN(l_i(t), j, F_{k_{d(t)}(t),c_i(t)}) AR(p, D_{i,g_i(t)}), & n_i(t+1) = EN_i \\ 0, & \text{otherwise} \end{cases} \tag{5.9}$$

where

$$v_i(t) = n_i(t+1) - (n_i(t) - l_i(t))$$

Note that in this buffering mechanism frames are only dropped due to buffer overflow (i.e., $j + p > EN_i - (n_i(t) - l_i(t))$). So, given $n_i(t)$, $c_i(t)$, $g_i(t)$, $d(t)$ and $k_{d(t)}(t)$, $(n_i(t) - l_i(t) + j + p - EN_i)^+$ data frames are dropped with probability

$$BN(l_i(t), j, F_{k_{d(t)}(t),c_i(t)})AR(p, D_{i,g_i(t)})$$

Considering every possibility of j and p, the expected number of data frame losses for traffic flow i in TTI t is

$$\eta_i(n_i(t)) = \sum_{j=0}^{l_i(t)} \sum_{p=0}^{\infty} AR(p, D_{i,g_i(t)})BN(l_i(t), j, F_{k(t),c_i(t)})(n_i(t) - l_i(t) + j + p - EN_i)^+ \quad (5.10)$$

Similar to the analysis of the C1 buffering, if $d(t) \neq i$, both the buffer dynamics and the expected frame drops do not depends on $k(t)$ and we have:

$$\begin{aligned} & p^{buf2*}(n_i(t+1), n_i(t)) \doteq P\{n_i(t+1)|n_i(t)\} \\ &= \begin{cases} AR(n_i(t+1) - n_i(t), D_{i,g_i(t)}), & n_i(t) \leq n_i(t+1) < EN_i \\ \sum\limits_{p=n_i(t+1)-n_i(t)}^{\infty} AR(p, D_{i,g_i(t)}), & n_i(t+1) = EN_i \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (5.11)$$

and

$$\eta_i^*(n_i(t)) = \sum_{p=0}^{\infty} AR(p, D_{i,g_i(t)})(n_i(t) + p - EN_i)^+ \quad (5.12)$$

5.4 Markov Decision Process for Scheduling and AMC Mode Selection

With the analysis above, now we construct an MDP to determine the optimal decision policy for scheduling traffic flows and selecting the AMC mode. We will show how QoS-CLS maximizes the channel throughput while supporting FLR guarantee and fair FLR distribution. Following the notations in [29], we describe the MDP as follows:

5.4.1 State Space

A system state is defined as the row vector of channel states, traffic states and buffer states of all traffic flows sharing the channel. The system state at the beginning of TTI t is

$$\begin{aligned} s(t) =&[c_1(t), \ldots, c_J(t), g_1(t), \ldots, g_J(t), \\ & b_{1,1}(t), \ldots, b_{1,EP_1}(t), b_{2,1}(t), \ldots, b_{J_1,EP_{J_1}}(t), \\ & n_{J_1+1}(t), \ldots, n_J(t)]. \end{aligned}$$

The state space S comprises all possible state vectors, where a state vector has a certain traffic state, channel state and buffer state for each of J traffic flows. So the state space is defined as

$$\begin{aligned} S=\{&s = [c, g, b, n] : \\ & c = [c_1, \ldots, c_J] \in \{1, 2, \ldots, L\}^J \\ & g = [g_1, \ldots, g_J] \in \Pi_{j=1}^{J} \{1, 2, \ldots, M_j\} \\ & b = [b_{1,1}, \ldots, b_{J_1,EP_{J_1}}] \in \Pi_{j=1}^{J} \{1, 2, \ldots, EB_{J_1}\}^{EP_j} \\ & n = [n_{J_1+1}, \ldots, n_J] \in \Pi_{j=J_1+1}^{J} \{1, 2, \ldots, EN_j\}\} \end{aligned} \tag{5.13}$$

5.4.2 Decision Epochs and Actions

The decision epochs are the beginning of the TTIs. An action for TTI t is described as $a(t) = [d(t), k_{d(t)}(t)]$, i.e., in TTI t, traffic flow $d(t)$ is chosen to transmit with AMC mode $k_{d(t)}(t)$. The action space is the set of all possible actions, which can be defined as:

$$A = \{a = [d, k] : d \in \{1, \ldots, J\}, k \in \{1, \ldots, K\}\} \tag{5.14}$$

5.4.3 State Dynamics

Since the probability of state $s(t+1)$ only depends on state $s(t)$ and action $a(t)$ as discussed in Section 5.3, the dynamics of the system state is a MDP, where the state transition probabilities are:

$$\begin{aligned} & p_{s(t)s(t+1)}(d(t), k_{d(t)}(t)) = P\{s(t+1)|s(t), d(t), k(t)\} \\ =& p^{buf1}(b_{d(t)}(t+1), b_{d(t)}(t)) \prod_{j=1\ldots J_1, j\neq d(t)} p^{buf1*}(b_j(t+1), b_j(t)) \\ & p^{buf2}(n_{d(t)}(t+1), n_{d(t)}(t)) \prod_{j=J_1+1\ldots J, j\neq d(t)} p^{buf2*}(n_j(t+1), n_j(t)) \\ & \prod_{j=1}^{J} \theta_{c_j(t), c_j(t+1)} \phi_{j, g_j(t), g_j(t+1)} \end{aligned} \tag{5.15}$$

It is clear that the embedded chain modelled above is a unichain [29].

5.4.4 Policy, Performance Criterion and Cost Function

The system must decide the action for any TTI based on the system state at the TTI. For deterministic decisions, an action $a \in A_s$ is chosen for each given state $s \in S$ according to a decision policy $\in \Psi$ where Ψ is defined as

$$\Psi = \{ \quad : S \to A| \quad _s \in A_s, \forall s \in S\}$$

To optimize the QoS of the system, we define the cost criterion as follows. For any deterministic policy $\in \Psi$ and an initial state $s(t_0)$, the average cost is

$$\Delta\ (s(t_0)) = \lim_{T\to\infty} \frac{1}{T}\mathbf{E}\Big\{ \sum_{t=0}^{T} \delta(s(t), a(t))\Big\} \tag{5.16}$$

where $\mathbf{E}\{\cdot\}$ denotes the expected value and $\delta(s(t), a(t))$ is the expected cost for a TTI given that the state at the beginning of the TTI is $s(t)$ and the action is $a(t)$. The aim here is to choose an optimum policy to minimizes $\Delta\ (s(t_0))$ for any initial state $s(t_0)$. As stated in above section, the aim of CLS-QoS is to optimize the channel throughput subject to certain QoS constraints or requirements. Note that, given the traffic source rates of the traffic flows, if less frames are lost, the channel throughput will be higher. So here we choose the following cost function to minimize the total data frame losses:

$$\delta(s(t), a(t)) = \begin{cases} \eta_{d(t)}(b_{d(t)}(t)) + \sum\limits_{j=1\ldots J_1, j\neq d(t)} \eta_j^*(b_j(t)) + \sum\limits_{j=J_1+1\ldots J} \eta_j^*(n_j(t)) & d(t) \leq J_1 \\ \eta_{d(t)}(n_{d(t)}(t)) + \sum\limits_{j=1\ldots J_1} \eta_j^*(b_j(t)) + \sum\limits_{j=J_1+1\ldots J, j\neq d(t)} \eta_j^*(n_j(t)) & d(t) > J_1 \end{cases} \tag{5.17}$$

5.4.5 Constraints

To achieve specific QoS requirements, we need to set up constraints to address the FLR guarantees and proportional distribution requirements. Without loss of generality, we assume that traffic flows $\beta_1, \beta_2, \ldots, \beta_\zeta$ need FLR guarantee $y_{\beta_1}, y_{\beta_2}, \ldots, y_{\beta_\zeta}$, and the remaining traffic flows $\xi_1, \xi_2, \ldots, \xi_\mu$ need to have their FLR proportionally distributed with weight $\varpi_{\xi_1}, \varpi_{\xi_2}, \ldots, \varpi_{\xi_\mu}$. Then for a C1 traffic flow $\tau \in \beta_1 \ldots \beta_\zeta$, we have

$$r_\tau = \lim_{T\to\infty} \frac{1}{T}\mathbf{E}\big\{\eta_\tau(b_\tau(t)) \times p\{d(t) = \tau\} + \eta_\tau^*(b_\tau(t)) \times (1 - p\{d(t) = \tau\})\big\} \leq D_\tau y_\tau \tag{5.18}$$

Also, for any two traffic flows $\tau_1, \tau_2 \in \xi_1 \dots \xi_\mu$, we have

$$\frac{r_{\tau_1}}{\varpi_{\tau_1} D_{\tau_1}} - \frac{r_{\tau_2}}{\varpi_{\tau_2} D_{\tau_2}} = 0 \tag{5.19}$$

5.4.6 Linear Programming Solution to The MDP

Based on the above discussion, the optimal policy of the MDP can be obtainted by solving the following linear programming.

$$\min_{x_{sa} \geq 0} \sum_{s \in S} \sum_{a \in A_s} x_{sa} \delta(s, a)$$

subject to

$$\sum_{a \in A_s} x_{sa} - \sum_{s' \in S} \sum_{a \in A_{s'}} x_{s'a} p_{s's}(a) = 0 \; \forall s \in S, \tag{5.20}$$

$$\sum_{s \in S} \sum_{a \in A_s} x_{sa} = 1, \tag{5.21}$$

$$\sum_{s \in S} \sum_{a \in A_s} x_{sa} \eta_\tau(s, a) \leq D_\tau \times y_\tau, \; \forall \tau \in \beta_1 \dots \beta_\zeta \tag{5.22}$$

$$\frac{\sum_{s \in S, a \in A_s} x_{sa} \eta_{\tau_1}(s, a)}{\varpi_{\tau_1} D_{\tau_1}} - \frac{\sum_{s \in S a \in A_s} x_{sa} \eta_{\tau_2}(s, a)}{\varpi_{\tau_2} D_{\tau_2}} = 0, \; \forall \tau_1, \tau_2 \in \xi_1 \dots \xi_\mu \tag{5.23}$$

Here if traffic flow τ uses the C1 buffering mechanism, we have $\eta_\tau(s, a) = \tau(b_\tau)$, otherwise we have $\eta_\tau(s, a) = \tau(n_\tau)$. Note that x_{sa} is the optimization variable, which is the probability that the system state is s and the action is a. Specifically, the action a will be chosen with probability $x_{sa} / \sum_{a' \in A_s} x_{sa'}$ when the system state is s.

Note that the above optimization objective makes sure that the number of data frame losses in the system is minimized subject to the QoS constraints. So the throughput of the system is effectively optimized. Also, constraint (5.22) ensures that the FLR of specific traffic flows are kept under their required thresholds. Constraint (5.23) ensures that the other traffic flows are served with FLR proportionally distributed.

Since the system cannot distinguish two system states which have the same buffer and channel states (i.e., even though they have different traffic states), the actions mapped from such two system states should be the same. We can add a constraint (5.24) addresses this practical requirement. However, the optimization is not linear anymore in this case.

$$x_{sa}/\sum_{a' \in A_s} x_{sa'} = x_{s'a}/\sum_{a' \in A_{s'}} x_{s'a'}, \text{ if } c = c', b = b', n = n'. \tag{5.24}$$

5.5 Implementation Issues

Due to the complexity of the MDP, there are implementation challenges for QoS-CLS. In practice, the scheduling decisions must be made quickly in real-time. In QoS-CLS, the scheduling of every TTI can be based on a simple table lookup. When a traffic flow arrives or departs, the system updates the look-up table by a precalculated scheduling policy and also prepares scheduling policies for situations where any traffic flow departs or another new traffic flow arrives, which will be used upon next arrival or departure. We also propose two approximate methods to reduce the complexity of the optimization.

5.5.1 Reduced Buffer State Space

Note that the state space of the MDP for QoS-CLS may become extremely large if some traffic flows have big buffers. So we first try to reduce the buffer states space of traffic flows with C1 buffering mechanism. For a C2 traffic flow i ($i \in \{J_1 + 1, \ldots, J\}$), the size of the buffer is EN_i. So in the original optimization, the buffer state $n_i(t)$ can be any number between 0 to EN_i. When EN_i is too big, we divide the $EN_i + 1$ states into a small number w_i of groups ($v_{i,1}$ to v_{i,w_i}), and we consider each group $v_{i,z}$ to represent a buffer state. As a result, the buffer state

$n'_i(t)$ can be any of the states $v_{i,z}$ ($z \in \{1, \ldots, w_i\}$). To apply the optimization based on the new buffer state definition, we need to have the buffer state transition probability $P\{n'_i(t+1)|n'_i(t)\}$. To make the illustration simple, we assume that $EN_i + 1$ can be divided by w_i and the buffer lengths $(EN_i + 1)/w_i \times (z-1)$ to $(EN_i + 1)/w_i \times z - 1$ belong to group $v_{i,z}$ ($z \in \{1, \ldots, w_i\}$) (a similar result can also be achieved if the buffer length is grouped in a different way). So the buffer state dynamics can be calculated as:

$$\begin{aligned} P\{n'_i(t+1)|n'_i(t)\} &= \sum_{n_i(t+1) \in n'_i(t+1)} P\{n_i(t+1)|n'_i(t)\} \\ &= \sum_{n_i(t+1) \in n'_i(t+1)} \sum_{n_i(t) \in n'_i(t)} P\{n_i(t+1)|n_i(t)\} \times P\{n_i(t)|n'_i(t)\}. \end{aligned} \tag{5.25}$$

Note that, when the buffer size EN_i is very large, the probabilities that the buffer has a specific length x or $x+1$ should be very close to each other. So we assume that the probability that the buffer is in a state $v_{i,z}$ is fairly distributed inside the group. We will show some results to validate this assumption later. As a result, we have:

$$P\{n_i(t)|n'_i(t)\} \approx \frac{w_i}{EN_i + 1}. \tag{5.26}$$

And the equation (5.25) becomes:

$$P\{n'_i(t+1)|n'_i(t)\} \approx \frac{w_i}{EN_i + 1} \sum_{n_i(t+1) \in n'_i(t+1)} \sum_{n_i(t) \in n'_i(t)} P\{n_i(t+1)|n_i(t)\}. \tag{5.27}$$

Note that $P\{n_i(t+1)|n_i(t)\}$ has been calculated in equation (5.11). Also, we can get the expected number of data frame losses for traffic flow i in TTI t as:

$$\eta'_i(n'_i(t)) = \sum_{n_i(t) \in n'_i(t)} P\{n_i(t)|n'_i(t)\}\eta_i(n_i(t)) \approx \frac{w_i}{EN_i + 1} \sum_{n_i(t) \in n'_i(t)} \eta_i(n'_i(t)) \tag{5.28}$$

With the above approximation, the state space and the computation complexity for the optimization presented in Section 5.4 can be greatly reduced.

5.5.2 Decomposition of The Optimization

Another approximate method proposed here is to decouple the optimization problem into two smaller problems and deal with them separately. Note that, given $c_i(t)$, $g_i(t)$, and $b_i(t)$ or $n_i(t)$ for every traffic flow i, we need to determine both the scheduled traffic flow and the selected AMC mode for the flow. Instead of optimizing them together, we first determine the AMC mode $k_i(t)$ for each traffic flow i if it needs to transmit data frames in TTI t.

When the aggregate traffic load is not too heavy, the data frame drop rate from each traffic flow should not be very large. In such a situation, the probability of channel access should be proportionally distributed to every traffic flow according to its average traffic rate. So we assume that the probability of $d(t) = i$ for any t in the optimization is:

$$P\{d(t) = i\} = \frac{D_i}{\sum_{i'=1}^{J_1+J_2} D_{i'}}. \tag{5.29}$$

As a result, we can get the buffer dynamics for traffic flow i as:

$$P\{b_i(t+1)|b_i(t)\} =$$

$$P\{d(t) = i\} \times p^{buf1}(b_i(t+1), b_i(t)) + (1 - P\{d(t) = i\}) \times p^{buf1*}(b_i(t+1), b_i(t)) \tag{5.30}$$

And the expected data frame drops is:

$$\eta_i'(n_i'(t)) = P\{d(t) = i\} \times \eta_i(n_i'(t)) + (1 - P\{d(t) = i\}) \times \eta_i^*(n_i'(t)) \tag{5.31}$$

Note that, with the above approximation, the buffer dynamics and expected data frame drops for traffic flow i do not depend on $d(t)$ any more. So we can follow the same steps described in Section 5.4 to formulate a MDP for the traffic flow i only. It can give a stochastic mapping between the current system state for traffic flow i ($c_i(t)$, $g_i(t)$, $b_i(t)$ or $n_i(t)$) and

$P\{k_i(t)|c_i(t), g_i(t), b_i(t)\}$, the probability of $k_i(t)$. Note that the dimension of this sub-problem is reduced by a factor of $\frac{c_i g_i EN_i}{\Pi_{j=1}^{J} c_j g_j \Pi_{j=1}^{J_1} EB_j \Pi_{j=J_1+1}^{J} EN_j}$ relative to the dimension of the original optimization problem.

Based on the above result, we can further optimize $d(t)$ using the MDP in Section 5.4. However, the actions are now only the decision for $d(t)$ in every TTI t since we already know the probability distribution of $k_i(t)$. So the dimension of the sub-problem is $\frac{1}{K}$ of the dimension of the original problem.

5.6 Results and Discussions

In this section, we study the performance of the proposed QoS-CLS algorithm. For the channel of each traffic flow, we assume that the average SINR of the channel is 15db. The Nakagami parameter m is 1. The channel is discretized into 5 levels with SINR bounds $[-\infty, 2, 5, 10, 20, \infty]$ db. We assume that the AMC modes employed in the system are the same as those listed in Table 5.1. The frame size is 1080 bits and the length of a TTI is 0.02 s. Each TTI can hold $[1, 2, 3, 6, 9]$ frames for modes 1 to 5. This setting provides a channel with a transmission rate ranging from 54kbps to 486kbps.

First, we study a system in which there is only one traffic flow. In this case, QoS-CLS optimizes the channel throughput by only selecting the most appropriate AMC mode for each TTI. The following system parameters are used in this study. The initial expiration counter for the C1 buffering mechanism is 2 and the buffer size for frames with the same deadline is 5. The buffer size for the C2 buffering mechanism is 10. The traffic flow has only one traffic state, i.e., it follows a Poisson arrival process. As a comparison, we also present the result for

a scheme where the AMC mode is simply chosen to guarantee the FER. Figure 5.1 shows the FLR of different traffic flows with varied buffering mechanisms and scheduling schemes. In the figure, "QoS-CLS" stands for results with the proposed QoS-CLS algorithm, "QoS-CLS, RBS" stands for results with the proposed algorithm and the reduced buffer state approximation, and "FER guaranteed" stands for results with conventional AMC mode selection scheme where the physical layer FER is kept below the threshold 0.01 (a very close result can be achieved with the threshold 0.001 and 0.0001 too). It is clear that the FLR achieved by QoS-CLS is much less than that achieved by the conventional FER guaranteed scheme. Also, we can observe that the results for C2 buffering scheme are much better than those for C1 buffering scheme, since C1 buffering mechanism drops frames not only where the buffer overflows but also when frame deadlines expire. However, for the FER-guaranteed scheme, when the traffic load exceeds 1.9 frame/TTI, the FLR is very close for both C1 and C2 buffering schemes. It is because the channel capacity is fixed and almost fully utilized for every TDMA frame so that the number of data frames transmitted should be the same for either C1 or C2 scheme, which results in the same FLR for both C1 and C2 schemes. Another fact we can observe in the figure is that the reduced buffer state approximation can provide results close to those of the original QoS-CLS. Figure 5.2 compares the channel throughput of QoS-CLS and the FER-guaranteed schemes. It is clear that with the increase of traffic load, the gain in throughput of QoS-CLS above that of the FER-guaranteed scheme increases. For example, if the FLR target is 0.01, QoS-CLS can achieve a throughput of 1.84 frame/TTI (with a traffic load of about 1.85) for C1 buffering and 2.3 frame/TTI (with a traffic load of about 2.32) for C2 buffering scheme. At the same time, the FER-guaranteed scheme can only achieve 1.4 frame/TTI (with a traffic load of about 1.4) for C1 buffering and 1.59 frame/TTI (with a traffic load of about 1.6) for C2 buffering

scheme. This represents a large improvement of 32% and 45% for C1 and C2 buffering scheme, respectively, relative to the performance of the FER-guaranteed scheme. Note that for the reduced buffer state approximation method, the maximum throughput for C2 buffing scheme with a 0.01 FLR target is about 2.23 frame/TTI, which is about 40% better than the maximum throughput of the FER-guaranteed scheme.

Next we consider a system with multiple traffic flows sharing the channel. We consider two types of traffic flows here. The first type of traffic flow is from a voice-over-IP (VoIP) session. The traffic can be represented by a two-state MMPP with an average arrival rate of 0 frame/TTI in state 1 (silent state) and 0.6 frames/TTI in state 2 (talk state). The average durations of state 1 and state 2 are 1.5s and 1s, respectively. The average traffic load of the traffic flow is 0.24 frames/TTI. This traffic flow employs the C1 buffering mechanism with an initial expiration counter of 1 TTI, i.e., a frame must be sent before the next frame arrives. The buffer size for frames with the same deadline is 6 (then the expected frame loss due to buffer overflow is about 0.0006 when the frame arrival rate is 1 frame/TTI). Also, this traffic flow has a FLR guarantee requirement of FLR_1 <0.01. The second type of traffic flow is represented by a Poisson process with the frame arrival rate 0.1 frame/TTI. All this type of traffic flows shares a buffer with size of 6 under C2 buffering mechanism. Type 2 traffic flows have no FLR guarantee but they need to have their FLR distributed proportionally.

To show the effectiveness of our proposed scheme, we compare QoS-CLS with some existing scheduling schemes. The first one we use as a benchmark for comparisons is opportunistic scheduling [30, 31] (denoted as "OS" in the figures). OS selects the traffic flow with the best channel condition to transmit in each TTI. Another scheme chosen for comparisons is the earliest deadline first [32] (denoted as "EDF" in the figures), which takes into account of transmission

deadlines. EDF schedules type 1 traffic flows for transmissions according to the numbers of head of line data frames, with those having the most head of line data frames scheduled for transmissions first. Type 2 traffic flows are only scheduled after all type 1 traffic flows have been scheduled, since type 2 data frames do not have transmission deadlines.

We first set up a system with one type 1 traffic flow and a variable number of type 2 traffic flows. The FLR of the type 1 traffic flow needs to be guaranteed to be less than 0.01. Figures 5.3 and 5.4 show the FLR for type 1 and type 2 traffic flows, respectively, when the number of the type 2 traffic flows is varied from 8 to 18. In Figure 5.3, results for "QoS-CLS" and "QoS-CLS, DO" show that the FLR of type 1 traffic flow can always be kept below 0.01 using these mechanisms. This demonstrates the correct operation of the proposed scheduling algorithm in terms of meeting QoS constraints. The results for OS show a much greater FLR for type 1 traffic flow since the scheme does not consider the stringent delay requirement and the buffering mechanism. Also, the EDF scheme achieve the better FLR result for the type 1 traffic flow than required because type 1 traffic flows always get access to the channel prior to other type 2 traffic flows. It actually wastes the channel resources and leads to a worse performance for the type 2 traffic flows. We can see clearly in Figure 5.4 that the FLR result of the EDF scheme is the worst one among all schemes. In Figure 5.4, we can also see that the OS scheme can provide a best performance for type 2 traffic flows because a type 2 traffic flow can transmit the most data frames in a TTI as long as its channel status is the best. If the target FLR is 1%, QoS-CLS can support 11 type 2 traffic flows and EDF can only support 8 type 2 traffic flows (a 38% capacity improvement). OS cannot achieve a good capacity in this situation at all for its poor support for real-time traffic flows. Also, the figures show that the approximation method DO can give similar results for type 1 traffic flows as the QoS-CLS scheme. But it produces

a little bit higher FLR for type 2 traffic flows compared with the QoS-CLS scheme. Figure 5.5 shows the corresponding average channel throughput for one TTI. It can be seen that the QoS-CLS scheme and its approximation scheme can provide a similar channel throughput that is a little bit better than both those of the OS and EDF schemes.

Next, we evaluate the effectiveness of fair FLR distribution supported by QoS-CLS. We still consider a system with two types of traffic flows as above. However, in this case the type 1 traffic flow does not require any FLR guarantee. Instead, the FLR is distributed fairly between the type 1 and type 2 traffic flows. The distribution weights of the two traffic flows are the same, which means that they should experience the same FLR according to the fairness requirement. We assume that there are one type 1 traffic flow and ten type 2 traffic flows. The frame arrival rate for a type 2 traffic flow is still 0.1 frames/TTI, but the frame arrival rate of the type 1 traffic flow in state 2 is varied from 0.5 to 1.5 frames/TTI. Figures 5.6 and 5.7 show the FLR and throughout results. In Figure 5.6, we can see that the FLR of the type 1 and 2 traffic flows are exactly the same for the QoS-CLS scheme and its approximation method. For the OS scheme, the FLR of type 1 traffic flow is very high and the FLR of type 2 traffic flows is very small. For the EDF scheme, the FLR of type 2 traffic flows are much higher than that of the QoS-CLS scheme or its approximation method, while the FLR of type 1 traffic flow is lower than that of the QoS-CLS scheme when the traffic load is relatively light, as EDF gives the type 1 traffic flow absolute priority over type 2 traffic flows. However, when the traffic load becomes heavy, the QoS-CLS scheme and its approximation method can still achieve a better FLR than that of EDF for the type 1 traffic flow, since the proposed schemes may choose a higher order AMC mode to transmit more frames per TTI to reduce the overall FLR. Figure 5.7 shows the corresponding throughout. Note that the QoS-CLS and its approximation method

can also provide higher throughput than those of the OS and EDF schemes, when fair FLR distribution is applied. In summary, the results from Figures 5.1 and 5.7 show that QoS-CLS can enhance the channel throughput as well as providing effective QoS provisioning.

5.7 Summary

In this chapter, we have proposed a novel scheduling algorithm called QoS-CLS for multimedia transmissions over wireless channels with adaptive modulation and coding. To satisfy user QoS requirements while maintaining fairness, a cross-layer framework is used for scheduling user transmissions and selecting the appropriate AMC mode. The scheduling problem is formulated as a MDP and solved by linear programming. As a result, the optimal scheduling policy can be pre-determined and stored in the system for traffic scheduling in real-time. Two approximation methods are also proposed to reduce the computational complexity of the optimization. Results show that QoS-CLS can achieve significant improvements in channel throughput compared to the opportunistic scheduling algorithm and the earliest deadline first scheme. Furthermore, it can effectively guarantee QoS (i.e., FLR) and maintain fairness.

Table 5.1: Parameters for Modulation and Coding Pairs in AMC

	1	2	3	4	5
Modulation	BPSK	QPSK	QPSK	16-QAM	64-QAM
Coding rate	1/2	1/2	3/4	3/4	3/4
Rate	0.50	1.00	1.50	3.00	4.50
ρ_k	274.7229	90.2514	67.6181	53.3987	35.3508
χ_k	7.9932	3.4998	1.6883	0.3756	0.0900
ι_k (db)	-1.5331	1.0942	3.9722	10.2488	15.9784

Table 5.2: List of Important Symbols

Symbol	**Desription**
W_0	Length of TTI of the downlink data channel
J	Number of traffic flow
K	Number of modulation and coding pairs
R_i	Transmission rate for mode i
α_i	Max number of frames per TTI for mode i
L	Number of channel states
Θ	Channel state transition matrix
$\theta_{i,j}$	probability for state to turn from i to j in one TTI
continued on next page	

Table 5.2: *continued*

Symbol	**Desription**
$F_{j,i}$	Average FER for channel state i in AMC mode j
M_i	Number of states for traffic flow i
$D_{i,j}$	Data frame arrival rate for traffic flow i in state j
Φ_i	Traffic transition matrix for traffic i
$\phi_{i,j,j'}$	Probability for traffic i to turn from state j to j'
$C1$	Class one buffering mechanism
$C2$	Class two buffering mechanism
J_1	Number of traffic flows for C1
J_2	Number of traffic flows for C2
EP_i	Initial expiration counter for C1 traffic flow i
EB_i	Buffer size for data frames with the same deadline for C1 traffic flow i
EN_i	Buffer size for C2 traffic flow i
$d(t)$	Traffic flow scheduled for TTI t
$k(t)$	Mode used for TTI t
$c_i(t)$	Channel state of traffic flow i at TTI t
$g_i(t)$	Traffic state of traffic flow i at TTI t
$b_i(t)$	Buffer state for C1 traffic flow i at TTI t
$b_{i,j}(t)$	Number of frames with expiration counter j for C1 traffic flow i
continued on next page	

Table 5.2: *continued*

Symbol	**Desription**
$n_i(t)$	Buffer state for C2 traffic flow i at TTI t

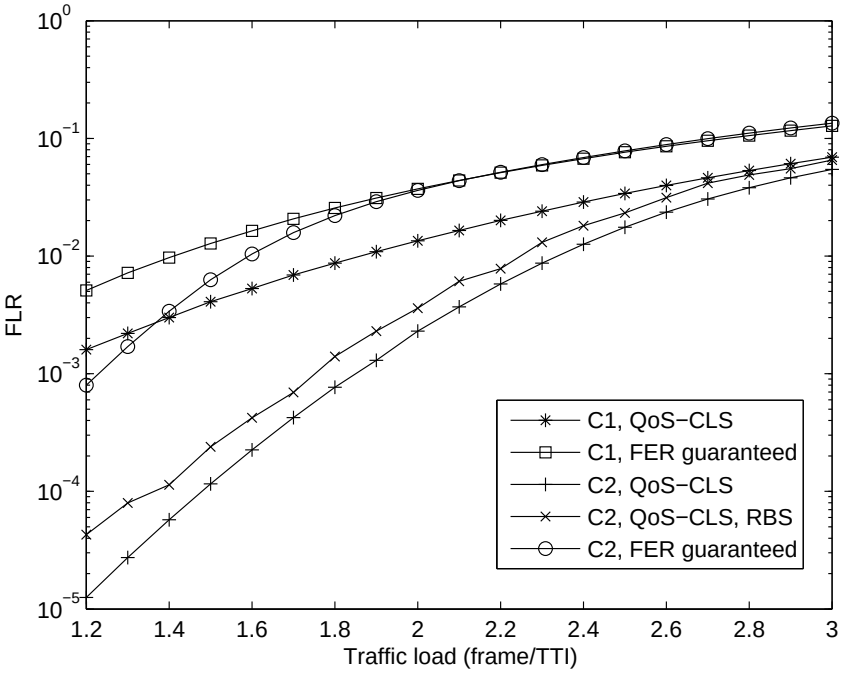

Figure 5.1: FLR comparison for single traffic flow.

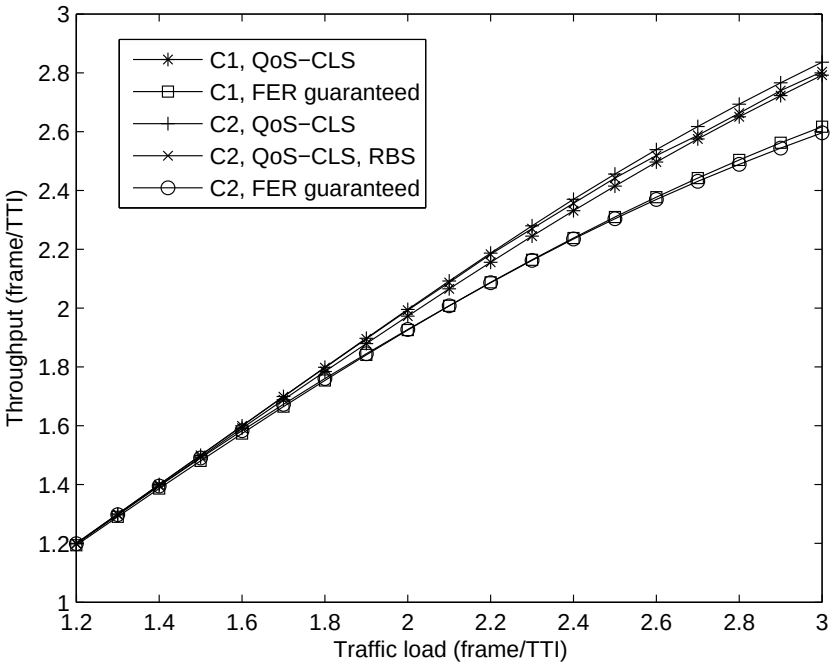

Figure 5.2: Throughput comparison for single traffic flow.

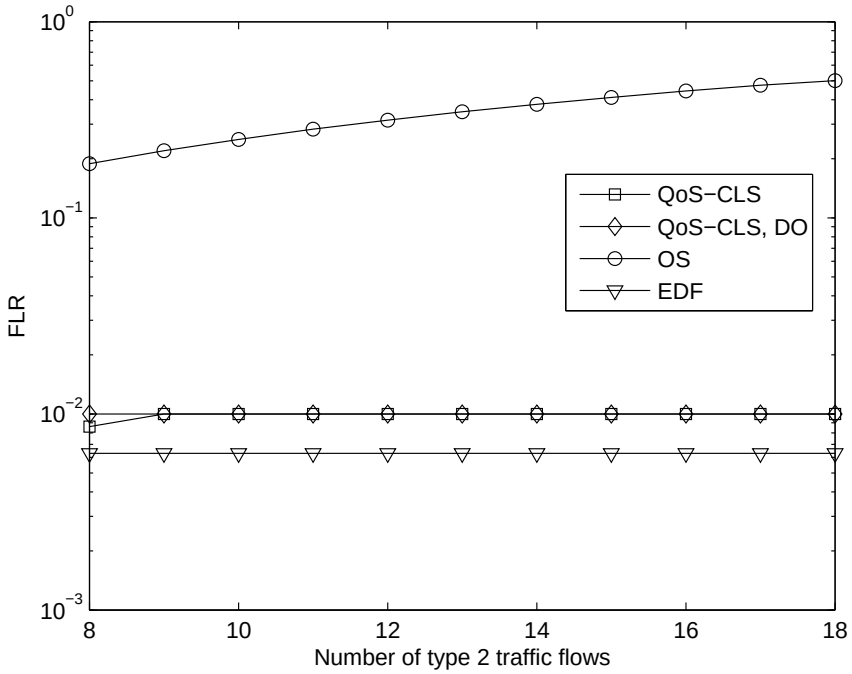

Figure 5.3: FLR comparison for the type 1 traffic flow in a system with FLR guarantee for the type 1 traffic flow.

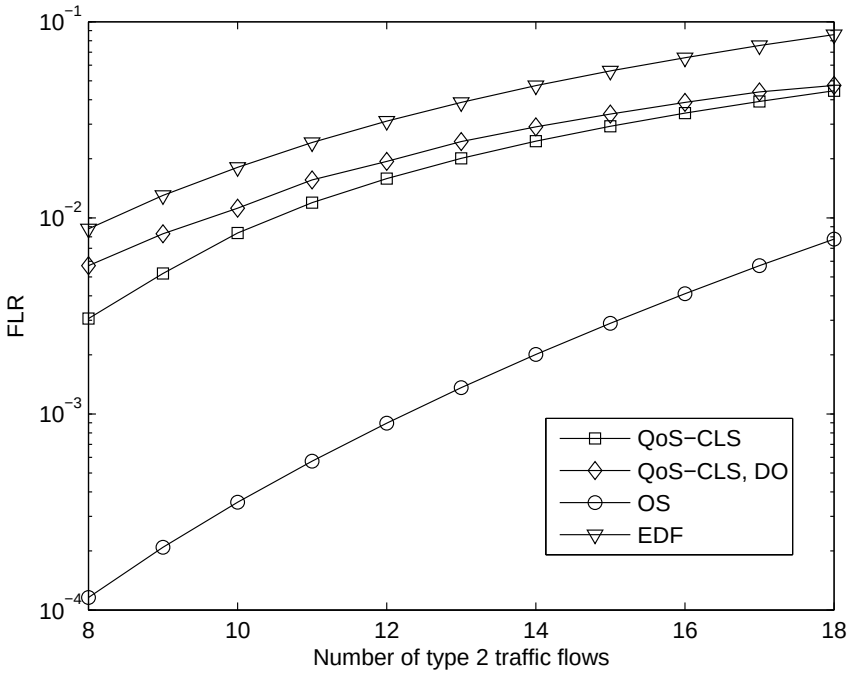

Figure 5.4: FLR comparison for the type 2 traffic flow in a system with FLR guarantee for the type 1 traffic flow.

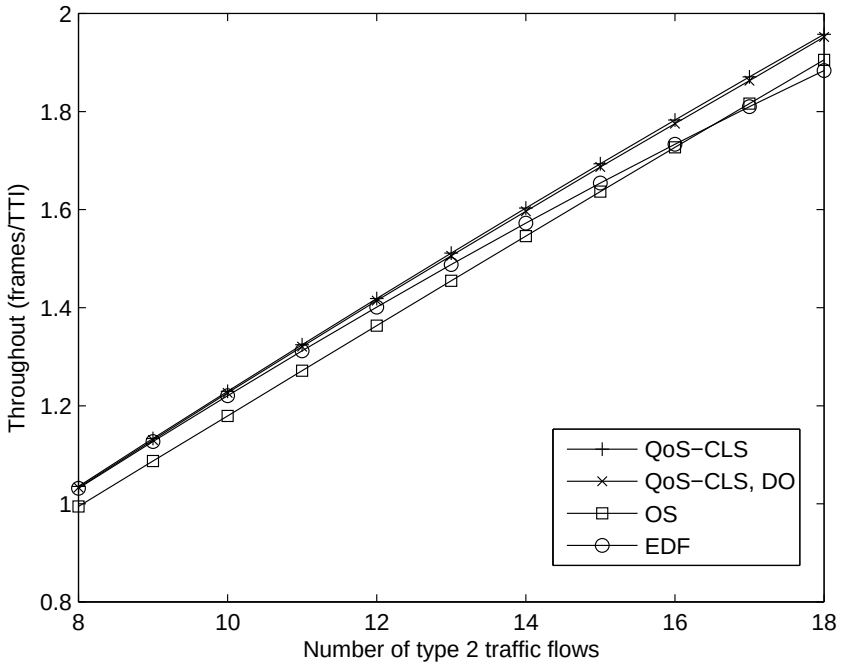

Figure 5.5: Throughput comparison for the system with FLR guarantee for the type 1 traffic flow.

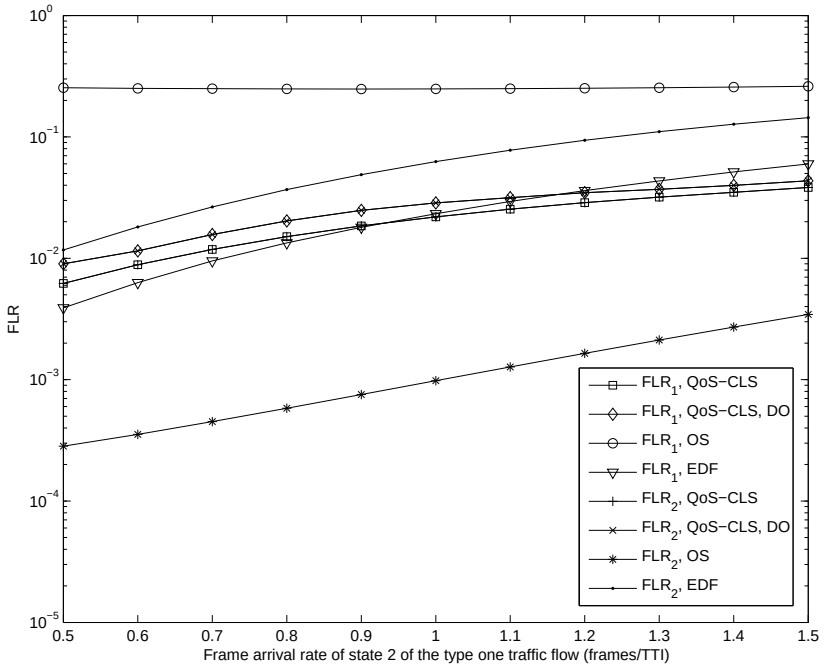

Figure 5.6: FLR comparison for the system with FLR sharing.

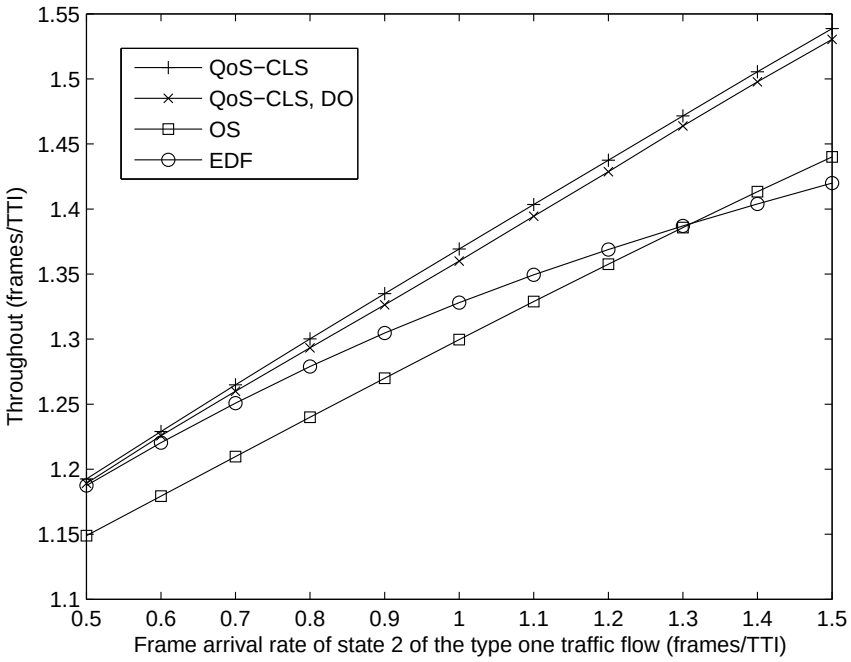

Figure 5.7: Throughput comparison for the system with FLR sharing.

Bibliography

[1] 3GPP, "Physical layer aspects of UTRA high speed downlink packet access," *3GPP TR 25.848*, V4.0.0, 2001.

[2] W. Xiao, A. Ghosh, D. Schaeffer, and L. Downing, "Voice over IP (VoIP) over Cellular: HRPD-A and HSDPA/HSUPA," *Proc. of IEEE VTC2005*, pp. 2785-2789, Sept. 2005.

[3] 3GPP2, "Physical layer standard for CDMA2000 spread spectrum systems," *3GPP2 C.S0002-0*, v1.0, July 1999.

[4] J. Yang, N. Tin and A. K. Khandani, "Adaptive modulation and coding in 3G wireless systems," *Proc. of IEEE VTC2002*, vol. 1, pp. 544-548, Sept. 2002.

[5] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, no. 5, pp.172-180, May 2002.

[6] IEEE Standard 802.11 Working Group, *IEEE 802.11a Physical Layer Specifications*, July 1999.

[7] F. Peng, J. Zhang and W. E. Ryan, "Adaptive modulation and coding for IEEE 802.11n," *IEEE Proc. WCNC 2007*, March 2007.

[8] H. Hu, Y. Zhang and J. Luo, "Distributed antenna systems open architecture for future wireless communications," Auerbach Publications, CRC Press, May 2007.

[9] IEEE Standard 802.16 Working Group, *IEEE standard for local and metropolitan area networks Part 16: Air Interface for Fixed Broadbandwireless Access Systems*, 2002.

[10] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Communications*, vol. 9, no. 5, pp. 76-83, Oct. 2002.

[11] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR: A hign efficiency-high data rate personal communication wireless system," *Proc. of IEEE VTC2000*, pp. 1854-1858, May 2000.

[12] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," *Proc. of Euro. Wireless*, pp. 2244-2248, Sept. 2004.

[13] G. barriac, and J. Holtzman, "Introducing delay sensitivity into the proportional fair algorithm for CDMA downlink scheduling," *Proc. of IEEE 7th Int'l. Symp. Spread Spectrum Techniques and Apps.*, vol. 3, pp. 652-656, Sept. 2002.

[14] H. Zeng et al., "Packet scheduling algorithm considering both the delay constraint and user throughput in HSDPA," *Proc. of Int'l. Conf. Commun., Circuits and Sys.*, vol. 1, pp. 387-92, May 2005

[15] A. Golaup, O. Holland, and A. Aghvami, "A packet scheduling algorithm supporting multimedia traffic over the HSDPA link based on early delay notification," *Proc. 1st Int'l. Conf. Multimeda Services Access Networks*, pp. 78-82, June 2005.

[16] B. Al-Manthari, H. Hassanein and N. Nasser, "Packet scheduling in 3.5G high-speed downlink packet access networks: Breadth and depth," *IEEE Network*, vol. 21, no. 1, pp. 41-46, Jan.-Feb. 2007.

[17] K. W. Choi, D. G. Jeong and W. S. Jeon, "Packet scheduler for mobile communications systems with time-varying capacity region," *IEEE Trans. Wireless Commun.*, vol. 6, no. 3, pp. 1034-1045, March 2007.

[18] L. B. Le, E. Hossain, and A. S. Alfa,"Service differentiation in multirate wireless networks with weighted round-robin scheduling and ARQ-based error control," *IEEE Trans. Commun.*, vol. 54, no. 2, pp. 208-215, Feb. 2006

[19] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantee in adaptive wireless networks," *IEEE J. Selected Areas Commun.*, vol. 23, no. 5, pp. 1056-1066, May 2005.

[20] C. Cicconetti, L. Lenzini, E. Mingozzi, and G. Stea, "An efficient cross layer scheduler for multimedia traffic in wireless local area networks with IEEE 802.11e HCCA," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 11, no. 3, pp. 31-46, July 2007.

[21] Q. Liu, S. Zhou and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, May 2005.

[22] Q. Liu, S. Zhou and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142-1153, May 2005.

[23] H. Chen, H. C. B. Chan and V. C. M. Leung, "Cross-layer enhanced real-time packet scheduling over CDMA networks," *Proc. of IEEE ICON2006*, pp. 1-6, Sept. 2006.

[24] P. Y. Kong, K. C. Chua, and B. Bensaou, "A novel scheduling scheme to share dropping ratio while guaranteeing a delay bound in a multicode-CDMA network," *IEEE/ACM, Trans. Networking*, vol. 11, no. 6, pp. 994-1006, Dec. 2003.

[25] A. T. Anderson and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 719-732, Jun. 1998.

[26] E. Biglieri, G. Caire, and G. Taricco, "Limiting performance of block-fading channels with multiple antennas," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1273-1289, May 2001.

[27] G. L. Stüber, *Principles of Mobile Communication*, 2nd ed. Norwell, MA: Kulwer, 2001.

[28] Y. L. Guan and L. F. Turner, "Generalized FSMC model for radio channels with correlated fading," *IEE Proc. Commun.*, vol. 146, no. 2, pp. 133-137, Apr. 1999.

[29] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley & Sons, New York, 1994.

[30] A. Farrokh and V. Krishnamurthy, "Opportunistic scheduling for streaming users in high-speed downlink packet access (HSDPA)," *Proc. of GLOBECOM2004*, pp. 4043-4047, Nov. 2004.

[31] V. Hassel, G. E. Oien, and D. Gesbert, "Throughput guarantees for wireless networks with opportunistic scheduling," *Proc. of IEEE GLOBECOM*, pp. 1-6, 2006.

[32] J. A. Stankovic, M. Spuri, K. Ramamrithan, and G. C. Buttazzo, *Deadline Scheduling for Real-Time Systems: EDF and Related Algorithms*. Norwell, MA: Kluwer, 1998.

[33] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair queuing in wireless networks: Issues and approaches," *IEEE Pers. Commun.*, vol. 6, no. 1, pp. 44-53, Feb. 1999.

[34] S. Lu, T. Nandagopal, and V. Bharghavan, "Design and analysis of an algorithm for fair service in error-prone wireless channels," *Wirel. Netw.*, vol. 6, no. 4, pp. 323-343, Jul. 2000.

Chapter 6

Conclusions and Future Work

6.1 Summary of Work Accomplished

This thesis has investigated several problems with regards to QoS provisioning in multiple access and packet scheduling schemes for multimedia transmissions over wireless communication systems. A new, multiple access scheme and several scheduling algorithms have been proposed to provide better QoS for multimedia traffic in wireless communication environments. A new approach for cross-layer enhanced QoS provisioning has been proposed and applied in the thesis. Specifically, we have addressed the following four major research problems, with the following contributions to the field:

- **A New MAC Scheme for Wireless Channels with MPR Capability:** In Chapter 2, we have proposed an MRMA scheme for MPR channels. The work uses an MPR channel model that is widely used in the literature. The proposed scheme accounts for the error-prone characteristic of the MPR channel and the specific design guarantees the correct operation of the reservation. MRMA also provides an adequate slot allocation scheme to provide good QoS for multimedia traffic flows. A Markov analysis model has been proposed, along with the MRMA, to evaluate the performance of the MRMA scheme, as well as to validate the simulation mode implemented in this work. The simulation

mode provides detailed performance analysis for MRMA. The results show that MRMA can provide effective channel access and good QoS for multimedia traffic. Compared to another scheme FPLS, the system capacity of voice traffic can be increased by 4% to 33% or 6%to 30% for SNR = 6db and SNR = 10db, respectively, in one of our simulation environment. Also, the data traffic has has only a slight and almost no impact to voice and video traffic, respectively. The MRMA shows details for the design of MAC schemes for wireless channels with MPR capacity and is also an effective MAC scheme with good QoS provisioning for multimedia transmission over wireless channels with MPR capacity.

- **Cross-Layer Enhanced Packet Scheduling for MC-CDMA:** In Chapter 3 of the thesis, we have proposed a cross-layer enhanced packet scheduling scheme for MC-CDMA networks. CEPS makes scheduling decisions by considering the joint FLR, instead of FDR and BER, separately. Essentially, CEPS assumes that the arrival of future data frames is not predictable, and decisions are made based on the historical FLR record, the current buffer status, and the QoS requirements of all different traffic flows. With this information, CEPS optimizes the sum of the weighted FLR of all traffic flows so that the channel throughput is optimized and the QoS is shared with predesigned weights among all traffic flows. We have also implemented a simulation model for CEPS. The results show that CEPS can effectively enhance the channel throughput, compared to other scheduling algorithms, and that the QoS is proportionally provisioned to traffic flows. Essentially, from the simulation result, CEPS enhanced the system capacity by 22% to 39% compared to the other two scheduling algorithms. CEPS shows the effectiveness of the cross-layer QoS consideration and is a practical candidate scheduling algorithm. It

provides flexible and effective QoS provisioning for multimedia transmission, so that it may be used in different wireless communication systems to enhance the users' experiences with multimedia applications.

- **Cross-Layer Optimization of the Maximum Number of Simultaneous Transmissions in MC-CDMA**: In Chapter 4 of the thesis, we have proposed a cross-layer optimization for MC-CDMA. By dynamically adapting the maximum number of transmissions for one slot, the cross-layer optimization minimizes the FLR experienced by the whole system. During optimization, we assume that a slot allocation scheme is used to minimize the FLR, given the scheduling decision and the maximum number of simultaneous transmissions for each time slot. Also, we assume that a scheduling algorithm is used for mapping the buffer status and the traffic status into the scheduling decision every frame. These assumptions are valid and other slot allocation and scheduling algorithms can also be used for the optimization. According to the traffic status and the buffer status of all traffic flows, the optimization determines the maximum number of simultaneous transmissions so that the FLR of the whole system is optimized. We have also proposed an approximation scheme to address the computational complexity issue for the optimization. Simulation results have been presented to show the large effects of the optimization. Compared to CEPS, the optimization can achieve a 10% capacity improvement for the system. The optimization verifies the effectiveness of the cross-layer QoS consideration and could be used to optimize channel capacity and QoS provisioning for MC-CDMA systems.

- **QoS-Based Cross-Layer Scheduling for AMC-Based Systems:** In Chapter 5 of the

thesis, we have proposed a cross-layer scheduling for AMC-based systems. The QoS-CLS method optimizes the QoS of different traffic flows by selecting appropriate modulation and coding pairs and the scheduling decisions. The delay requirement for multimedia traffic is met by a buffering mechanism. The FLR requirements of multimedia traffic are satisfied by the optimization, and for traffic flows not requiring the FLR guarantee, the FLR experienced is proportionally distributed according to a predesigned weight for the traffic type. The optimization is based on the information of the traffic status, the buffer status, and the QoS requirements of all traffic flows. Simulation result shows that the system capacity is increased by 38% and event more compared by the EDF and OS algorithms. To address the computational complexity of the optimization, two approximation methods are proposed. QoS-CLS provides a good solution for the scheduling problem in AMC-based systems for multimedia transmissions. It can achieve precise QoS provisioning and, at the same time, optimize the channel capacity. It could be a candidate scheduling algorithm for future wireless communication systems with AMC. Moreover, the scheduling also verified the effectiveness of the cross-layer QoS consideration in AMC-based wireless communication systems.

6.2 Future Work

The work described in this thesis can be the subject of future research, in several aspects, as described below:

- **Combination of the MRMA and CEPS**: We have proposed a multiple access scheme MRMA for MPR channel in Chapter 2. Although a simple, round-robin slot allocation

mechanism is used in MRMA, it is also possible to apply other scheduling algorithms to further provide the QoS differentiation. Studies of the performance of the MRMA with an advanced scheduling algorithm would be of interest. We have also proposed a CEPS scheme in Chapter 3 for MC-CDMA systems. When CEPS is applied for the uplink scheduling, a signaling channel is required for user terminals to notify the base station and for the base station to inform the scheduling decision and the transmission result. One of our future studies is to apply CEPS on MRMA, and to investigate how the signaling function could be taken over by the MRMA protocol and how better QoS differentiation among different traffic flows could be provisioned.

- **Admission Control for Multimedia Traffic with Cross-Layer QoS Consideration**: We have applied the cross-layer QoS consideration in Chapters 3, 4, and 5 to optimize the QoS provisioning and system throughput. In addition, the result can be used to study the performance improvement in the admission control by the cross-layer QoS consideration. The admission control in MPR systems for multimedia traffic is another interesting topic. The cross-layer optimizations proposed in Chapters 3, 4, and 5 can further improve the channel capacity. Still, the design of an effective and efficient admission control scheme becomes more complicated to allow the admitted traffic flows to meet their QoS requirement while channel utilization is maximized. In our future work, we intend to investigate the design of an admission control scheme with cross-layer QoS considerations and optimizations.

- **Cross-Layer QoS Consideration in Other New Wireless Technologies**: Although the cross-layer QoS consideration can be generally applied to any wireless link, some

advanced wireless technologies, such as MIMO, multi-carrier CDMA, and OFDMA, are sophisticated and not easily modeled as MC-CDMA or AMC channels. Thus, to apply the cross-layer QoS consideration in such wireless channels for achieving better performance by adjusting the physical and MAC layer schemes would be a challenging undertaking. In future work, we intend to implement the cross-layer schemes for these sophisticated channels to improve system performance.

- **Better Solutions for Proposed Optimizations**: Except for Chapter 3, all of our work on the cross-layer QoS consideration is based on the optimization modeled by MDP. We have solved these MDP optimizations with the linear programming method. When the state space of the optimization becomes very large, however, the computational complexity becomes a practical problem. We have proposed several approximation methods to reduce the state space of the optimization in Chapter 5, but still need better ways for solving the optimization so that it can be easily implemented in large systems. Reinforcement learning could be a potential solution for the MDP. In future work, we intend to find a better way to model the optimization problem for the cross-layer QoS consideration and find better ways to solve MDP-based optimizations.

- **Higher Layer QoS**: In proposing the cross-layer QoS optimization, we have used a number of assumptions. Some of these stipulated the specific design of higher layers. For example, we assumed that every data frame has an expiration deadline on the wireless link. This assumption dictates the higher layer design, which may affect the higher layer QoS. It would be of interest to see how the cross-layer QoS consideration on the wireless link can affect the end-to-end QoS of the wireless communication system. In the future,

we will investigate the relationship between the wireless link QoS and the end-to-end QoS of the wireless communication system, and investigate the impact of the cross-layer QoS consideration on the higher layer QoS.

Bibliography

[1] H. Chen, F. Yu, H. C. B. Chan and V. C. M. Leung, "A novel multiple access scheme over multi-packet reception channels for wireless multimedia networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1501-1511, April 2007.

[2] H. Chen, H. C. B. Chan, V. C. M. Leung and J. Zhang, "Cross-layer enhanced uplink packet scheduling for multimedia traffic over MC-CDMA networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 986-992, Feb. 2010.

[3] H. Chen, H. C. B. Chan, and V. C. M. Leung, "Two cross-layer optimization methods for transporting multimedia traffic over multicode CDMA networks," *Proc. of IEEE WCNC'07*, pp. 288-293, March 2007.

[4] H. Chen, H. C. B. Chan, and V. C. M. Leung, "A cross-layer scheduling scheme for wireless multimedia transmissions with adaptive modulation and coding," *Proc. of IEEE Broadnets2008*, pp. 249-256, Sep. 2008.

Appendix A

Calculations for State Transition Probabilities for MRMA

A.1 Calculation of The Transition Probability T_1

Section 2.4 discusses the state transitions at the frame boundary between frame n and $n+1$. We have:

$$Pst = 1 - e^{\alpha T_f}, \qquad Pts = 1 - e^{\beta T_f} \tag{A.1}$$

Suppose that there are i CON UTs and j IDL UTs changing their states; we have:

$$j = N_{CON,n+1} - N_{CON,n,h_n} + i \tag{A.2}$$

Therefore, the transition probability can be found as (A.3)

$$T_1(N_{CON,n+1}, N_{REV,n+1} \mid N_{CON,n,h_n}, N_{REV,n,h_n}) =$$

$$\begin{cases} \displaystyle\sum_{i=\max\{0,N_{CON,n,h_n}-N_{CON,n+1}\}}^{\min\{N_{CON,n,h_n},Nv-N_{CON,n+1}-N_{REV,n,h_n}\}} B(N_{CON,n,h_n}, i, Pts) \times B(N_{IDL,n,h_n}, j, Pst) & \\ & N_{REV,n+1} = N_{REV,n,h_n} \\ 0 & \\ & N_{REV,n+1} \neq N_{REV,n,h_n} \end{cases} \tag{A.3}$$

where

$$B(x, y, z) = \binom{x}{y} z^y (1-z)^{x-y}$$

$$N_{IDL,n,h_n} = Nv - N_{CON,n,h_n} - N_{REV,n,h_n}$$

Note that the number of REV UTs remains unchanged at $N_{REV,n+1} = N_{REV,n,h_n}$ during this transition. Also, the variables i and j in (A.3) must satisfy the following conditions:

$$0 \leq i \leq N_{CON,n,h_n}, \quad 0 \leq j \leq N_{IDL,n,h_n} \tag{A.4}$$

Based on these conditions, the upper and lower limits of i can be determined as shown in (A.3).

A.2 Calculation of The Transition Probability T_2

Here we first analyze the transition from REV to IDL. The number of REV UTs changing to IDL at the end of frame $n+1$ is $N_{CON,n+1} + N_{REV,n+1} - N_{CON,n+1,h_{n+1}} - N_{REV,n+1,h_{n+1}}$. Given that there are $N_{REV,n+1}$ REV UTs at the beginning of frame $n+1$, the probability that there are $N_{CON,n+1} + N_{REV,n+1} - N_{CON,n+1,h_{n+1}} - N_{REV,n+1,h_{n+1}}$ REV UTs changing to IDL in frame $n+1$ is:

$$B(N_{REV,n+1}, N_{CON,n+1} + N_{REV,n+1} - N_{CON,n+1,h_{n+1}} - N_{REV,n+1,h_{n+1}}, Pts) \tag{A.5}$$

Next, we analyze the transition from CON to REV by a recursive method. Such a transition depends on the number of mini-slots (i.e., h_{n+1}) in the frame, the permission probability p, the initial number of CON uses at the beginning of the frame (i.e., $N_{CON,n+1}$), and the C matrix. In frame $n+1$, if there are y CON UTs at the end of the i-th mini-slot, the probability that

there are x *CON* UTs at the end of the $(i+1)$-th mini-slot is:

$$\begin{aligned} &P(N_{CON,n+1,i+1} = x \mid N_{CON,n+1,i} = y) \\ &= \sum_{z=y-x}^{y} B(y,z,p) \times C_{z,y-x} \end{aligned} \tag{A.6}$$

Here, z is the number of *CON* UTs who send reservation requests via the $(i+1)$-th mini-slot.

By considering all the possible values of y, we have:

$$\begin{aligned} &P(N_{CON,n+1,i+1} = x) \\ &= \sum_{y=x}^{N_{CON,n+1}} P(N_{CON,n+1,i} = y) \sum_{z=y-x}^{y} B(y,z,p) \times C_{z,y-x} \end{aligned} \tag{A.7}$$

Note that the initial condition is:

$$P(N_{CON,n+1,0} = y) = \begin{cases} 1, & y = N_{CON_n+1} \\ 0, & \text{otherwise} \end{cases} \tag{A.8}$$

Furthermore, we have:

$$h_{n+1} = t \times \left(F - \lceil \frac{N_{REV,n+1}}{K_{vo}} \rceil \right)^{+} \tag{A.9}$$

Denote $P(N_{CON,n+1,h_{n+1}} \mid N_{CON,n+1}, N_{REV,n+1})$ as the probability that, if there are $N_{CON,n+1}$ *CON* UTs and $N_{REV,n+1}$ *REV* UTs at the beginning of frame $n+1$, there will be $N_{CON,n+1,h_{n+1}}$ *CON* UTs at the end of frame $n+1$. This probability can be found by running the recursive equation (A.7) from $i = 0$ to $i = h_{n+1}$ based on the initial condition defined by (A.8).

Finally, we can get the transition probability as follows:

$$\begin{aligned} &T_2(N_{CON,n+1,h_{n+1}}, N_{REV,n+1,h_{n+1}} \,|\, N_{CON,n+1}, N_{REV,n+1}) \\ =&B(N_{REV,n+1}, N_{CON,n+1} + N_{REV,n+1} - N_{CON,n+1,h_{n+1}} \\ &- N_{REV,n+1,h_{n+1}}, Pts) \times \\ &P(N_{CON,n+1,h_{n+1}} \,|\, N_{CON,n+1}, N_{REV,n+1}) \end{aligned} \tag{A.10}$$

Appendix B

List of Publications

Journal Papers

- H. Chen, F. Yu, H. C. B. Chan and V. C. M. Leung, "A novel multiple access scheme over multi-packet reception channels for wireless multimedia networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1501-1511, April 2007.

- H. Chen, H. C. B. Chan, V. C. M. Leung and J. Zhang, "Cross-layer enhanced uplink packet scheduling for multimedia traffic over MC-CDMA networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 986-992, Feb. 2010.

- H. Chen, H. C. B. Chan, and V. C. M. Leung, "Cross-layer optimization for multimedia transport over multicode CDMA networks," submitted to *IEEE Transactions on Mobile Computing*, Dec. 2009.

- H. Chen, H. C. B. Chan, and V. C. M. Leung, "A Cross-layer Scheduling Scheme for Wireless Multimedia Transmissions with Adaptive Modulation and Coding," in preparation for journal submission.

Conference Papers

- H. Chen, F. Yu, H. C. B. Chan and V. C. M. Leung, "A novel multiple access scheme in

wireless multimedia networks with multi-packet reception," in *Proc. of ACM WMuNeP 2005*, pp. 24-31, 2005.

- H. Chen, H. C. B. Chan and V. C. M. Leung, "Cross-layer enhanced real-time packet scheduling over CDMA networks," *Proc. of IEEE ICON2006*, pp. 1-6, Sept. 2006.

- H. Chen, H. C. B. Chan, and V. C. M. Leung, "Two cross-layer optimization methods for transporting multimedia traffic over multicode CDMA networks," *Proc. of IEEE WCNC'07*, pp. 288-293, March 2007.

- H. Chen, H. C. B. Chan, and V. C. M. Leung, "A cross-layer scheduling scheme for wireless multimedia transmissions with adaptive modulation and coding," *Proc. of IEEE Broadnets2008*, pp. 249-256, Sep. 2008.