

國立交通大學

電機工程學系

博士論文

無線網路之媒體接取控制與排程機制

Medium Access Control and Scheduling  
Schemes for Wireless Networks

研究生：顏志明

指導教授：張仲儒 博士

中華民國 九十九 年 一 月

# 無線網路之媒體接取控制與排程機制

學生：顏志明 指導教授：張仲儒

國立交通大學電機工程學系

## 摘要

在無線網路中保障服務品質是很重要的議題。現今，不同的網路對不同的服務最佳化，在每個網路中提供多樣性的服務變的越來越急迫了。為了解決這個問題，網路中的媒體控制協定與排程機制需要再重新設計，來滿足多媒體服務的品質並提升系統效能。

在無線區域網路中，媒體接取控制機制的目的是解決使用者間相互競爭的問題。它可利用不同的仲裁訊框間隔 (Arbitration Inter Frame Space) 來區分不同的服務以提升服務品質，但是效能並不是很好。為了要在無線區域網路保證多媒體服務的品質，提出一個適應性p-persistent (APP) 基礎的媒體接取控制法。APP媒體接取控制法依照不同的服務種類給予不同的允諾機率來服務多媒體使用者並利用傳送的允諾機率來區分使用者。當使用者有較大的延遲封包他會有較高的允諾機率。從分析與模擬結果中証實，APP可以降低使用者間的延遲變異達15%，並降低高優先權服務的失敗率。

在IEEE 802.16 都會型網路中，基地台要服務大量的使用者與提供不同種類的服務型態，所以需要排程機制提升系統效能，並兼顧使用者的服務品質需求。我們在IEEE 802.16 上鏈路提出了一個基於動態優先次序的資源分配(dynamic priority-based resource allocation, DPRA)機制，在下鏈路提出一個最大化系統產出並

降低複雜度的功效式排程機制 (utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling)。

DPRA 機制對於急迫性較高的服務，我們給予較高的優先次序值(Priority value)，使具有較高優先次序的使用者能優先被分配系統資源做傳輸。我們也會根據每一種服務在不同時間的急迫性，動態調整其優先次序。我們提出的DPRA機制會在子通道(subchannel), 調變方法(modulation order), 以及能量(power)三方面找尋最佳化的資源分配方法，並且對同一個使用者做一致性分配(consistent allocation)。由模擬結果顯示，我們提出的方法可以達到傳輸速率最佳化以及QoS的滿足，並且能減少標頭傳輸(transmission overhead)以及降低運算複雜度。

U\_TMCR程機制不只在保證服務品質 (QoS) 的情形下最大化系統效能，同時也降低計算複雜度，並針對多媒體使用者做通道配置、天線選擇與決定調變方法。U\_TMCR 機制根據通道品質和使用者的服務品質需求，為每個使用者設計效用函數 (utility function)，將排程問題轉成考慮系統限制對整個系統最佳化效用的問題。U\_TMCR 機制也提出一個低複雜度演算法來解決所提出的最佳化問題。由模擬驗證，U\_TMCR可以提升8%的系統效能並降低6.25%~29.2%的計算複雜度。

# Medium Access Control and Scheduling Schemes for Wireless Networks

Student: Chih-Ming Yen    Advisor: Chung-Ju Chang

Institute of Electrical Engineering  
National Chiao Tung University

## Abstract

To guarantee the quality of service (QoS) in wireless network is an important issue. Currently, the different networks are optimized for different services, but it becomes urgent to provide varied service in wireless networks. To address this problem, there need to re-design a medium access control (MAC) or scheduling scheme to satisfy the QoS of multimedia service and to enhance the network utilization.

In the WLAN, the goal of the medium access control (MAC) protocol is to deal with the contention of stations. It uses the different arbitration inter frame space to differentiate the services to promote the service quality, but the QoS satisfaction is not good enough. In order to support multimedia services in the WLAN, an *adaptive p-persistent-based* (APP) MAC scheme for IEEE 802.11 WLAN is investigated. The APP MAC scheme can further differentiate priorities of access categories by the initial permission probabilities and adaptively adjust permission probabilities to transmission stations according to its transmission state. Numerical and simulation results show that the APP MAC scheme can reduce the dropping probability of high priority service and effectively reduce the delay variance by 15%.

The base station has to serve the massive users with different service type in IEEE

802.16 WiMAX system. Therefore, it needs to elaborately design the scheduling scheme to enhance QoS satisfaction with high system efficiency. In the IEEE 802.16 system, a dynamic priority resource allocation (DPRA) scheme for uplink and a utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling scheme for downlink are investigated.

The DPRA scheme dynamically gives priority values to different services based on the urgency degrees and allocates system radio resources according to the priority values. It can maximize the system throughput and satisfy differentiated QoS requirements. Also, the DPRA scheme performs consistent allocation to conform the uplink frame structure of IEEE 802.16, to fulfill QoS requirement, and to reduce the computational complexity. Simulation results show that the proposed DPRA scheme performs very close to the optimal method, which is by exhaustive search, in system throughput; and it outperforms the conventional EFS algorithm [39] in the performance measures such as system throughput, rtPS packet dropping rate, ratio of unsatisfied nrtPS, and average transmission rate of BE.

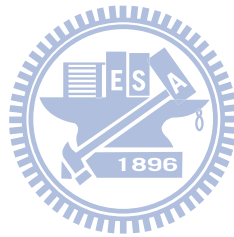
The goals of the U\_TMCR scheme are not only to maximize system throughput under QoS guarantee but also to reduce computational complexity. Based on channel quality and QoS requirements of each user, the U\_TMCR scheme designs a utility function for every user and formulates the scheduling into an optimization problem of overall system utility function subject to system constraints. It also contains a heuristic TMCR algorithm to *efficiently* solve the optimization problem. Simulation results show that the U\_TMCR scheme can improve the system throughput by 8% and reduce the computational complexity by 6.25%~29.2%.

# Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Chung-Ju Chang, for his great care and guidance at research and the methodology of schedule throughout my school time.

I am also indebted to all friends in BCN Lab. Thank you all for your kind help.

This dissertation is dedicated to my family for their understanding and encouragement.



# Contents

Chinese Abstract.....	i
Abstract .....	iii
Acknowledgments.....	v
Contents.....	vi
List of Figures .....	viii
List of Tables.....	ix
Chapter 1 .....	1
1.1 Motivation .....	1
1.2 Literature Survey.....	3
1.3 Dissertation Organization.....	7
Chapter 2 .....	9
2.2 System Models.....	13
2.3 The Adaptive P-Persistent (APP) MAC Scheme .....	14
2.4 Analysis .....	16
2.4.1 System Throughput .....	20
2.4.2 Delay .....	21
2.4.3 The Optimal Value of $P_0$ .....	22
2.5 Numerical and Simulation Results.....	22
2.5.1 Data Only Environment .....	22
2.5.2 Multimedia Service Environment .....	28
2.6 Concluding Remarks .....	34
Chapter 3 .....	36
3.1 Introduction .....	36
3.3 Dynamic Priority-based Resource Allocation Scheme .....	45
3.3.1 Problem Formulation.....	45
3.3.2 DPRA Scheme.....	47
3.4 Simulation Results.....	54
3.4.1 Simulation Environment .....	54



3.4.2 Source Model and QoS Requirements .....	54
3.4.3 Performance Evaluation .....	56
3.5 Concluding Remarks .....	63
Chapter 4 .....	64
4.1 Introduction .....	64
4.2 System Model.....	68
4.2.1 System Assumptions .....	70
4.3 Utility-based TMCR Scheduling Scheme .....	73
4.3.1 Utility Function .....	73
4.3.2 Problem Formulation.....	75
4.3.3 Heuristic TMCR Algorithm .....	76
4.4 Simulation Results.....	80
4.5 Concluding Remarks .....	89
Chapter 5 .....	91
Bibliography.....	96
Vita .....	103



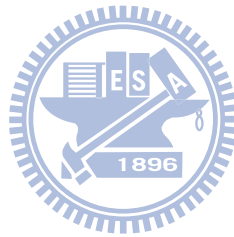


# List of Figures

Figure 2.1 State transition diagrams for the APP MAC scheme .....	18
Figure 2.2 Collision probabilities of APP, BEB, and DDFC.....	24
Figure 2.3 System throughputs of APP, BEB, and DDFC.....	25
Figure 2.4 Mean delays of APP, BEB, and DDFC .....	26
Figure 2.5 Delay variances of APP, BEB, and DDFC.....	27
Figure 2.6 performance of APP with optimal $P_0^*$ and BEB with $W_{opt}$ .....	28
Figure 2.8 System throughput.....	32
Figure 2.9 (a) Mean delay and (b) delay variance of low priority packet.....	33
Figure 3.1 Flow chart of the DPRA scheme.....	53
Figure 3.2 System Throughput.....	58
Figure 3.3 Voice Packet Dropping Rate.....	59
Figure 3.4 Video Packet Dropping Rate.....	59
Figure 3.5 Ratio of Unsatisfied HTTP Users.....	60
Figure 3.6 FTP Average Transmission Rate.....	61
Figure 3.7 Average Number of Iterations.....	62
Figure 4.1. System configuration of the downlink MIMO-OFDMA system.....	69
Figure 4.2. Flow chart of the heuristic TMCR algorithm .....	77
Figure 4.3 System Throughput.....	84
Figure 4.4 (a) Voice packet dropping rate; (b) Video packet dropping rate .....	85
Figure 4.5 (a) Mean delay of voice packet; (b) Mean delay of video packet.....	86
Figure 4.6 Guaranteed ratio of HTTP packets.....	87
Figure 4.7 <i>BE</i> throughput.....	88

# List of Tables

Table 2.1 Parameter Settings for a WLAN Environment.....	23
Table 3.1 System-Level Parameters .....	54
Table 3.2 QoS Requirements of each service type.....	55
Table 4.1 System-Level Parameters .....	81
Table 4.2 QoS requirements of each service .....	81



# Chapter 1

## Introduction

---

### 1.1 Motivation

The success of second-generation (2G) mobile system in the pass decades gives rise to the development of third-generation (3G) mobile system. For example, global system for mobile communication (GSM ) and IS-95 from 2G system designed to carry speech and low-bit-rate data has been upgraded to provide higher data-rate services on their 3G version. Besides, the range of wireless system, such as general packet radio service (GPRS), IMT-2000, Bluetooth, and IEEE 802.11 wireless local area networks (WLANs) have as well developed from 2G to 3G. All those system were separately designed to meet the needs of a variety of service types, data rates, and users. Having just one system can never be sufficient enough to be substituted for all other technologies since every single system has its own merits and relative drawbacks.

Mobile cellular networks have progressed to 3G within two decades due to the advancement of wireless technologies and the emergence of multimedia data services. People have pictured the prospects of wireless networks as the integration of lots of existing and newly developed wireless access networks such as WLAN, IEEE 802.16

wireless metropolitan area networks (WMANs), GPRS, and universal mobile telecommunications system (UMTS). The tendency towards packet-switched technologies, and the increasing use and acceptance of the internet protocol (IP) illustrates that an IP core network is going to connect different wireless access networks together. Therefore, based on that conception, several overlapping IP-based wireless access network domains will be the future of all-IP wireless networks.

The future wireless networks have some evident features listed as the following. Firstly, all-IP based hybrid networks which allow users to use any system at anytime and anywhere are the future of wireless networks. Multiple wireless networks provide users carrying an integrated terminal with a wide range of applications. Secondly, not only telecommunications services but also data and multimedia services are provided by the future wireless systems. High-data-rate services with good system reliability will be provided in order to support multimedia services, and, simultaneously, the cost of a low per-bit transmission will stay maintained. Thirdly, the future network will provide personalized service. It is anticipated that users in widely different locations, occupations, and economic classes will use the services while the future wireless services are launched. For the reason of catering the needs of diverse users, personal and customized services should be designed. Finally, facilities for integrated services will also be provided by the future wireless systems. Users can use multiple services at a time from any service provider.

There are two types of existing wireless systems: IP-based and non-IP-based. IP-based systems are usually optimized for data services (e.g., IEEE 802.11 WLAN). As for non-IP-based systems, many of which are highly optimized for voice delivery (e.g., GSM, cdma2000, and UMTS). In order to integrate the multimedia service, one major challenge in the future wireless systems is radio resource management (RRM). When

particularly considering time-sensitive or multimedia applications, the quality of service (QoS) guarantee for end-to-end services needs to be stated.

## 1.2 Literature Survey

In personal area network (PAN), high transmission rate and low design complexity in medium access control (MAC) protocol are the benefits that wireless local area networks (WLAN) possess. Hot spot cells and indoor environments widely apply it for diverse applications. Due to channel sharing, the MAC protocol is the key role to determine the efficiency and performance of the WLAN. The Abramson proposed an elegant MAC protocol, called ALOHA [1]. In ALOHA, high collision probability result form stations allowed transmitting immediately upon receiving data from upper layers. To decrease the collision probability, carrier sense multiple access (CSMA) scheme [2] requires stations to transmit until the medium becomes idle. When a station detects the channel is idle, it can transmit with a probability of 1 or  $p$  ( $0 < p < 1$ ). The former is called 1-persistent CSMA and the latter is  $p$ -persistent CSMA. In the IEEE 802.11 [3], it adapt the CA scheme of DCF further reduces frame collision probability by requiring each backlogged station to perform binary exponential backoff (BEB) after the medium becomes idle. In BEB, if a station successfully transmits a frame, its contention window will be reset to an initial value. However, if the transmission fails, the window size is doubled. Nevertheless, owing to the situation that collided stations would have smaller probability to access the medium than new stations, it generally entails larger delay variance on stations and results in unfairness.

The important and challenging issue is to divide the channel fairly among stations. Therefore, the design of efficient MAC protocols with high-throughput performance, as well as a high degree of fairness performance, has been a major focus in WLAN

research areas [4], [5], [49]-[51]. There have been many studies of backoff algorithms [6]-[13], [16], [45], but all of them did not address the problem of larger delay variance.

For real-time applications, higher delay variance leads to larger amount of dropped packets due to excess delay. For non-real-time data applications, on the other hand, higher delay variance usually causes larger requirement of storage buffer or more probability of buffer overflow. Hence, in order to reduce the delay variance, the radio resource over WLAN interface is necessarily to be fairly shared by an effective MAC protocol. The fairness problem of MAC in WLAN that some algorithms solve was proposed [14], [15]. However, channel throughput is decreased because of the high collision probabilities.

Dynamic contention window (*CW*) schemes [16]-[18], different maximum packet length scheme [18], and various interframe space (IFS) schemes [18]-[20] are usually adopted to design the priority differentiation in order to support multimedia services for the IEEE 802.11e [52] WLAN. However, owing to the backoff scheme, large delay variance in the same access category (AC) would still be arisen by these solutions. Obviously, larger probability of quality-of-service (QoS) violation of multimedia traffic is brought about by higher delay variance because of excess delay.

In WMANs, the problem of future wireless communication is resolved by multiple-input multiple-output (MIMO) based orthogonal frequency division multiplexing (OFDM) because it helps achieve high system capacity and provide transmit/receiver diversity for reliable communication link. Downlink resource management for multiuser OFDM (MU-OFDM) systems has recently been investigated [21]-[25], in which topics were emphasized on transmission power allocation, subcarrier allocation, bit allocation, or adaptive modulation and coding (AMC). The goal of the design is to maximize system capacity, minimize total transmission power, provide

fairness, or guarantee QoS requirements.

Many papers investigated the downlink resource allocation [21]-[32] but few papers probed into the uplink resource allocation. Both downlink and uplink perform the resource allocation primarily through the base station (BS). The power distribution over the selected set of subcarriers for every user is included in the algorithm in [35] so that it minimizes the total power being used. In [36], a greedy subcarrier allocation algorithm, based on a marginal rate function, and an iterative water-filling power allocation algorithm were proposed. A practical algorithm and the optimization problem were presented in [37]. All of them can nearly reach an optimal solution but they did not focus more on the QoS requirement. The power saving in IEEE 802.16 OFDMA systems via an efficient uplink resource allocation was shown in [38]. While guaranteeing BER, it minimizes the required transmissions power through adaptively adjusting the modulation and coding scheme. However, they don't attend to their differentiated QoS requirements and multiple services. [39] exhibited an efficient and fair scheduling (EFS) algorithm for each time slot in IEEE 802.16 OFDMA/TDD system. A fixed priority scheme which gives priorities to service traffic according to their QoS requirements is applied to design the EFS algorithm.

The bandwidth is allocated according to channel quality and queue state of the traffic for SS with real-time and non-real-time polling services in [40]. From the previously mentioned works, people either omitted or simplified the QoS requirements and fairness issues. The QoS requirement usually refers to a predefined weight which corresponds to the fixed priority scheme or a minimum required transmission data rate. However, the design of radio resource allocation for practical applications should include the delay bound and the packet dropping rate, regarded as essential QoS requirements to provide multimedia real-time traffic. In addition, buffer conditions of

different traffic types and realistic traffic models should be taken into account.

For MIMO-OFDM systems, the exponential increases in the computational complexity on radio resource scheduling for downlink multiuser is proportional to the number of subcarrier, multiuser, transmitting antenna, and receiving antenna. The multiuser scheduling algorithms for system throughput maximization with reduced complexity in a downlink MIMO/OFDMA system were proposed [26]-[29]. They decoupled the multiuser scheduling problem into frequency and spatial domains. Multiple parallel independent single-user MIMO channels are decomposed from the multiuser downlink MIMO channel by the preprocessing scheme. However, the number of transmitting antennas restrains the number of simultaneously transmitted users. Computational complexity of the scheduling algorithm is still too high and the QoS requirements and user demand were not considered in the scheduling algorithm.

A fixed priority algorithm was proposed [30] in relation to the QoS requirement in MU-MIMO-OFDMA system. The non real-time traffic's rate of transmission is too low to fulfill the requirement rate while the real-time traffic can be provided in time at low traffic intensity. A dynamic priority scheduling scheme was proposed in [32]. With that scheme, not only is high priority given for urgent users but also the priority of users is dynamically adjusted frame by frame. However, the ARRA does not give the clear differentiation of real-time service from non-real time one but depends on the time to expiration while adjusting the priority.

This is also an important issue: the tradeoff between system performance and computational complexity. Optimal solution [41] can be achieved by the greedy algorithm which performs symbol by symbol allocation, but it causes high computational complexity. The symbol-by-symbol allocation algorithm costs high transmission overhead according to the frames structure of DL-MAP and UL-MAP



defined in IEEE 802.16 for downlink and uplink, respectively. Moreover, most resource allocation algorithms are not only designed for downlink but also claimed to be compatible with uplink. Even so, both the uplink frame structure (UL-MAP) and downlink frame structure (DL-MAP) have different definitions in IEEE 802.16 specifications [42]. Therefore, to meet its individual frame structures, a design of an efficient and feasible resource allocation algorithm for either downlink or uplink is particularly needed.

### 1.3 Dissertation Organization

In this dissertation, the radio resource allocation schemes in wireless network are studied, and the Qos guarantee radio resource management schemes are investigated in both PAN and WMAN.

For IEEE 802.11 WLAN to provide low delay variance, an adaptive p-persistent-based (APP) medium access control (MAC) scheme [46]-[48] is presented in Chapter 2. Permission probabilities of transmission for stations being incurred with different packet delays can be differentiated through the APP MAC scheme and it is designed as a function of the numbers of retransmissions and re-backoffs so that stations with larger packet delay can have higher permission probability. Moreover, the scheme is modeled by a Markov-chain and successfully analyzed, in which the system throughput and delay are derived from. For multimedia services, the APP MAC scheme adaptively gives transmission stations which are in different access category and with various waiting delay differentiated permission probabilities.

For uplinks in IEEE 802.16 wireless communication systems, a dynamic priority resource allocation (DPRA) scheme [33]-[34] is proposed in Chapter 4. The DPRA scheme dynamically gives four types of service traffic based on their urgency degrees

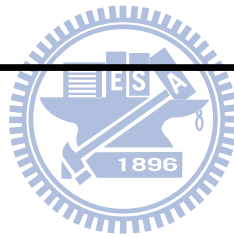
priority values and allocates system radio resources according to their priority values. It can satisfy differentiated QoS requirements and maximize the system throughput. Also, in order for packets of users to conform the uplink frame structure of IEEE 802.16 to fulfill QoS requirement and reduce the computational complexity, the DPRA scheme performs consistent allocation.

For downlink multiuser MIMO-OFDMA systems, a utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling scheme [43]-[44] is proposed in Chapter 3. The U\_TMCR scheme allocates subchannels, antenna sequence, and modulation order to multimedia users with goals not only to reduce computational complexity but also to maximize system throughput under QoS guarantee. Based on each user's channel quality and QoS requirements, both a utility function for every user is designed and the scheduling is formulated into an optimization problem of overall system utility function subject to system constraints by the U\_TMCR scheme. It also contains a heuristic TMCR algorithm for efficiently solving the optimization problem.

Finally, the conclusive statements and future research topics are addressed in Chapter 5.

# Chapter 2

## Analysis of an Adaptive $P$ -Persistent MAC Scheme for WLAN Providing Delay Fairness



### 2.1 Introduction

In the IEEE 802.11, the fundamental mechanism to access the medium is called distributed coordination function (DCF), which is based on carrier sense multiple access with collision avoidance (CSMA/CA) protocol. Retransmissions of collided packets are managed by *binary exponential backoff* (BEB) rules. The IEEE standard also defines an optional point coordination function (PCF), which is a centralized MAC protocol to support collision free and time bounded services. Both the DCF and PCF can operate concurrently with the same basic service set (BSS) to alternate contention and contention-free periods.

In the DCF mode, if a station has a new packet to transmit, it will sense the channel

state firstly. The station transmits only if the channel is idle for a period of time equal to a DCF inter frame space (DIFS). Otherwise, the station persists to monitor the channel until the measured idle period equals a DIFS. Additionally, the DCF also adopts a BEB scheme to avoid the occurrence of packet collision.

Traditional MAC scheme accompanying with the BEB algorithm is one of the most widely used scheme for data transmission, because of its simplicity and high channel utilization. However, the fairness of the BEB algorithm is very poor in some cases. For example: considering a WLAN with  $n$  stations using the DCF mode to access channel, and the stations always have packets to transmit. When one station transmits successfully, it will decrease the size of the contention window to the size of the initial contention window ( $W_0$ ). Before it transmits a next packet, it has to uniformly choose a backoff counter in the backoff interval,  $(0, W_0-1)$ . At that instance, other stations which had experienced collided transmission have a larger backoff interval. As a result, a station with new packet in queue has higher probability to access channel than other waiting stations; that is, unfairness occurs.

Some algorithms to solve the fairness problem of MAC in WLAN were proposed [14], [15]. A multiplicative increase linear decrease (MILD) scheme was proposed in MACAW protocol for WLAN [14]. In the MILD scheme, the contention window of a collided station was increased by multiplying an amount of 1.5, while the contention window of a successful station is decreased by one step. Here, the step was defined as the transmission time of a packet. In the MACAW protocol, the current backoff interval information was included in each transmitted packet, and also a *backoff interval copy mechanism* implemented in each station copied the backoff intervals of the overheard successful transmitters.

Although MILD scheme addresses the problem of unfairness in the BEB algorithm,

it incurs a new problem. Consider the same example as above. When a station successfully transmits a packet with large contention window, other stations waiting to transmit packets change their backoff interval to the large contention window because of the backoff interval copy mechanism applied in MILD scheme. This algorithm works well when many stations happen to transmit at the same time because the probability of collision decreases. But, it results in long channel idle time and decreases channel utilization if only few stations contend for the wireless channel.

Haas and Deng proposed a new MAC scheme [45] named sensing backoff algorithm (SBA). The SBA is an optimized version of MILD algorithm in slotted ALOHA networks. In the SBA, upon collision, stations multiplied their contention windows by  $\alpha$  ( $\alpha > 1$ ). The backoff intervals of the transmitting station and the receiving station, after each successful transmission, were multiplied by  $\theta$  ( $\theta < 1$ ). The contention windows of all active stations sensing a successful transmission were decreased by  $\beta$  steps. Once the value of  $\alpha$  was chosen, the optimization parameters for SBA can be accordingly determined. The SBA guarantees that the successful transmission probabilities of other stations are the same as that of the previously successful station; that is, the stations have the same transmission probability regardless of the number of retransmission. Unfortunately, SBA does not resolve the problem of large delay variance among stations.

Yamada, Morikawa, and Aoyama proposed a decentralized delay fluctuation control (DDFC) MAC mechanism [15], where the contention window is changed according the packet waiting time. The larger the packet waiting time is, the smaller the contention window will be. The DDFC in nature lessens variance of waiting time from enqueueing to successful transmission. Unfortunately, the channel utilization in DDFC is still low due to the small contention windows and high collision probabilities.

To support multimedia services for the IEEE 802.11e WLAN, dynamic contention window ( $CW$ ) schemes [16]-[18], different maximum packet length scheme [18], and various interframe space (IFS) schemes [18]-[20] are usually adopted to design the priority differentiation. However, these solutions would still cause large delay variance in the same access category (AC) because of the backoff scheme. Noticeably, higher delay variance results in larger probability of quality-of-service (QoS) violation of multimedia traffic due to excess delay.

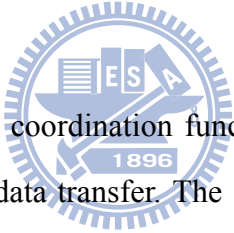
This chapter proposes and analyzes an *adaptive p-persistent-based* (APP) MAC scheme for the IEEE 802.11 WLAN proposed in [46], [47]. The APP MAC scheme, installed in a station, dynamically adjusts the permission probability of transmission for the station itself, and sets the permission probability as a function of the numbers of retransmissions and re-backoffs. The station with longer packet delay, implying larger numbers of retransmissions and re-backoffs, is given higher permission probability. Therefore, the packet delay variance of station for each access can be decreased and the WLAN can provide good delay fairness for stations in each access. The Markov-chain model [20], [49]-[51], is adopted to analyze the proposed APP MAC scheme. The performance measures such as collision probability, system throughput, and mean delay are successfully obtained. Numerical and simulation results show that the APP MAC scheme can effectively reduce the delay variance and thus achieve the delay fairness. The collision probability is decreased and the system throughput is enhanced, compared to conventional schemes. Moreover, discrepancy between numerical and simulation results is provided to corroborate the analyses. These results reveal that the analyses are quite accurate.

For multimedia services, the various initial contention window ( $CW_{min}$ ) and DCF interframe space (DIFS) assigned to each AC, the APP MAC scheme gives different

initial permission probabilities to various ACs to further differentiate their priorities. Moreover, it adaptively adjusts the permission probability of stations in each AC according to their respective waiting delays to reduce the delay variance of stations within the same AC.

The rest of the chapter 2 is organized as follows. Section 2.2 describes the system model, and section 2.3 introduces the APP MAC scheme. The mathematical analysis of the APP MAC scheme is given in section 2.4. Section 2.5 illustrates the performance comparisons of the APP MAC scheme and other conventional methods, such as BEB MAC and DDFC MAC, by numerical and simulation results. Finally, concluding remarks are given in section 2.6.

## 2.2 System Models

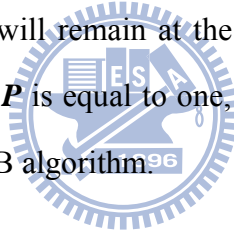


The IEEE 802.11 distributed coordination function (DCF) adopts the CSMA/CA protocol to support asynchronous data transfer. The station can start to transmit only if the medium is sensed idle for a time interval equal to DCF interframe space (DIFS). Otherwise, the transmission is deferred and the BEB algorithm is invoked. In the BEB algorithm, the station chooses a backoff counter from contention window ( $W$ ), before transmitting. At the first transmission attempt,  $W$  is set to the initial contention window,  $W_0$ ; otherwise,  $W$  depends on the number of transmissions failed for the packet. The backoff counter is decremented by one at the end of each slot time,  $\sigma$ , as long as the medium is sensed idle, and suspended otherwise. It will be reactivated when the medium is again sensed idle for a period longer than DIFS. When the backoff counter reaches to zero, the station transmits immediately. A collision will occur when two or more stations transmit simultaneously. This kind of scheme is called *1-persistent*.

An acknowledgement packet sending from the destination station is used to

response to its origination station to denote that the transmitted packet has been successfully received. If the acknowledge packet is not received, it assumes that the transmission has been corrupted. For an unsuccessful transmission,  $W$  is doubled until it reaches to the maximum value of the contention window,  $W_{\max}$ . For a successful transmission, if the station still has packets queued for transmission, it enters a new backoff procedure.

In the APP MAC scheme, its backoff procedure is similar to that of the traditional CSMA/CA MAC scheme with BEB backoff algorithm, except when the backoff counter of a station in a backoff stage decreases to zero. At this instant, the station with the APP MAC scheme may transmit packet with a permission probability  $P$  or enter into a re-backoff procedure with a probability  $(1 - P)$ . Here, the re-backoff procedure is defined as the process of that the station will remain at the same backoff stage with the same contention window. Noticeably, if  $P$  is equal to one, the APP MAC scheme turns to the CSMA/CA MAC scheme with BEB algorithm.



## 2.3 The Adaptive P-Persistent (APP) MAC Scheme

The adaptive p-persistent (APP) MAC scheme is based on the CSMA/CA protocol with a novel APP transmission algorithm. In which, the value of the permission probability  $P$  is adaptively adjusted, according to the state of its packet transmission, which is a function of the number of retransmissions (backoff stages), denoted by  $RT$ , and the number of re-backoffs, denoted by  $RB$ . It is because  $RT$  and  $RB$  can be regarded as measures of delay time of packet transmission. If a station enters into the re-backoff procedure one time, the value of  $RB$  will be added one until up to  $RB_{\max}$ , where  $RB_{\max}$  is the maximum number of re-backoff times. When the value of  $RB$  is equal to  $RB_{\max}$  and the station enters into the re-backoff procedure again, the value of  $RB$  will not be



increased anymore. If a station suffers a collision, the value of  $RT$  will be added one until up to  $BS_{max}$  and the value of  $RB$  will be set to zero, where  $BS_{max}$  is the maximum number of backoff stage. When the value of  $RT$  is equal to  $BS_{max}$  and the station collides again, the station will remain with the value of  $RT$  equal to  $BS_{max}$ . If a station achieves a successful transmission, values of both  $RT$  and  $RB$  will be set to zero. Consequently, the APP MAC scheme can make a station obtain a higher permission probability  $P$  at the same backoff stage if the station has a larger  $RB$ ; it will make a station obtain a lower permission probability  $P$  if the station is in the state with a smaller  $RT$ .

More in details, for a station with the APP algorithm,  $RT$  and  $RB$  are initially zero, and  $P$  is assigned to be  $P_0$  which is the initial permission probability chosen for the first transmission of a ready packet. Afterwards,  $P$  will be adaptively adjusted according to the function designed by



$$P = P_0 + \frac{1 - P_0}{BS_{max}} * \left[ RT + \frac{RB}{1 + RB_{max}} \right], \quad 0 \leq RT \leq BS_{max}, \quad 0 \leq RB \leq RB_{max}. \quad (2.1)$$

The philosophy behind Eq. (2.1) is that a station having larger  $RT$  and  $RB$  should be promoted to have a larger permission probability  $P$ . Also, it is expected that the average waiting time spent at any  $RB$  for a given  $RT$  would be less than that spent at  $(RT+1)$  and  $RB = 0$ . Therefore, it is reasonable that  $P$  is increased by  $(1-P_0)/BS_{max}$  if one more retransmission and  $(1-P_0)/[BS_{max}*(1+RB_{max})]$  if one more re-backoff procedure. The pseudo-code for the APP algorithm in this APP MAC scheme is shown below.

[The APP Algorithm]

if ( $RT = 0$  and  $RB = 0$ )

{

$$P = P_0$$

}

Else

{

$$P = P_0 + (1 - P_0) / BS_{max} * [RT + RB / (RB_{max} + 1)]$$

} °

## 2.4 Analysis

For any station with the APP MAC scheme, define  $s(m)$ ,  $r(m)$ , and  $b(m)$  to be random processes of the backoff stage, the number of re-backoff, and the value of backoff counter, at time  $m$ , respectively, where  $0 \leq s(m) \leq BS_{max}$ ,  $0 \leq r(m) \leq RB_{max}$ , and  $0 \leq b(m) \leq W_i - 1$ ,  $W_i = 2^i W_0$ ,  $W_i$  is the contention window  $W$  of the  $i$ th backoff stage. Also, define  $(s(m), r(m), b(m))$  as the state of system. Assume that there are  $n$  contending stations in the system, and each station is operated in a saturation condition, denoting it always has a ready packet to transmit. The discrete-time observation points are embedded at the end of each slot time, which follows the medium if sensed idle longer than DIFS interval. The three-dimensional random process  $\{(s(m), r(m), b(m))\}$  is a discrete-time Markov chain under the assumptions that both the collision probability and the packet transmission probability of a station are indifferent to its backoff procedure [49]. The collision probability of a station, denoted by  $p_c$ , is the probability of that a station transmits and at least one of the other  $n - 1$  stations transmits; the transmission probability of a station, denoted by  $p_\tau$ , is the probability of that a station transmits at a randomly selected time slot. It is intuitive that this assumption would be more accurate as long as  $W_0$  and  $n$  get larger. Under this assumption,  $p_c$  is supposed to be a constant value. We can obtain the state transition diagram for a station shown in Fig. 2.1 and state transition probabilities given by

$$P\{(i, j, k) | (i, j, k+1)\} = 1, \quad 0 \leq i \leq BS_{\max}, 0 \leq j \leq RB_{\max}, 0 \leq k \leq W_i - 2, \quad (2.2)$$

$$P\{(i, j, k) | (i, j-1, 0)\} = (1 - P_{i,j-1}) \frac{1}{W_i}, \quad 0 \leq i \leq BS_{\max}, 1 \leq j \leq RB_{\max}, 0 \leq k \leq W_i - 1, \quad (2.3)$$

$$P\{(i, 0, k) | (i-1, j, 0)\} = P_{i-1,j} p_c \frac{1}{W_i}, \quad 1 \leq i \leq BS_{\max}, 0 \leq j \leq RB_{\max}, 0 \leq k \leq W_i - 1, \quad (2.4)$$

$$P\{(0, 0, k) | (i, j, 0)\} = P_{i,j} (1 - p_c) \frac{1}{W_0}, \quad 0 \leq i \leq BS_{\max}, 0 \leq j \leq RB_{\max}, 0 \leq k \leq W_i - 1, \quad (2.5)$$

$$P\{(BS_{\max}, 0, k) | (BS_{\max}, 0, 0)\} = P_{BS_{\max},0} p_c \frac{1}{W_{BS_{\max}}}, \quad 0 \leq k \leq W_{BS_{\max}} - 1, \quad (2.6)$$

where  $P\{(i, j, k) | (i', j', k')\} = \text{Prob}\{(s(m) = i, r(m) = j, b(m) = k) | (s(m-1) = i', r(m-1) = j', b(m-1) = k')\}$ , and  $P_{i,j}$  is the permission probability  $\mathbf{P}$  at state  $(i, j, 0)$ . Eq. (2.2) describes the fact that the backoff counter is decremented by 1 at the beginning of each slot time. Eq. (2.3) accounts for the situation that the station re-backoffs again. Eq. (2.4) indicates the case that an unsuccessful retransmission occurs at backoff stage  $i-1$  thus the backoff stage is increased and the new backoff counter is uniformly chosen in the range  $(0, W_i - 1)$ . Eq. (2.5) denotes what a successful packet transmission happens, thus a new packet starts with backoff stage 0 and the initial backoff counter is randomly chosen in the range  $(0, W_0 - 1)$ . Finally, Eq. (2.6) stands for that  $\mathbf{RT}$  is not increased in subsequent packet transmissions, when the backoff stage reaches the value  $BS_{\max}$ .

Define  $\lim_{m \rightarrow \infty} (s(m), r(m), b(m))$  as the system state at steady state. Let  $b_{i,j,k} = \lim_{m \rightarrow \infty} \text{Prob}\{(s(m), r(m), b(m)) = (i, j, k)\}$  be the steady-state probability of the state  $(s(m), r(m), b(m)) = (i, j, k)$ . The state transition equations for  $b_{i,j,k}$  can be obtained by

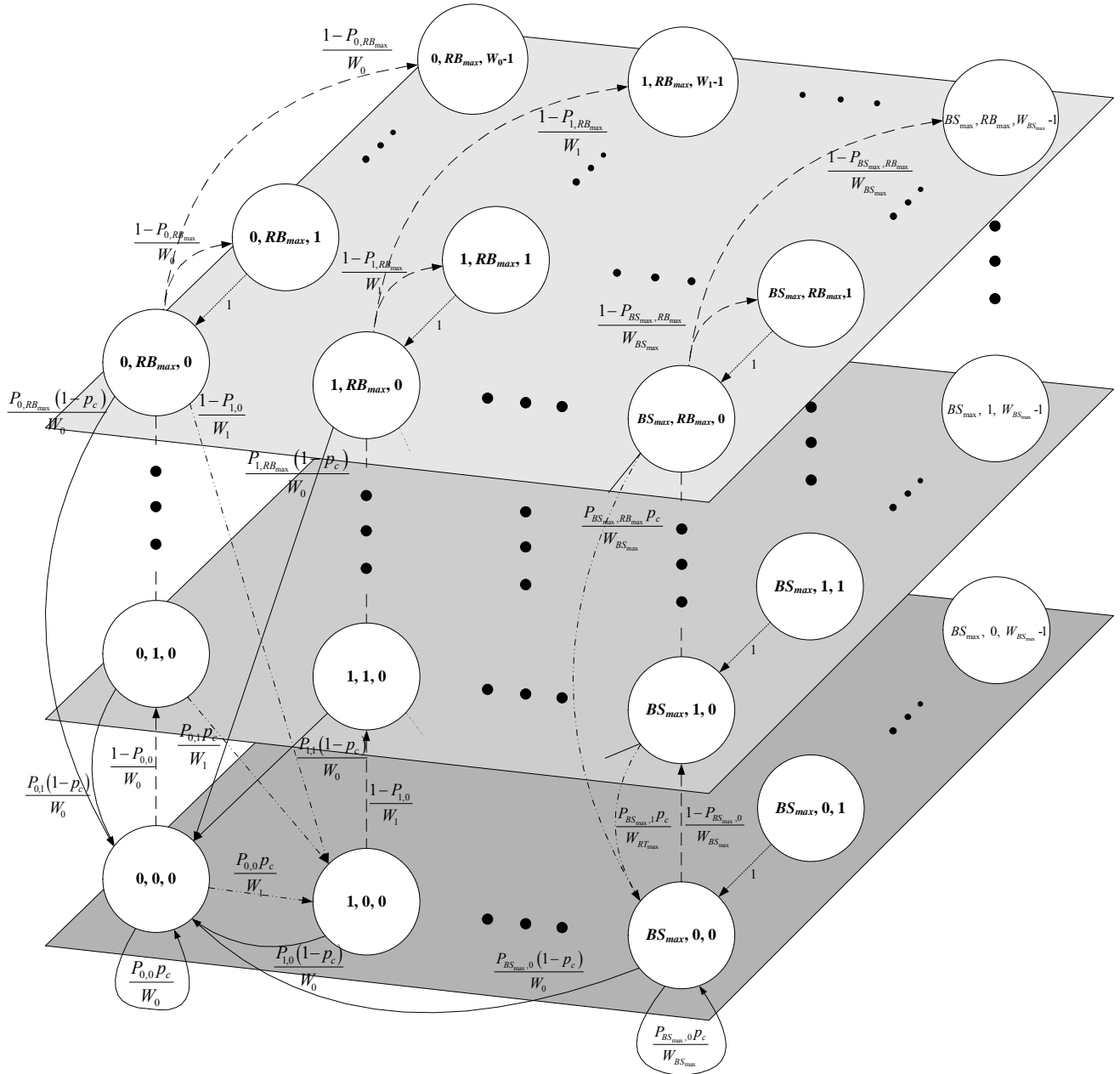


Figure 2.1 State transition diagrams for the APP MAC scheme

$$\left\{ \begin{array}{l}
b_{0,0,k} = \frac{1-p_c}{W_0} \sum_{i=0}^{BS_{\max}} \sum_{j=0}^{RB_{\max}} P_{i,j} b_{i,j,0} + b_{0,0,k+1}, \quad 0 \leq k \leq W_0 - 2, \\
b_{0,0,W_0-1} = \frac{1-p_c}{W_0} \sum_{i=0}^{BS_{\max}} \sum_{j=0}^{RB_{\max}} P_{i,j} b_{i,j,0}, \\
b_{i,j,k} = b_{i,j,k+1} + \frac{1-P_{i,j-1}}{W_i} b_{i,j-1,0}, \quad 0 \leq i \leq BS_{\max} - 1, 1 \leq j \leq RB_{\max}, 0 \leq k \leq W_i - 2, \\
b_{i,j,W_i-1} = \frac{1-P_{i,j-1}}{W_i} b_{i,j-1,0}, \quad 0 \leq i \leq BS_{\max} - 1, 1 \leq j \leq RB_{\max} - 1, \\
b_{i, RB_{\max}, k} = \frac{1}{W_i} \left[ (1-P_{i, RB_{\max}}) b_{i, RB_{\max}, 0} + (1-P_{i, RB_{\max}-1}) b_{i, RB_{\max}-1, 0} \right] + b_{i, RB_{\max}, k+1}, \\
\hspace{20em} 0 \leq i \leq BS_{\max} - 1, 0 \leq k \leq W_{BS_{\max}} - 2, \\
b_{i, RB_{\max}, W_i-1} = \frac{1}{W_i} \left[ (1-P_{i, RB_{\max}}) b_{i, RB_{\max}, 0} + (1-P_{i, RB_{\max}-1}) b_{i, RB_{\max}-1, 0} \right], \quad 0 \leq i \leq BS_{\max} - 1, \\
b_{i,0,k} = \frac{p_c}{W_i} \sum_{j=0}^{RB_{\max}} P_{i-1,j} b_{i-1,j,0} + b_{i,0,k+1}, \quad 1 \leq i \leq BS_{\max} - 1, 0 \leq k \leq W_i - 2, \\
b_{i,0,W_i-1} = \frac{p_c}{W_i} \sum_{j=0}^{RB_{\max}} P_{i-1,j} b_{i-1,j,0}, \quad 1 \leq i \leq BS_{\max} - 1, \\
b_{BS_{\max},0,k} = b_{BS_{\max},0,k+1} + \frac{p_c}{W_{BS_{\max}}} \left[ \sum_{j=0}^{RB_{\max}} P_{BS_{\max}-1,j} b_{BS_{\max}-1,j,0} + P_{BS_{\max},0} b_{BS_{\max},0,0} \right], \quad 0 \leq k \leq W_{BS_{\max}} - 2, \\
b_{BS_{\max},0,W_{BS_{\max}}-1} = \frac{p_c}{W_{BS_{\max}}} \left[ \sum_{j=0}^{RB_{\max}} P_{BS_{\max}-1,j} b_{BS_{\max}-1,j,0} + P_{BS_{\max},0} b_{BS_{\max},0,0} \right].
\end{array} \right. \quad (2.7)$$

Via algebraic manipulation of Eq. (2.7), we can obtain

$$\left\{ \begin{array}{l}
b_{i,j,k} = \frac{W_i - k}{W_i} b_{i,j,0}, \quad 0 \leq i \leq BS_{\max} - 1, 0 \leq j \leq RB_{\max}, 0 \leq k \leq W_i - 1, \\
b_{i,j,0} = \prod_{r=0}^{j-1} (1-P_{i,r}) b_{i,0,0}, \quad 0 \leq i \leq BS_{\max} - 1, 1 \leq j \leq RB_{\max}, \\
b_{i,0,0} = \prod_{m=0}^{i-1} \left( p_c \sum_{r=0}^{RB_{\max}} P_{m,r} \prod_{s=-1}^{r-1} (1-P_{m,s}) \right) b_{0,0,0}, \quad 1 \leq i \leq BS_{\max},
\end{array} \right. \quad (2.8)$$

where  $P_{i,-1}$  is set to be zero. Also from Eq. (2.8),  $b_{i,j,k}$  can be obtained in terms of  $b_{0,0,0}$ , permission probability  $P_{i,j}$ , and collision probability  $p_c$ , by

$$b_{i,j,k} = \frac{W_0 2^i - k}{W_0 2^i} \prod_{h=-1}^{j-1} (1-P_{i,h}) \prod_{m=-1}^{i-1} \left[ p_c \sum_{r=0}^{RB_{\max}} P_{m,r} \prod_{s=-1}^{r-1} (1-P_{m,s}) \right] b_{0,0,0}, \quad (2.9)$$

where  $P_{-l,j}$  is defined to be 1. By using the normalization condition for stationary state probabilities, the  $b_{0,0,0}$  can be yielded as

$$b_{0,0,0} = \frac{1}{\sum_{i=0}^{BS_{\max}} \sum_{j=0}^{RB_{\max}} \sum_{k=0}^{W_i-1} \frac{W_0 2^i - k}{W_0 2^i} \prod_{h=-1}^{j-1} (1 - P_{i,h}) \prod_{m=-1}^{i-1} \left[ p_c \sum_{r=0}^{RB_{\max}} P_{m,r} \prod_{s=-1}^{r-1} (1 - P_{m,s}) \right]}. \quad (2.10)$$

Afterwards, the transmission probability of a station,  $p_\tau$ , can be derived as

$$\begin{aligned} p_\tau &= \sum_{i=0}^{BS_{\max}} \sum_{j=0}^{RB_{\max}} P_{i,j} b_{i,j,0} \\ &= \sum_{i=0}^{BS_{\max}} \sum_{j=0}^{RB_{\max}} \left\{ P_{i,j} \prod_{n=-1}^{j-1} (1 - P_{i,n}) \prod_{m=-1}^{i-1} \left[ p_c \sum_{r=0}^{RB_{\max}} P_{m,r} \prod_{s=-1}^{r-1} (1 - P_{m,s}) \right] \right\} b_{0,0,0}, \end{aligned} \quad (2.11)$$

and the collision probability of station,  $p_c$ , is given by

$$p_c = 1 - (1 - p_\tau)^{n-1}. \quad (2.12)$$

## 2.4.1 System Throughput

For the derivation of system throughput, we consider that the time span is partitioned into three categories: the idle slot time, denoted by  $T_\sigma$ , the successful transmission time, denoted by  $T_s$ , and the collision time, denoted by  $T_c$ . Proportionally, the idle slot time would be with a portion of  $(1 - P_{tr})$ , the successful transmission time would be with a portion  $P_{tr}P_s$ , and the collision time would be with a portion of  $P_{tr}(1 - P_s)$ . The  $P_{tr}$  is the probability of that at least one transmission occurs in a slot time, and it is given by

$$P_{tr} = 1 - (1 - p_\tau)^n. \quad (2.13)$$

The  $P_s$  is the probability of that a successful transmission occurs, conditioned on the fact that at least one station transmits, and accordingly,

$$P_s = \frac{np_\tau(1-p_\tau)^{n-1}}{P_r}. \quad (2.14)$$

Therefore, for a successful transmission of a packet in time  $T_s$ , the system throughput, denoted by  $S$ , can be obtained by

$$S = \frac{P_r P_s B}{(1-P_r)T_\sigma + P_r P_s T_s + P_r(1-P_s)T_c}, \quad (2.15)$$

where the denominator denotes the average time interval taken for this successful transmission, and  $B$  is the average payload size of a packet.

Values of  $T_s$  and  $T_c$  are given by, if the basic access mechanism is adopted,

$$\begin{cases} T_s = H + B_t + SIFS + \delta + ACK + DIFS + \delta, \\ T_c = H + B_t + \delta + DIFS, \end{cases} \quad (2.16)$$

where  $H$  is the time required to transmit PHY and MAC frame headers;  $B_t$  is the average time that a payload is transmitted; SIFS is the duration of SIFS;  $\delta$  is the propagation delay; ACK is the time required to transmit the acknowledgement packet; and DIFS is the duration of DIFS. They are given by, if the RTS/CTS access mechanism is used,

$$\begin{cases} T_s = RTS + SIFS + \delta + CTS + SIFS + \delta \\ \quad + H + B_t + SIFS + \delta + ACK + DIFS + \delta, \\ T_c = RTS + DIFS + \delta. \end{cases} \quad (2.17)$$

Note that collision is assumed to be occurred at RTS frame transmitted.

## 2.4.2 Delay

As those described for Eq. (2.15), the average time interval taken for a successful transmission of a packet is  $(1-P_r)T_\sigma + P_r P_s T_s + P_r(1-P_s)T_c$ , and its probability is  $P_r P_s$ . If the  $n$  contending stations are identical, the average delay of a station, denoted by  $T_D$ , can be obtained by

$$T_D = \frac{n \times [(1 - P_{tr})T_\sigma + P_{tr}P_sT_s + P_{tr}(1 - P_s)T_c]}{P_{tr}P_s}. \quad (2.18)$$

### 2.4.3 The Optimal Value of $P_0$

In WLAN, the number of stations  $n$  is not a directly controllable variable. The way to achieve optimal performance is to employ adaptive techniques to tune the value of  $W_0$  based on an estimated value of  $n$  [49]. Bianchi stated in [49] that the maximum system throughput can be achieved if the optimal initial contention window in BEB, denoted by  $W_{opt}$ , is given by

$$W_{opt} \approx n\sqrt{2T_c/\sigma}. \quad (2.19)$$

In contrast, the initial contention window of the APP MAC scheme since is equivalent to  $W_0/P_0$ , the optimal value of  $P_0$ , denoted by  $P_0^*$ , can be obtained by

$$\begin{aligned} P_0^* &= W_0/W_{opt} \\ &\approx W_0/n\sqrt{2T_c/\sigma}. \end{aligned} \quad (2.20)$$

## 2.5 Numerical and Simulation Results

### 2.5.1 Data Only Environment

Table 2.1 lists system parameters of a considered WLAN environment and values of PHY-related parameters, which are referred to specifications of IEEE 802.11 [3]. In the simulations, we compare the APP scheme with the BEB and the DDFC [15] schemes. In the BEB scheme, two initial contention windows,  $W_0=16$  and  $W_0=32$ , are assumed. In the DDFC scheme, the setting parameters are  $t_0=100\text{ms}$ ,  $t_s=10\text{ms}$  and  $W_0=16$ . Since stations are operated in a saturation condition and the queueing time is not considered in the simulation, the packet waiting time by the DDFC scheme is accounted from the



beginning of packet contention, not as the primary usage defined in [15]. In the following figures, results of APP are shown by numerical and/or simulation, while results of BEB and DDFC are given by simulation.

Table 2.1 Parameter Settings for a WLAN Environment

Slot Time, $\sigma$	20 $\mu$ s
DIFS	60 $\mu$ s
SIFS	10 $\mu$ s
Propagation Delay	1 $\mu$ s
Bit Rate	11 Mbps
PHY Overhead	192 $\mu$ s
MAC Header	28 byte
ACK Length	14 byte
Data Packet Payload, $B$	1028 byte
Max Backoff Stage, $BS_{max}$	4
Initial Contention Window, $W_0$	16
Transmission Retry Limit	$\infty$

Figure 2.2 illustrates the collision probability  $p_c$  of the APP, BEB, and DDFC MAC schemes. It reveals that APP with  $P_0 = 1/4$  achieves an improvement of collision probability by 40% (38.8%) over DDFC (BEB with  $W_0=16$ ), when the number of stations is 8. The reason is that the proposed APP MAC scheme assigns every packet a permission probability  $P$ . When two stations count to zero simultaneously, the collision probability of APP is equal to  $P^2$ . Thus, APP has smallest collision probability; and the smaller the  $P_0$  is, the lower the collision probability would be. This phenomenon is equivalent to making the initial contention window larger. The figure also exhibits that the discrepancy between numerical and simulation results is less than 3.5%, thus this corroborates the collision probability analysis.

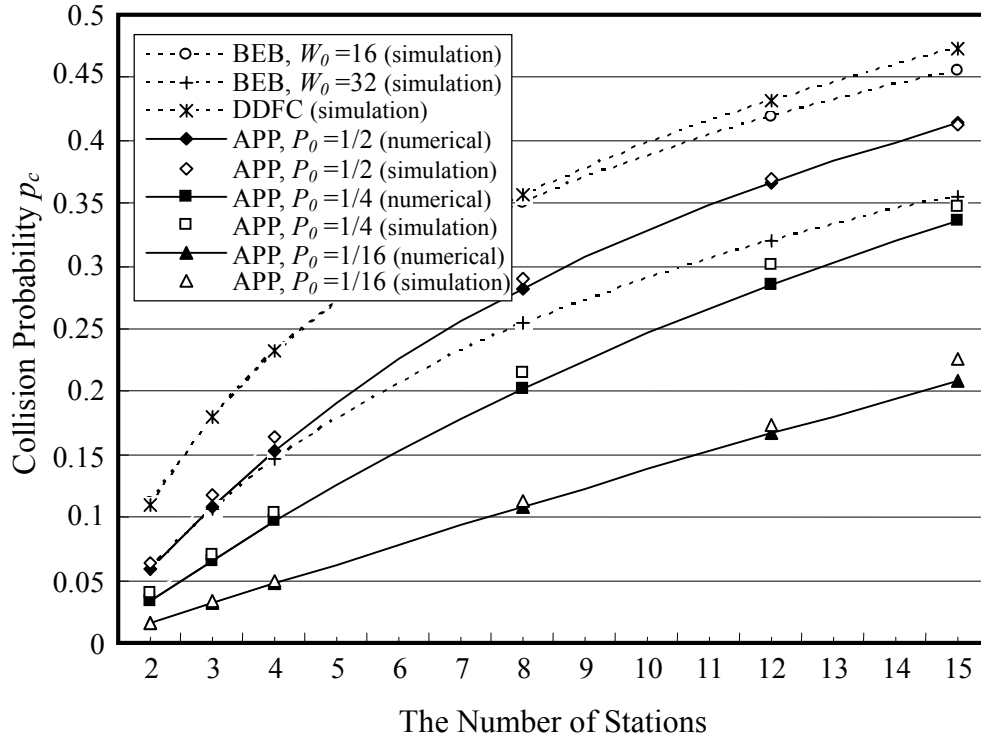


Figure 2.2 Collision probabilities of APP, BEB, and DDFC

Figure 2.3 depicts the system throughputs of the APP, BEB, and DDFC MAC schemes. It can be seen that the throughput increases first and then decreases. It is because increasing the number of stations not only raises the channel utilization but also enlarges the packet collision probability as shown in Fig. 2.2, so the throughput increases first and decreases due to high collision probability. Also, APP with  $P_0=1/4$  achieves an improvement of throughput by 7% (6.5%) over DDFC (BEB with  $W_0=16$ ) when the number of stations is 8. The reason is that APP can reduce the collision probability and increase the transmission efficiency consequently. It can also be found that the smaller  $P_0$  will cause a lower system throughput when fewer stations are in the system. It is because the smaller  $P_0$  is equal to making a larger initial contention window. This will increase the channel idle time and decrease the channel utilization. Noticeably, the difference between numerical and simulation results is also less than 3.5%, which

justifies the validity of the throughput analysis.

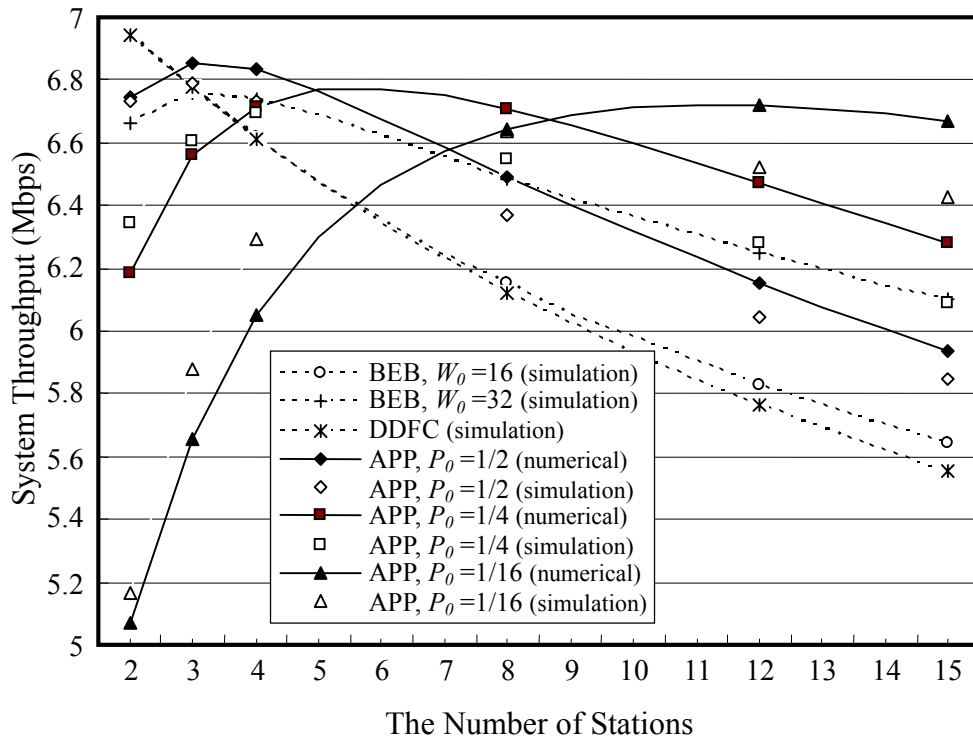


Figure 2.3 System throughputs of APP, BEB, and DDFC

Figure 2.4 shows the mean delays of the APP, BEB, and DDFC MAC schemes. It indicates that the APP with  $P_0=1/4$  achieves an improvement of mean delay by 6.6% (6.1%) over DDFC (BEB with  $W_0=16$ ), when the number of stations is 8. It is because the APP enhances the channel utilization. It can also be found that the smaller  $P_0$  has a larger delay time when there are fewer stations in the system but a smaller delay time when there are more stations in the system. Also, the difference between numerical and simulation results is less than 3.23%, and this substantiates the delay analysis.

Figure 2.5 shows delay variances of the APP, BEB, and DDFC MAC schemes versus the number of stations by simulations. It can be found that the APP MAC scheme possesses the lowest delay variation, while the BEB MAC scheme (BEB with  $W_0=16$ )

the highest. For example, the APP with  $P_0=1/4$  achieves improvement of delay variation over DDFC (BEB with  $W_0=16$ ) by 76.4% (79.4%), at the number of stations is 8. Also, the smaller the  $P_0$  is, the more the improvement of delay variation would be. The reason is the proposed APP scheme adaptively determines the permission probability of transmission according to a function of the number of retransmission (**RT**) and the number of re-backoff (**RB**). The APP scheme lets the ready packet with the longest delay time transmit first and delays the new packet, this makes the delay time of packet be close to the mean value.

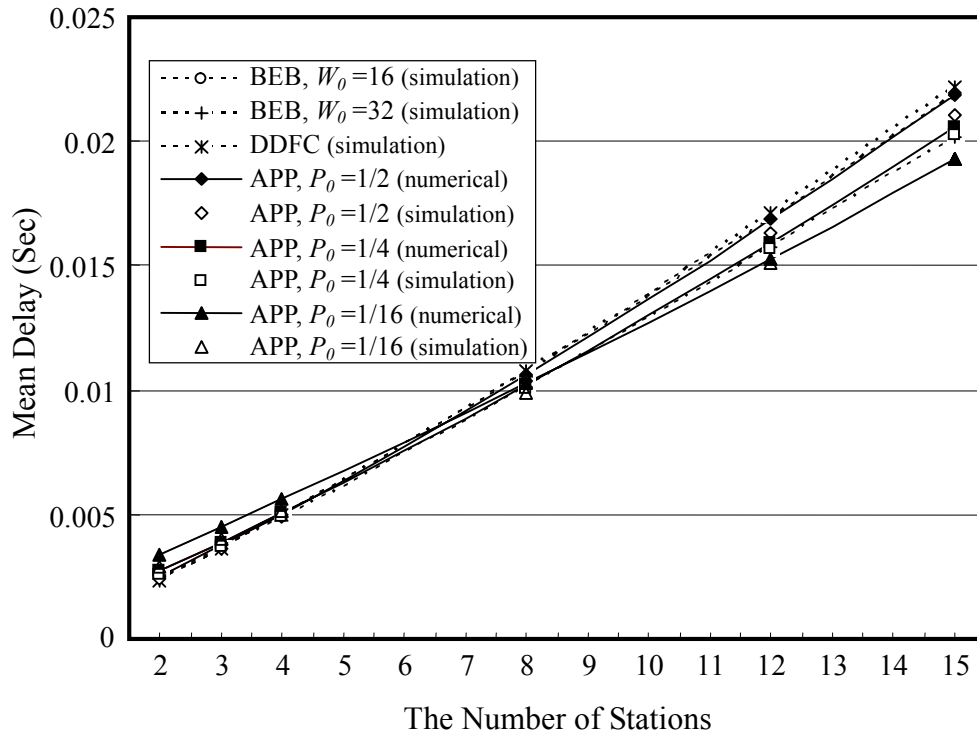


Figure 2.4 Mean delays of APP, BEB, and DDFC

Besides, making  $P_0$  smaller is equivalent to making the  $W_0$  larger, thus lower collision probability. However, the large  $W_0$  in the BEB cannot greatly decrease delay variance and it would cause the system performance degrade (see BEB with  $W_0 = 32$  in

Figs. 2.2 – 2.4). It is because the APP scheme is not actually increase the size of  $W_0$ , but provides another dimension (permission probability  $P$ ) to avoid collision and makes the transmission efficiency, and thus the APP scheme has the smallest mean delay and highest system throughput.

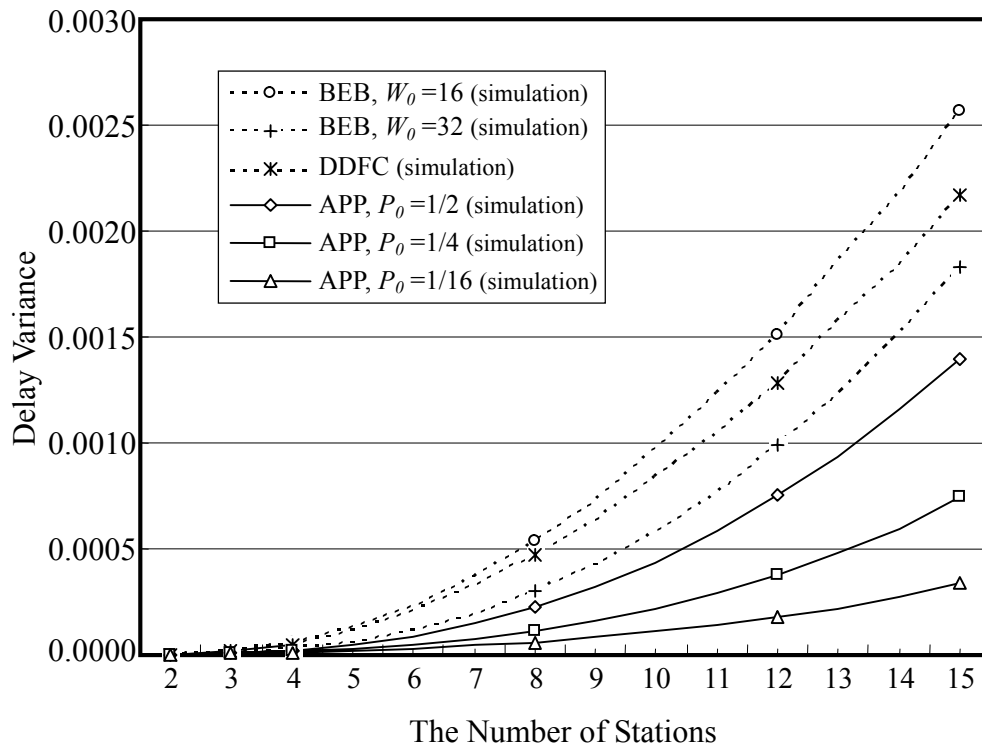


Figure 2.5 Delay variances of APP, BEB, and DDFC

Figure 2.6 shows the system throughput and delay variance of APP with optimal  $P_0^*$  and BEB with  $W_{opt}$  given in [6] by simulations, where the BEB operates with  $W_{opt}$  to obtain the maximum system throughput and the APP uses the optimal  $P_0$  with fixed  $W_0$ . It can be found that APP with optimal  $P_0^*$  loses the system throughput by 1.3% but gains an improvement of delay variation by 15%, compared to BEB with  $W_{opt}$ . This shows that the APP MAC scheme can achieve maximum system throughput and support good delay fairness.

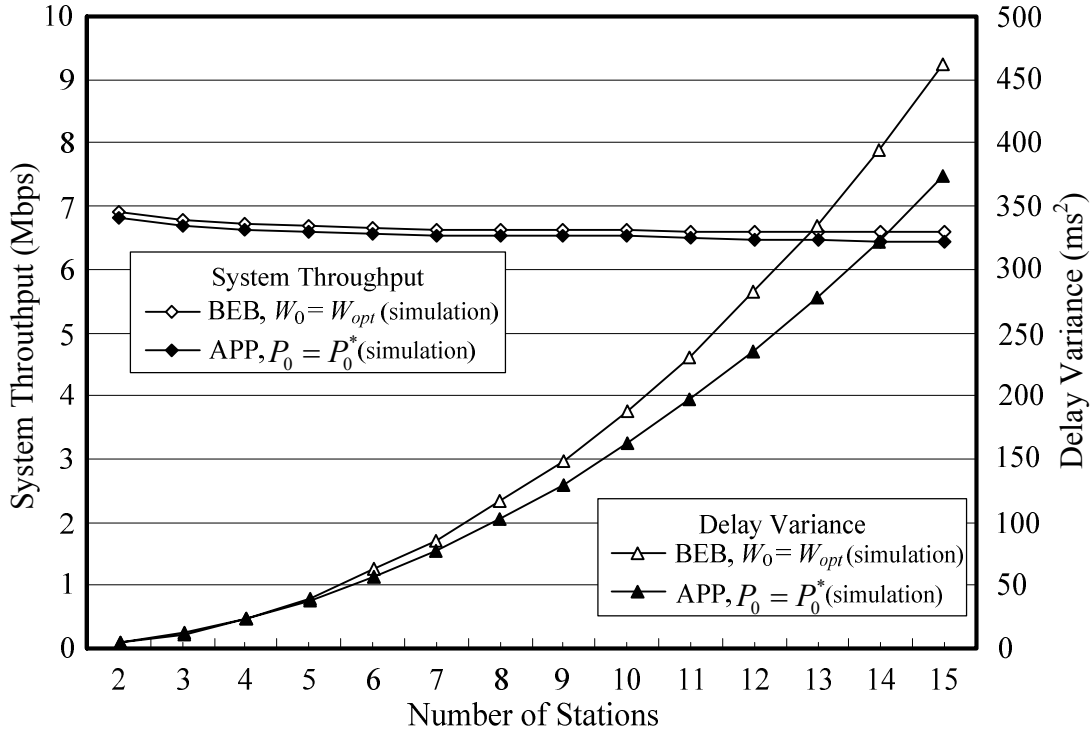


Figure 2.6 performance of APP with optimal  $P_0^*$  and BEB with  $W_{opt}$

## 2.5.2 Multimedia Service Environment

In the simulations, the multimedia WLAN considers three kinds of ACs: high, medium, and low priorities. High (low) priority AC is for voice (data) service, and medium priority AC is for multimedia message service (MMS). Packets generated from high priority AC stations are modeled in an on-off behavior; medium and low priority AC stations are assumed to be in the saturation mode. The packet payload size of high (medium, low) priority AC is 59 (528, 1028) bytes. The value of  $BS_{max}(RB_{max})$  is 5 (5). Also, parameters of the WLAN are set as follows: slot time = 20  $\mu$ s, DIFS for high (medium, low) priority AC = 60 (80, 80)  $\mu$ s, SIFS=10  $\mu$ s, propagation delay = 1  $\mu$ s, bit rate = 11 Mbps, PHY overhead = 192  $\mu$ s, MAC header = 28 bytes, and ACK length = 14 bytes. Values of PHY-related parameters are referred to specifications of IEEE 802.11e [52]. The number of medium (low) priority AC stations is set to be 10 (30), while the

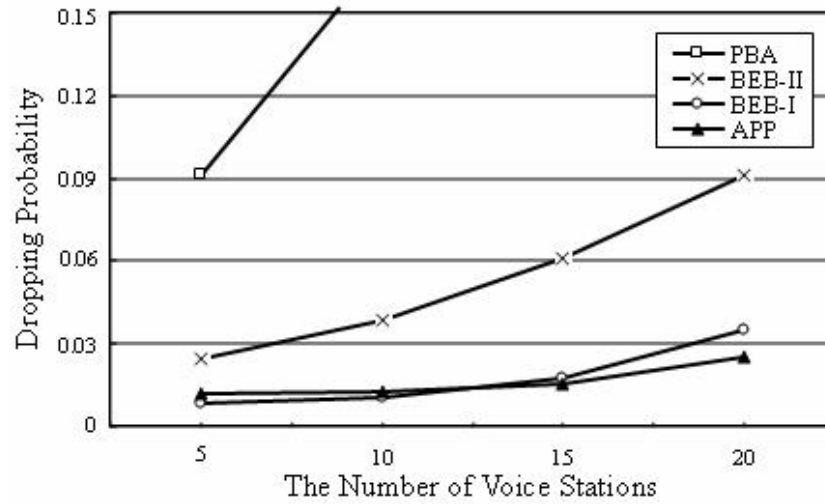
number of high priority AC stations is altered to indicate various traffic load conditions.

The BEB in [52] and the priority backoff algorithm (PBA) in [17] are selected for comparison. In PBA, each station computes the average quantity, in unit of bytes, of successful transmission data of the system. When a station has packet to transmit, it calculates  $CW$  based on the average system quantity and its priority. If the quantity of successful transmission data of the station itself is higher (smaller) than the average system quantity, the station should choose a larger (smaller)  $CW$  to let other station (itself) have higher possibility to access the channel, otherwise it uses the same  $CW$  to select backoff counter.

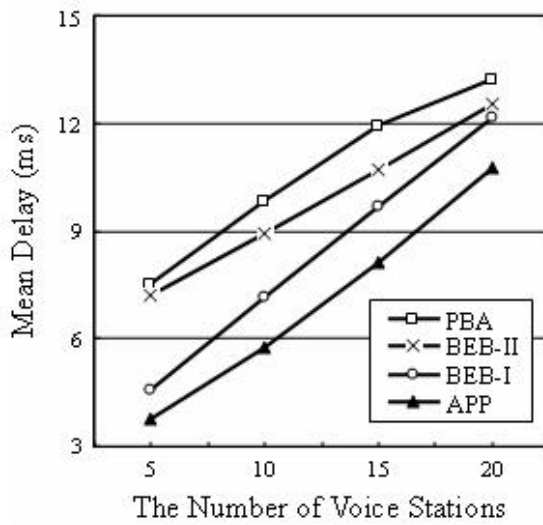
The  $P_0$  ( $CW_{min}$ ) for high, medium, and low priority AC stations in the APP MAC scheme is assumed to be 1/2 (8), 1/16 (24), and 1/32 (32), respectively. The  $CW_{min}$  of all priorities in PBA is set to be 16. The BEB with  $CW_{min}$  equal to 8, 24, and 32 (16, 24, and 32) for high, medium, and low priority AC stations, respectively, is called BEB-I (BEB-II). Define the delay time of a voice packet as the time elapsed between the instant of the packet generation and the instant of the packet reception. A voice packet will be dropped if its delay time is larger than 40 ms. Also, the QoS requirement of voice service is defined as the voice packet dropping probability, which is set to be 3%.

Fig. 2.7 depicts (a) dropping probability, (b) mean delay, and (c) delay variance of voice packets in APP, BEB and PBA versus the number of high priority AC stations. It can be found that the voice packet dropping probabilities of the APP and BEB-I schemes are much smaller than those of the BEB-II and PBA schemes. Also, under the QoS requirement of voice service, APP can accommodate more than 20 voice stations, while BEB-I, BEB-II and PBA can have 18, 7 and 0 voice stations, respectively. The APP performs even better than the BEB-I. The reasons are that the APP further differentiates priorities of ACs by the initial assignment of  $P_0$ , and gives voice service stations a

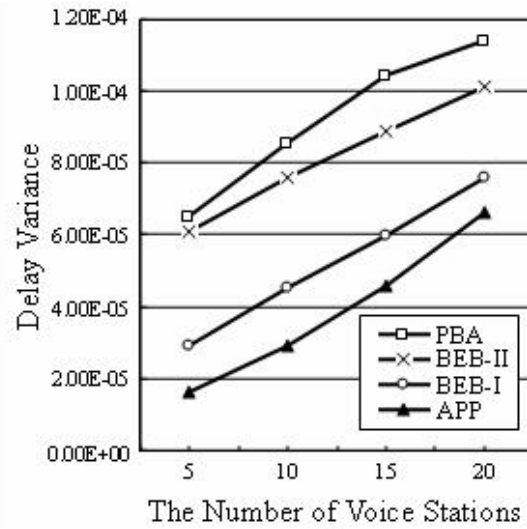
largest  $P_0$  to have a highest priority. Thus, the APP has the least mean delay, which is shown in Fig. 2.7 (b). Moreover, the APP has both the capability of adaptive adjustment of permission probability and the effect of re-backoff procedure. Thus the APP achieves



(a)



(b)



(c)

Figure 2.7 (a) Dropping probability (b) mean delay and (c) delay variance of voice packets

the station's transmission delay approaching to the mean value, and it has the smallest delay variance, which is given in Fig. 2.7 (c). On the other hand, the BEB-II cannot differentiate the priority of voice service from the other two ACs by  $CW_{min}$  more greatly



than APP and BEB-I. Therefore, the increasing of the number of high priority stations would enlarge the collision probability of system. This causes BEB-II has higher mean delay, delay variance, and dropping probability of voice packets. The PBA changes  $CW_{min}$  of high priority stations without considering the number of high priority stations and the various payload size of different priority. In this simulation scenario, the payload size of voice (high priority) packet is much smaller than that of medium and low priority packets, thus the quantity of successful transmission data of high priority station is less than that of the average system. This leads the high priority stations to change their  $CW_{min}$  to a small one and then results in a high collision probability. The phenomenon would make PBA have the highest mean delay, delay variance, and dropping probability of voice packets.

Figure 2.8 shows the system throughput versus the number of high priority stations. It can be seen that APP performs the best and BEB-I performs the worst. When the number of high priority stations is 15, APP achieves an improvement of system throughput over BEB-I, BEB-II, and PBA by 24.1%, 9.9%, and 16.4%, respectively. The reasons are that the APP owns  $P_0$  to differentiate the priority, which can reduce collision probability among stations of different priorities; the APP adaptively adjusts the permission probabilities, which can decrease collision probability among stations in the same AC. Consequently, APP enlarges the channel utilization and enhances the system throughput. Noticeably, when the number of high priority stations is larger than 18, the system throughputs of PBA and BEB-II are a little bit higher than that of APP. That is because APP devotes most of the channel bandwidth to sustain the voice QoS requirement, while PBA and BEB-II violate the voice QoS requirement, which is illustrated in Fig. 2.7 (a).

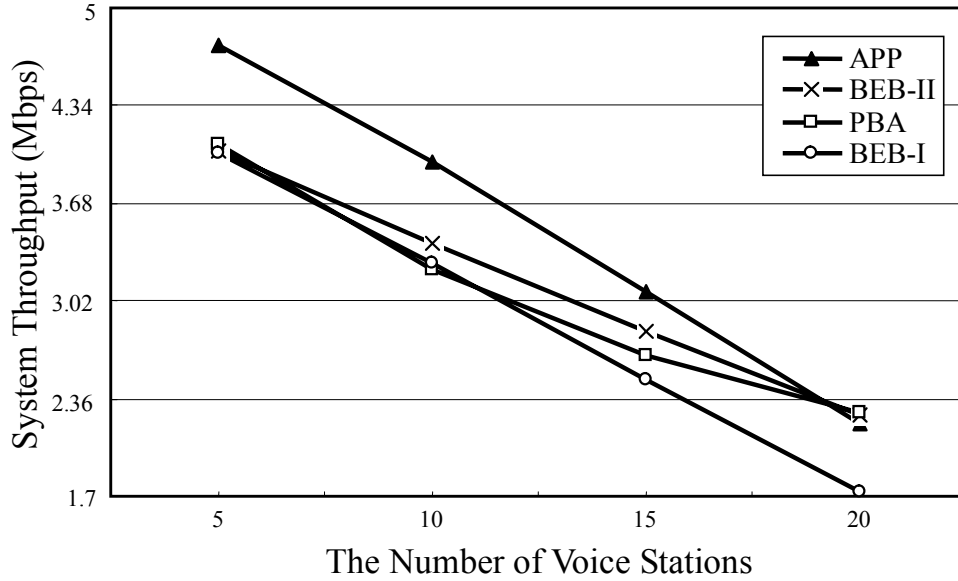


Figure 2.8 System throughput

Figure 2.9 presents the (a) mean delay and (b) delay variance of low priority packets versus the number of high priority stations. It can be found that the APP scheme has the smallest mean delay and delay variance of low priority packet. When the number of high priority stations is 15, the APP achieves by 21.6% (83.5%), 9.6% (78.3%), and 11.1% (16.9%) improvement of mean delay (delay variance) of low priority packet over the BEB-I, BEB-II, and PBA, respectively. Also, these two delay measures for medium priority packets with APP, BEB-I, BEB-II, and PBA have almost the same results as those of low priority packets, which are not shown here. The reason is that  $P_\theta$  in APP provides another dimension to avoid collision and makes the transmission efficiency, and thus APP has the smallest mean delay for medium and low priority packets. Also, both the adaptive adjustment of permission probability and re-backoff procedure of APP for the medium and low priority stations work well, therefore their delay variance is the smallest. On the other hand, the BEB-I differentiates priority more greatly by setting a smaller  $CW_{min}$  for voice stations than the BEB-II. This makes voice stations of BEB-I use a larger portion of channel bandwidth. Therefore medium and low priority stations

with BEB-I cannot access the channel more probabilistically and have mean

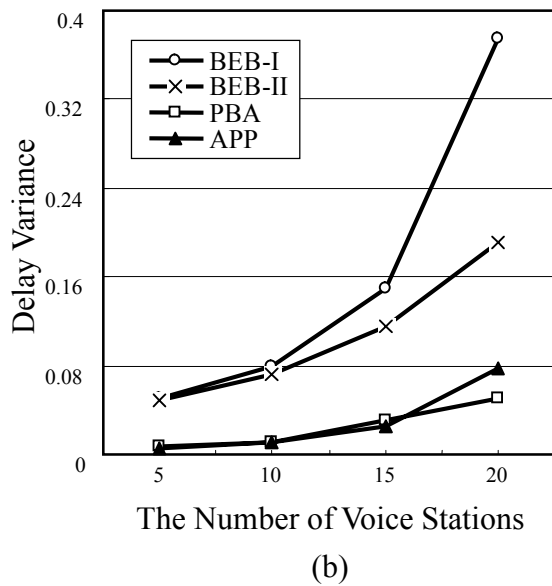
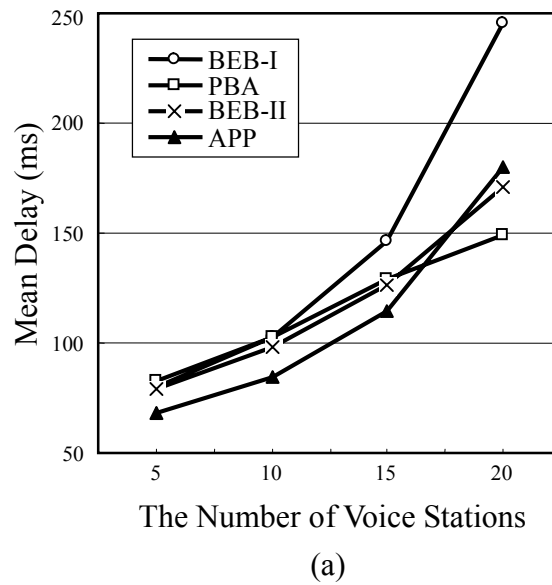


Figure 2.9 (a) Mean delay and (b) delay variance of low priority packet

delay and delay variance higher than those with BEB-II. In PBA, the payload sizes of medium and low priority packets are large, thus the quantity of successful transmission data of medium and low priority stations are larger than system average quantity. These medium and low priority stations would change  $CW_{min}$  up to maximal contention

window to reduce the collision probability of medium and low priority stations. Therefore, their delay and delay variance are smaller than those of BEB-I and BEB-II.

## 2.6 Concluding Remarks

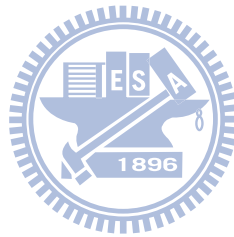
This chapter proposed and analyzed an adaptive p-persistent (APP) MAC scheme for IEEE 802.11 WLAN to achieve fairness in the sense of low delay variance. The APP MAC scheme resolves the fairness problem at each access of stations by adaptively determining the permission probability of station according to the state of packet transmission of the station. It differentiates the permission probabilities of stations with various waiting delay, and assigns a higher priority (probability) to stations with larger packet delay. The chapter analyzes the APP MAC scheme by Markov-chain model and successfully obtains the collision probability, the system throughput, and the mean delay. Results show that the discrepancy between the numerical results and the simulation results is very small, and the analyses are quite correct. Besides, the APP MAC scheme can effectively reduce the delay variance and enhance the system throughput.

The initial permission probability  $P_0$  is an important design parameter in the APP MAC scheme. It can be determined by considering the system design objective which is to reduce the delay variance or enhance the system throughput. Besides, the initial permission probability  $P_0$  can be adaptively determined according to the system load. For example,  $P_0$  could be set to be 1/16 (1/2) when the system is in heavy (light) load.

For multimedia environment, the APP MAC scheme can differentiate stations with various AC of services in multimedia WLAN by setting different initial permission probabilities. Also, it dynamically determines the permission probability of a station in the same AC, according to its transmission state, to reduce the delay variance of the station. Simulation results show that the APP MAC scheme can enhance the

performance of multimedia WLAN; it effectively improves the capacity of high priority stations, reduces the mean delay, enhances the mean throughput, and achieves lower delay variance, compared to conventional algorithms.

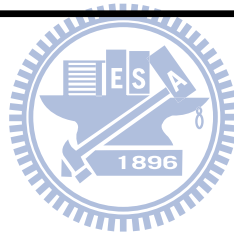
In realistic implementation, the number of rebackoffs (***RB***) and the number of retransmissions (***RT***) are statistical data recorded by stations. The current *CW* of a station can indicate ***RT***, thus only a register is needed in the station to store the value of ***RB***. Also, the value of  $P_0$  ( $RB_{max}$ ) for an AC would be set larger (smaller) if the AC is with more delay sensitive service, for the configuration of WLAN MAC.



# Chapter 3

## Dynamic Priority Resource Allocation for Uplinks in IEEE 802.16 Wireless Communication Systems

---



### 3.1 Introduction

Orthogonal frequency division multiplexing (OFDM) has been proposed as a promising technique for future multimedia wireless communication systems due to its ability to mitigate frequency selective fading, intersymbol interference (ISI) and its flexibility for adaptive modulation on each subcarrier. Orthogonal frequency division multiple access (OFDMA) has been adopted for IEEE 802.16 broadband wireless access (BWA) system. Although the medium access control (MAC) signaling has been well defined in the IEEE 802.16 specifications [42], resource management and scheduling are still remained as open issues. Since the wireless channel condition varies with time, adaptive resource allocation has been viewed as one of the key technologies to provide efficient utilization of the limited system resource in multiuser wireless communication system. Also, for a system with multimedia traffic provisioning, diverse quality of

service (QoS) requirements should be taken into account when developing an efficient resource allocation algorithm. Therefore, an effective resource allocation scheme is required to exploit frequency diversity, multiuser diversity, time diversity, and QoS requirement diversity so that the overall system resource can be efficiently utilized and QoS requirement can be guaranteed.

Subcarrier, bit, and power allocation algorithms for multiuser OFDMA systems to maximize the overall data rate or minimize the total transmitted power under some constraints have been studied in many literatures. Wong et al. [21] proposed a Lagrangian-based algorithm to minimize the total transmission power consumption under user's QoS requirements, which were defined by a specified data transmission rate and bit error rate (BER). However, a high computational complexity renders it impractical. To reduce the complexity, Zhang and Letaief [25] proposed a near optimum dynamic multiuser subcarrier-and-bit allocation algorithm to maximize the overall spectral efficiency.

Many papers considered the downlink resource allocation [24], [25], [28], but a few papers investigated the uplink resource allocation. Resource allocation of both downlink and uplink is primarily performed by the base station (BS). Das and Mandyam [35] considered the uplink transmission of the OFDMA system and developed an efficient algorithm for subcarrier and bit allocation of each user. The algorithm includes the power distribution over the selected set of subcarriers for every user so that the total used power is minimized. Kim, Han, and Kim proposed a joint subcarrier and power allocation scheme for uplink OFDMA systems to maximize the rate-sum capacity based on Shannon capacity formula [36], where a greedy subcarrier allocation algorithm, based on a marginal rate function, and an iterative water-filling power allocation algorithm were proposed. The scheme was shown to achieve a near optimal solution. Jang and Lee

[59] concluded that the equal power allocation algorithm over assigned subcarriers for each user can achieve similar performance to the water-filling scheme. Hosein [37] assumed that subchannels made up of a group of contiguous subcarriers are assigned to users in unit of time slots. Also the CSI on subchannels of each SS is assumed to be reported periodically. Then the optimization problem using a utility function was formulated and a practical algorithm was provided to obtain a near-optimal solution. Singh and Sharma [39] also developed an efficient and fair scheduling (EFS) algorithm for each time slot in IEEE 802.16 OFDMA/TDD system. The EFS algorithm is designed with a fixed priority scheme which gives priorities to service traffic according to their QoS requirements. Chen and Chang proposed a dynamic uplink channel allocation strategy to select a better channel for each SS depending on SS's SNR value [65]. However, the QoS requirements and power constraint are not considered. Also, an efficient uplink resource allocation for power saving in IEEE 802.16 OFDMA systems was proposed in [38]. It adaptively adjusts the modulation and coding scheme to minimize the required transmissions power while guaranteeing BER. However, multiple services and their differentiated QoS requirements are not taken in account.

In the previous works mentioned above, the QoS requirements and fairness issues are either omitted or simplified. A minimum required transmission data rate or a predefined weight which corresponds to the fixed priority scheme is usually adopted as the QoS requirements. However, with the provision of multimedia real-time traffic, the delay bound and the packet dropping rate, regarded as essential QoS requirements, should be included in the design of radio resource allocation for practical applications. Besides, realistic traffic models and buffer conditions of different traffic types should be considered. Niyato and Hossain proposed a queue-aware uplink bandwidth allocation scheme for SS with real-time and non-real-time polling services in IEEE 802.16 system



[40]. The bandwidth is allocated according to channel quality and queue state of the traffic.

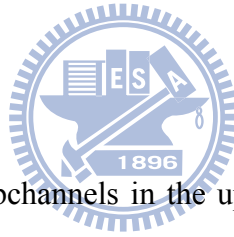
Also, the tradeoff between system performance and computational complexity is an important issue. The greedy algorithm which performs symbol by symbol allocation can achieve optimal solution [41], but it results in high computational complexity. According to the frame structures of DL-MAP and UL-MAP defined in IEEE 802.16 for downlink and uplink, respectively, the symbol-by-symbol allocation algorithm costs high transmission overhead. Besides, most resource allocation algorithms are designed for downlink and claimed to be compatible with uplink as well. However, the downlink frame structure (DL-MAP) and uplink frame structure (UL-MAP) are differently defined in IEEE 802.16 specifications [42]. Thus an efficient and feasible resource allocation algorithm for either downlink or uplink needs to be specifically designed to meet its individual frame structure.

In this chapter, we propose a dynamic priority resource allocation (DPRA) scheme for IEEE 802.16 uplink communication systems. The goal of the proposed DPRA scheme is to maximize system throughput while satisfying various QoS requirements of multimedia traffic. Four types of service traffic for users are taken into account, including unsolicited grant service (UGS), real-time polling service (rtPS), non-real-time polling service (nrtPS), and best effort (BE) service. A priority value for every service type of each user is defined and adaptively adjusted frame by frame according to its urgency related with individual QoS requirements and buffer condition. Then the BS will dynamically allocate the uplink subchannel, modulation order, and power to each SS according to its priority value and the CSI. Furthermore, in order to meet the uplink frame structures defined in IEEE 802.16 specifications and reduce the computational complexity and the transmission overhead, a consistent allocation mechanism is

designed in the proposed DPRA scheme. Simulation results show that the proposed DPRA scheme performs close to the optimal method, which is by exhaustive search, in system throughput. Also, it outperforms the EFS conventional algorithm [39] in system throughput and rtPS packet dropping rate. Besides, the DPRA scheme can take much less computational complexity than the optimal method and the EFS algorithm, where the DPRA scheme is just 1/1000 of the optimal method and 1/10 of the EFS algorithm.

The chapter 4 is organized as follows. The system model of the considered uplink OFDMA system is introduced in section 4.2. Section 4.3 presents the details of the proposed DPRA scheme. Section 4.4 discusses the performance of the DPRA scheme, compared to the efficient and fair scheduling [39]. Finally, conclusions are given in section 4.5.

## 3.2 System Model



Suppose that there are  $N$  subchannels in the uplink of the IEEE 802.16 OFDMA system, and each subchannel consists of  $q$  adjacent subcarriers. There are  $K$  subscriber stations (SSs) going to communicate with one BS in one cell. Each SS can be viewed as a single user containing different service types of traffic to transmit and each service type in an SS has its individual queue. Also, based on IEEE 802.16 uplink rectangle frame structure, traffic data are transmitted in fixed length of frames and each frame contains  $L$  OFDMA slots. Then, the total number of resource units in each frame is  $L \times N$  slots, which is in sequence from the most left of the top subchannel to the most right of the bottom subchannel.

IEEE 802.16 defines the following four service types, and each of them has different QoS requirements: (i) *Unsolicited Grant Service (UGS)*: The UGS supports real-time traffic that generates fixed size of data packets periodically. Thus BS generally

allocates a fixed amount of bandwidth for this type of service. (ii) *Real-time Polling Service (rtPS)*: It is designed to support real-time service which generates variable size data packets. It is a delay sensitive traffic so that the delay requirement is an important QoS issue. The amount of bandwidth granted to this type of service needs to be determined dynamically according to its priority based on the QoS requirements and traffic models. (iii) *Non-real-time Polling Service (nrtPS)*: It is designed to support delay-tolerant data streams while a minimum data transmission rate is required. Also the bandwidth granted to nrtPS needs to be determined dynamically according to its priority based on the QoS requirement and the buffer condition. (iv) *Best effort (BE)*: *BE* service is designed to support data streams which have no QoS requirement. It will be transmitted when system resource is available. Thus the bandwidth left after serving the *UGS*, *rtPS* and *nrtPS* traffic is allocated to *BE* service.

The priority value of service type  $s$  ( $s \in \{UGS, rtPS, nrtPS, BE\}$ ) for user  $k$ , denoted by  $\gamma_{k,s}$ , is here defined in term of the minimum number of bits required to transmit per frame. The  $\gamma_{k,UGS}$  remains constant in each frame since the system needs to grant a constant amount of bandwidth to *UGS*. The  $\gamma_{k,rtPS}$  and  $\gamma_{k,nrtPS}$  are dynamically adjusted frame by frame so that the QoS requirements can be satisfied and the radio resource will be efficiently utilized. The  $\gamma_{k,BE}$  is set to be  $0 \leq \gamma_{k,BE} \leq \gamma_{k,s}$ ,  $s \in \{UGS, rtPS, nrtPS\}$  since there is no delay or transmission rate requirement. Usually,  $\gamma_{k,BE}$  is the smallest and  $\gamma_{k,UGS}$  is the largest but not necessarily.

For *rtPS*, denote  $D_k^*$  the maximum delay tolerance of user  $k$  with a *rtPS* head-of-line (HOL) packet and  $D_k$  the current delay of the *rtPS* HOL packet of user  $k$  experienced, which is the time duration from the arrival frame of the packet to the present frame. Both  $D_k^*$  and  $D_k$  are in unit of frame. The remaining time for the *rtPS* HOL packet of user  $k$  before being dropped, denoted by  $\Delta D_k$ , is given by

$$\Delta D_k \equiv D_k^* - D_k. \quad (3.1)$$

If  $\Delta D_k \leq 0$ , the packet will be dropped and not considered. Therefore, the  $\gamma_{k,rtPS}$  is defined as

$$\gamma_{k,rtPS} = \begin{cases} B_{k,rtPS} & \text{if } 1 \leq \Delta D_k \leq D_{th}, \\ \frac{B_{k,rtPS}}{\Delta D_k + \log(\Delta D_k)} & \text{if } D_{th} < \Delta D_k, \end{cases} \quad (3.2)$$

where  $B_{k,rtPS}$  is the number of residual bits of the *rtPS* HOL packet buffered at the queue of user  $k$ , and  $D_{th}$  is a predefined delay threshold for warning in order to guarantee QoS requirements. If  $\Delta D_k$  is smaller than or equal to the threshold  $D_{th}$ , it means that the *rtPS* HOL packet of user  $k$  is very urgent and all of the residual bits in the buffer had better finish transmission in the current frame. Otherwise, the priority value  $\gamma_{k,rtPS}$  can be set lower based on the average transmission rate,  $B_{k,rtPS} / \Delta D_k$ . We further add its denominator with a bias of  $\log(\Delta D_k)$  to lessen its served transmission bits and make room for other possible high priority users since its residual time before QoS violation is still long. Note that the larger the  $D_{th}$ , the earlier the warning and the better the QoS satisfaction of *RT* services. But in this situation, the system will reserve or consume more resource to protect these *RT* services, and the system throughput will be reduced. Therefore, the  $D_{th}$  should be properly set.

For *nrtPS*, the average transmission rate should be larger than the requirement of the minimum transmission rate, denoted by  $R_{k,nrtPS}^*$ . Denote  $B_{k,nrtPS}$  the number of residual bits of the user  $k$ 's *nrtPS* HOL packet buffered at the current frame and  $\Delta T_k$  the maximum number of frames left for the *nrtPS* HOL packet of user  $k$  at the current frame so that the requirement  $R_{k,nrtPS}^*$  can be fulfilled. Similarly, the  $\gamma_{k,nrtPS}$  is designed as

$$\gamma_{k,nrtPS} = \begin{cases} a \cdot B_{k,nrtPS} & \text{if } \Delta T_k \leq T_{th}, \\ \frac{a \cdot B_{k,nrtPS}}{\Delta T_k + \log(\Delta T_k)} & \text{if } \Delta T_k > T_{th}, \end{cases} \quad (3.3)$$

where  $T_{th}$  is a predefined threshold for *nrtPS* to make obvious priority distinction, the  $\log(\Delta T_k)$  is a bias by the same concept as that for *rtPS*, and  $a$  is a weighting constant,  $0 < a \leq 1$ , which is used to depress the priority of *nrtPS* traffic as compared to that of *rtPS* traffic. Similar to  $D_{th}$ , the  $T_{th}$  should be properly determined.

The transmitted signal of user  $k$  on subchannel  $n$  at the  $\ell$  th OFDMA slot, denoted by  $s_{k,n}^{(\ell)}$ , is given as

$$s_{k,n}^{(\ell)} = \sqrt{\rho_{k,n}^{(\ell)}} \cdot d_{k,n}^{(\ell)}, \quad 1 \leq k \leq K, \text{ and } 1 \leq n \leq N, \quad (3.4)$$

where  $\rho_{k,n}^{(\ell)}$  is the power allocated to user  $k$  on subchannel  $n$ , and  $d_{k,n}^{(\ell)}$  is the transmitted data symbol of user  $k$  on subchannel  $n$  at the  $\ell$  th slot. Note that the normalized  $M$ -QAM modulation is used so that the data symbol has unitary mean energy.

We assume that the coherence time of the wireless channel is larger than the duration of one frame. Hence the CSI is assumed to remain constant over one frame. Besides, perfect estimation of CSI on each subchannel of each user is assumed in this paper. Since in IEEE 802.16 uplink system, the SSs only report the uplink CSI on each subchannel, the channel gain of each adjacent subcarrier which a subchannel contains is assumed to be the same. Let  $h_{k,n}$  be the uplink channel gain between user  $k$  and the considered base station on subchannel  $n$ . Note that the channel gain is not a function of slot time  $\ell$  since it remains fixed during one frame time. The received signal of user  $k$  on subchannel  $n$  at the  $\ell$  th OFDMA slot, denoted by  $y_{k,n}^{(\ell)}$ , is given by

$$y_{k,n}^{(\ell)} = h_{k,n} \sqrt{\rho_{k,n}^{(\ell)}} d_{k,n}^{(\ell)} + \sum_{k' \in K'} h_{k',n} \sqrt{\rho_{k',n}^{(\ell)}} d_{k',n}^{(\ell)} + z_{k,n}^{(\ell)}, \quad (3.5)$$

where  $K'$  is the set of users which use the same subchannel  $n$  at the  $\ell$  th OFDMA slot in other cells, and  $z_{k,n}^{(\ell)}$  is the complex white Gaussian noise of user  $k$  on subchannel  $n$  with zero mean and variance  $\sigma^2$ . The second term at the right-hand side of (3.5) is the co-channel interference from other cells. Therefore, the received signal-to-interference-plus-noise-ratio (SINR) of user  $k$  on subchannel  $n$  at the  $\ell$  th OFDMA slot, denoted by  $SINR_{k,n}^{(\ell)}$ , can be obtained as [66]

$$SINR_{k,n}^{(\ell)} = \frac{\rho_{k,n}^{(\ell)} |h_{k,n}|^2}{\sum_{k' \in K'} \rho_{k',n}^{(\ell)} |h_{k',n}|^2 + \sigma^2}. \quad (3.6)$$

An approximated bit error rate (BER) when using  $M$ -QAM modulation has been given by [60]

$$BER \approx 0.2e^{-1.5 \frac{SINR}{M-1}}. \quad (3.7)$$

From (3.7), the minimum required SINR of user  $k$ , denoted by  $SINR_k^*$ , can be obtained by

$$SINR_k^* = -\frac{\ln(5BER_k^*)}{1.5}(M-1). \quad (3.8)$$

Therefore, based on the  $BER_k^*$  of user  $k$ , the minimum power allocated to user  $k$  on each subcarrier of subchannel  $n$  can be obtained by

$$\rho_{k,n}^{(\ell)} = \frac{-\ln(5BER_k^*)}{1.5|h_{k,n}|^2}(M-1)\sigma^2. \quad (3.9)$$

Besides, the allocated power on each subcarrier of subchannel  $n$  will be equally distributed. Thus the total allocated power to user  $k$  on subchannel  $n$ , which contains  $q$  subcarriers, at the  $\ell$  th OFDMA symbol, denoted by  $p_{k,n}^{(\ell)}$ , can be obtained by

$$p_{k,n}^{(\ell)} = q \cdot \rho_{k,n}^{(\ell)}. \quad (3.10)$$

## 3.3 Dynamic Priority-based Resource Allocation Scheme

### 3.3.1 Problem Formulation

The goal of the proposed dynamic priority-based resource allocation (DPRA) scheme is to maximize overall system throughput and satisfy QoS requirements. Here the allocation problem is mathematically formulated into optimization equations first. Let  $x_{k,n}^{(\ell)}$  be the assignment variable indicating the number of bits carried by each subcarrier of subchannel  $n$  with modulation order assigned to user  $k$  at the  $\ell$  th OFDMA slot. The  $x_{k,n}^{(\ell)}$ ,  $1 \leq k \leq K$ ,  $1 \leq n \leq N$ ,  $1 \leq \ell \leq L$ , is given by

$$x_{k,n}^{(\ell)} = \begin{cases} 6, & \text{if 64-QAM modulation,} \\ 4, & \text{if 16-QAM modulation,} \\ 2, & \text{if QPSK modulation,} \\ 0, & \text{if not assigned.} \end{cases} \quad (3.11)$$

The assignment vector for the  $\ell$  th OFDMA slot over  $K$  users and  $N$  subchannels, denoted by  $\mathbf{x}^{(\ell)}$ , is defined as

$$\mathbf{x}^{(\ell)} \equiv (x_{1,1}^{(\ell)}, \dots, x_{1,N}^{(\ell)}, \dots, x_{k,1}^{(\ell)}, \dots, x_{k,n}^{(\ell)}, \dots, x_{k,N}^{(\ell)}, \dots, x_{K,1}^{(\ell)}, \dots, x_{K,N}^{(\ell)})^T, \quad (3.12)$$

and the assignment vector over the frame, denoted by  $\mathbf{x}$ , can be obtained by

$$\mathbf{x} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(L)}]. \quad (3.13)$$

In such a case, the total number of bits allocated to user  $k$  in one frame, denoted by  $R_k$ , can be calculated by

$$R_k \equiv R_k(\mathbf{x}) = \sum_{\ell=1}^L \sum_{n=1}^N q \cdot x_{k,n}^{(\ell)}. \quad (3.14)$$

The optimization equations for the proposed DPRA scheme is then formulated by

- Objective

$$\text{Throughput maximization: } \mathbf{x}^* = \arg \max_{\mathbf{x}} \sum_{k=1}^K R_k(\mathbf{x}), \quad (3.15)$$

- Allocation Strategy

(i) First allocate the user who has the largest priority value in order to fulfill QoS requirements.

(ii) Consistently allocate the chosen user in order to conform the UL-MAP structure.

- Allocation Constraints

(i) Power constraint:  $\sum_{n=1}^N p_{k,n}^{(\ell)} \leq p_{k,\max}, \quad \forall \ell \text{ and } k. \quad (3.16)$

(ii) Buffer constraint:  $R_k \leq B_k, \quad \forall k. \quad (3.17)$

(iii) Slot allocation constraint:  $\sum_{k=1}^K \text{sgn}(x_{k,n}^{(\ell)}) = 1, \quad \forall n \text{ and } \ell. \quad (3.18)$

In order to fulfill the QoS requirements, the allocation strategy (i), which is the QoS fulfillment strategy, is to first serve the user with the highest priority value. Due to UL-MAP structure, the allocation strategy (ii), which is the consistent allocation strategy, is to continuously allocate slots to the user in the same subchannel. If the allocated slot is to the end slot of the subchannel, it will be continuously allocated from the first slot of the next subchannel. On the other hand, the  $p_{k,\max}$  in (3.16) is the maximum allowable uplink transmission power of user  $k$ . For each OFDMA slot, the total transmission power of each SS should have a cap. The  $B_k$  in (3.17) is the total number of residual bits in the buffer of all service types of user  $k$ . In order not to waste the system resource, the allocated bit to user  $k$  in each frame should not be larger than the total buffer occupancy. To avoid the cochannel interference, the (3.18) indicates that each OFDMA slot can only be allocated to one user. It is a basic constraint for the single-antenna system.



### 3.3.2 DPRA Scheme

The proposed DPRA scheme performs subchannel selection, modulation order, and power allocation assignment for uplink users. The DPRA scheme is also designed with a *consistent allocation mechanism*. Once a subchannel with a modulation order and power is assigned to a selected user at a certain slot, the next consecutive slots at the same subchannel will be consistently given to the selected user until its required transmission of bits completes. Notice that the allocation for each user in a frame by the DPRA scheme takes execution only one time rather than symbol by symbol [36]. Consequently, the DPRA scheme can not only meet the uplink frame structure defined in IEEE 802.16 [42] but also reduce the computational complexity and fulfill the QoS requirement.

The DPRA scheme is a heuristic algorithm which contains six steps of functions to solve the optimization problem given in (3.15)-(3.18). At the beginning, the assignment vector  $\mathbf{x}^{(\ell)}$  and allocated power  $p_{k,n}$  are initialized to be zero, which means that all resources are free. Denote  $\ell_n$  the  $\ell$ th slot of subchannel  $n$  which has been allocated, and  $\delta_k$  the total number of slots allocated to user  $k$ . They are initialized to be zero too. Denote  $N^{free}$  the set of free subchannels of the system and  $N^k$  the set of subchannels allocated to user  $k$ . Initially,  $N^{free} = \{n | 1 \leq n \leq N\}$  and  $N^k = \{\emptyset\}$ ,  $\forall k$ . Also, let  $\Omega$  be the set of backlogged users whose buffers are not empty. The six steps of functions are described as the following.

Step 1) – *User-Subchannel Selection*:

The DPRA scheme selects users from the set of backlogged users  $\Omega$  according to the priority values. Let  $\Omega_h$  be the set of backlogged users having the service with the highest priority value, which is denoted by  $\gamma_{max}$ . In order to achieve the goal of system throughput maximization, the DPRA scheme selects the subchannel with the best CSI in

free subchannel set  $N^{free}$  for the acceptable highest modulation order. In other words, from  $\Omega_h$  and  $N^{free}$ , an optimal pair of user and subchannel  $(k^*, n^*)$  would be chosen to maximize system throughput and fulfill QoS requirement. The function is shown below.

**Function:** *User-Subchannel Selection*

$$\gamma_{\max} = \max_{k \in \Omega} \gamma_{k,s}, \forall s \in \{UGS, rtPS, nrtPS\}$$

$$\Omega_h = \left\{ k \mid k = \arg \max_{k \in \Omega} \gamma_{k,s}, \forall s \in \{UGS, rtPS, nrtPS\} \right\}$$

$$(k^*, n^*) = \arg \max_{k \in \Omega_h, n \in N^{free}} h_{k,n}$$

$$N^{k^*} = N^{k^*} + \{n^*\} \quad \blacksquare$$

Step 2) –*Highest Modulation Order Assignment:*

Once an optimal pair of user and subchannel  $(k^*, n^*)$  is selected, the highest modulation order assignment, its associated power allocation given in (3.10) for user  $k^*$  on subchannel  $n^*$ , and the largest  $x_{k^*,n^*}^{(\ell_{n^*})}$  are performed. Note that the power allocation gets its maximum power constraint, denoted by  $p_{k^*,\max}$ . The selected user  $k^*$  will be removed from the backlogged users set if even QPSK cannot be assigned for power constraint violation. The function is given below.

**Function:** *Highest Modulation Order Assignment*

$$\text{while } p_{k^*,n^*}(BER_{k^*}^*, x_{k^*,n^*}^{(\ell_{n^*})} + 2) < p_{k^*,\max}$$

$$x_{k^*,n^*}^{(\ell_{n^*})} = x_{k^*,n^*}^{(\ell_{n^*})} + 2$$

end while

$$\text{if } x_{k^*,n^*}^{(\ell_{n^*})} = 0, \text{ then } \Omega = \Omega - \{k^*\}, N^{k^*} = N^{k^*} - \{n^*\}, \text{ and go to step 1. } \blacksquare$$

Step 3) –*Allocation Slots Calculation:*

The number of bits that user  $k^*$  can transmit on subchannel  $n^*$  in the first assigned slot is given by  $x_{k^*,n^*}^{(\ell_{n^*})} \cdot q$  since each subchannel contains  $q$  subcarriers. Let  $\alpha_{k^*}$  be the number of slots required for user  $k^*$  to transmit residual bits of HOL

packets of all service types and  $\bar{\alpha}_{k^*,n^*}$  be the number of allocation slots that the system can assign to user  $k^*$  when the allocation starts in subchannel  $n^*$ . Note that if  $\gamma_{\max} > \gamma_{k,BE}$ , this means that there are some more urgent services in user  $k^*$  and its *BE* data will not be considered for allocation this time. Because slots  $(1 \sim \ell_{n^*} - 1)$  of subchannel  $n^*$  have already been allocated to other users before, if remaining slots  $(L - \ell_{n^*} + 1)$  in subchannel  $n^*$  that can be allocated to user  $k^*$  is not sufficient, it is assumed that the available slots of the two subchannels next to subchannel  $n^*$  will be considered for allocation to user  $k^*$  due to the proposed consistent allocation mechanism.

**Function:** *Allocation Slot Calculation*

if  $\gamma_{\max} > \gamma_{k,BE}$

$$\alpha_{k^*} = \left\lceil \frac{\gamma_{k^*,UGS} + B_{k^*,rtPS} + B_{k^*,nrPS}}{x_{k^*,n^*}^{(\ell_{n^*})} \cdot q} \right\rceil$$

else

$$\alpha_{k^*} = \left\lceil \frac{\gamma_{k^*,UGS} + B_{k^*,rtPS} + B_{k^*,nrPS} + B_{k^*,BE}}{x_{k^*,n^*}^{(\ell_{n^*})} \cdot q} \right\rceil$$

$$\bar{\alpha}_{k^*,n^*} = \alpha_{k^*}$$

if  $\bar{\alpha}_{k^*,n^*} > L - \ell_{n^*} + 1$

if  $\bar{\alpha}_{k^*,n^*} - (L - \ell_{n^*} + 1) > L - \ell_{n^*+1}$

if  $\bar{\alpha}_{k^*,n^*} - (L - \ell_{n^*} + 1) - (L - \ell_{n^*+1}) < L - \ell_{n^*+2}$

$$\bar{\alpha}_{k^*,n^*} = \alpha_{k^*}$$

else  $\bar{\alpha}_{k^*,n^*} = (L - \ell_{n^*} + 1) + (L - \ell_{n^*+1}) + (L - \ell_{n^*+2})$

else  $\bar{\alpha}_{k^*,n^*} = \alpha_{k^*}$

else  $\bar{\alpha}_{k^*,n^*} = \alpha_{k^*}$  ■

Step 4) – *Power Rechecking:*

If  $\bar{\alpha}_{k^*,n^*} > L$ , user  $k^*$  can possibly transmit on the same slots of more than one subchannel simultaneously. Since we only check the transmission power on the

subchannel  $n^*$  at Step 2, the power constraint should be rechecked if it is still satisfied. If the power constraint is violated, the number of slots allocated to user  $k^*$  will be decreased until power constraint is fulfilled. The function is given below.

**Function:** *Power Rechecking*

$$c = \lceil \bar{\alpha}_{k^*,n^*} / L \rceil$$

$$\text{if } \bar{\alpha}_{k^*,n^*} > L$$

$$\text{while } p_{k^*,n^*}^{(\ell_{n^*})}(BER_{k^*}^*, x_{k^*,n^*}^{(\ell_{n^*})}) \cdot c > p_{k^*,\max}$$

$$\bar{\alpha}_{k^*,n^*} = L \cdot (c - 1)$$

$$c = c - 1$$

end while

end if ■

Step 5) – *Maximum Available Slots Finding*:

The  $x_{k^*,n^*}^{(\ell_{n^*})} \cdot q \cdot \bar{\alpha}_{k^*,n^*}$  would be the total number of available bits that the system can allocate to user  $k^*$  from subchannel  $n^*$ . If it is smaller than the actual number of required bits, which is  $x_{k^*,n^*}^{(\ell_{n^*})} \cdot q \cdot \alpha_{k^*}$ , the QoS requirements will not be fulfilled. Thus we will search for other subchannels and choose the one for user  $k^*$ . The function is shown below.

**Function:** *Maximum Available Slots Finding*

$$\text{if } x_{k^*,n^*}^{(\ell_{n^*})} \cdot q \cdot \bar{\alpha}_{k^*,n^*} < x_{k^*,n^*}^{(\ell_{n^*})} \cdot q \cdot \alpha_{k^*}$$

$$\text{if } N^{free} - N^{k^*} \neq \{\emptyset\}$$

$$n^* = \arg \max_{n \in N^{free} - N^{k^*}} h_{k^*,n}, N^{k^*} = N^{k^*} + \{n^*\}$$

$$\text{while } p_{k^*,n^*}^{(\ell_{n^*})}(BER_{k^*}^*, x_{k^*,n^*}^{(\ell_{n^*})} + 2) < p_{k^*,\max}$$

$$x_{k^*,n^*}^{(\ell_{n^*})} = x_{k^*,n^*}^{(\ell_{n^*})} + 2$$

end while

end if

$$\text{if } x_{k^*,n^*}^{(\ell_{n^*})} > 0, \text{ go to step 3}$$

$$\text{else } n^* = \arg \max_{n \in N^{k^*}} \left( x_{k^*,n}^{(\ell_{n^*})} \cdot \bar{\alpha}_{k^*,n} \right)$$

else

$$n^* = \arg \max_{n \in N^{k^*}} \left( x_{k^*,n}^{(\ell_{n^*})} \cdot \bar{\alpha}_{k^*,n} \right). \quad \blacksquare$$

Step 6) – *Remapping*:

Due to the consistent allocation, slots allocated to user  $k^*$  must be from a certain  $(\ell_{n^*})$  slot of subchannel  $n^*$  sequentially to slots of the next subchannel ( $n^*+1$  or  $n^*+2$ ). However, slots of the next subchannel may have been allocated to other users. In order to fulfill the slot allocation constraint, the slots which have been allocated to other users had better shift to the neighboring available slots. This process of shifting is called “remapping”. Let  $\sigma_{k^*}$  be the length that needs to be shifted in unit of slot for user  $k^*$ . Thus the user besides  $k^*$  who are originally allocated at the  $\ell$  th slot on subchannel  $n^*$  will be shifted to the  $(\ell + \sigma_{k^*})$  th slot. Note that if  $(\ell + \sigma_{k^*}) > L$ , it will be shifted to the slot of the next subchannel.

**Function:** *Remapping*

$$x_{k^*,n^*}^{(\ell_{n^*+1})} = x_{k^*,n^*}^{(\ell_{n^*+2})} = \dots = x_{k^*,n^*}^{(L)} = x_{k^*,n^*}^{(\ell_{n^*})}$$

$$\sigma_{k^*} = \bar{\alpha}_{k^*,n^*} - (L - \ell_{n^*} + 1)$$

for  $n = n^* + 1$  to  $N$

for  $\ell = 1$  to  $L$

$$\text{if } \sum_{k \in K} x_{k,n}^{(\ell)} > 0$$

$$\bar{k} = \arg \max_{k \in K} (x_{k,n}^{(\ell)})$$

if  $\ell + \sigma_{k^*} \leq L$  and  $n = n^* + 1$

$$x_{\bar{k},n}^{(\ell + \sigma_{k^*})} = x_{\bar{k},n}^{(\ell)}, x_{\bar{k},n}^{(\ell)} = 0, x_{k^*,n}^{(\ell)} = x_{k^*,n^*}^{(\ell_{n^*})}$$

else  $\ell + \sigma_{k^*} > L$  and  $n = n^* + 1$

$$\text{if } \sum_{k \in K} x_{k,n+1}^{(\ell + \sigma_{k^*} - L)} > 0$$

$$\bar{k} = \arg \max_{k \in K} (x_{k,n+1}^{(\ell + \sigma_{k^*} - L)})$$

$\bar{k} = \bar{k}$   
 $\ell^* = \max \{ \ell \mid x_{\bar{k},n+1}^{(\ell)} > 0 \}$   
 $\text{if } \sum_{k \in K} x_{k,n+1}^{(\ell^*+1)} > 0$   
 $\bar{k} = \arg \max_{k \in K} (x_{k,n+1}^{(\ell^*+1)}), \ell^{**} = \max \{ \ell \mid x_{\bar{k},n+1}^{(\ell)} > 0 \},$   
 $x_{\bar{k},n+1}^{(\ell^{**}+1)} = x_{\bar{k},n+1}^{(\ell^*+1)}, x_{\bar{k},n+1}^{(\ell^*+1)} = x_{\bar{k},n}^{(\ell)}, x_{\bar{k},n}^{(\ell)} = 0, x_{k^*,n}^{(\ell)} = x_{k^*,n}^{(\ell_{n^*})}$   
*else*  
 $x_{\bar{k},n+1}^{(\ell^*+1)} = x_{\bar{k},n}^{(\ell)}, x_{\bar{k},n}^{(\ell)} = 0, x_{k^*,n}^{(\ell)} = x_{k^*,n}^{(\ell_{n^*})}$   
*else*  
 $\ell^{***} = \max \{ \ell \mid x_{\bar{k},n+1}^{(\ell)} > 0 \}$   
 $x_{\bar{k},n+1}^{(\ell^{***}+1)} = x_{\bar{k},n+1}^{(\ell+\sigma_{k^*}-L)}, x_{\bar{k},n+1}^{(\ell+\sigma_{k^*}-L)} = x_{\bar{k},n}^{(\ell)}, x_{\bar{k},n}^{(\ell)} = 0, x_{k^*,n}^{(\ell)} = x_{k^*,n}^{(\ell_{n^*})}$   
*else*  
 $\bar{k} = \arg \max_{k \in K} (x_{k,n}^{(\ell)})$   
 $x_{\bar{k},n+1}^{(\ell+\sigma_{k^*}-L)} = x_{\bar{k},n}^{(\ell)}, x_{\bar{k},n}^{(\ell)} = 0, x_{k^*,n}^{(\ell)} = x_{k^*,n}^{(\ell_{n^*})}$   
 end if  
 end  
 end  
 $\Omega = \Omega - \{k^*\}$  ■

Fig. 3.1 shows the flow chart of the proposed DPRA scheme. The DPRA scheme will be continuously executed until there are no free subchannels or no backlogged users. Note that the DPRA scheme is a kind of greedy algorithm, which can find a near optimal solution for the optimization equations given in (3.15)-(3.18) [41]. Also, the functions in Step 3) ~ Step 6) are the consistent allocation customized for IEEE 802.16 systems.

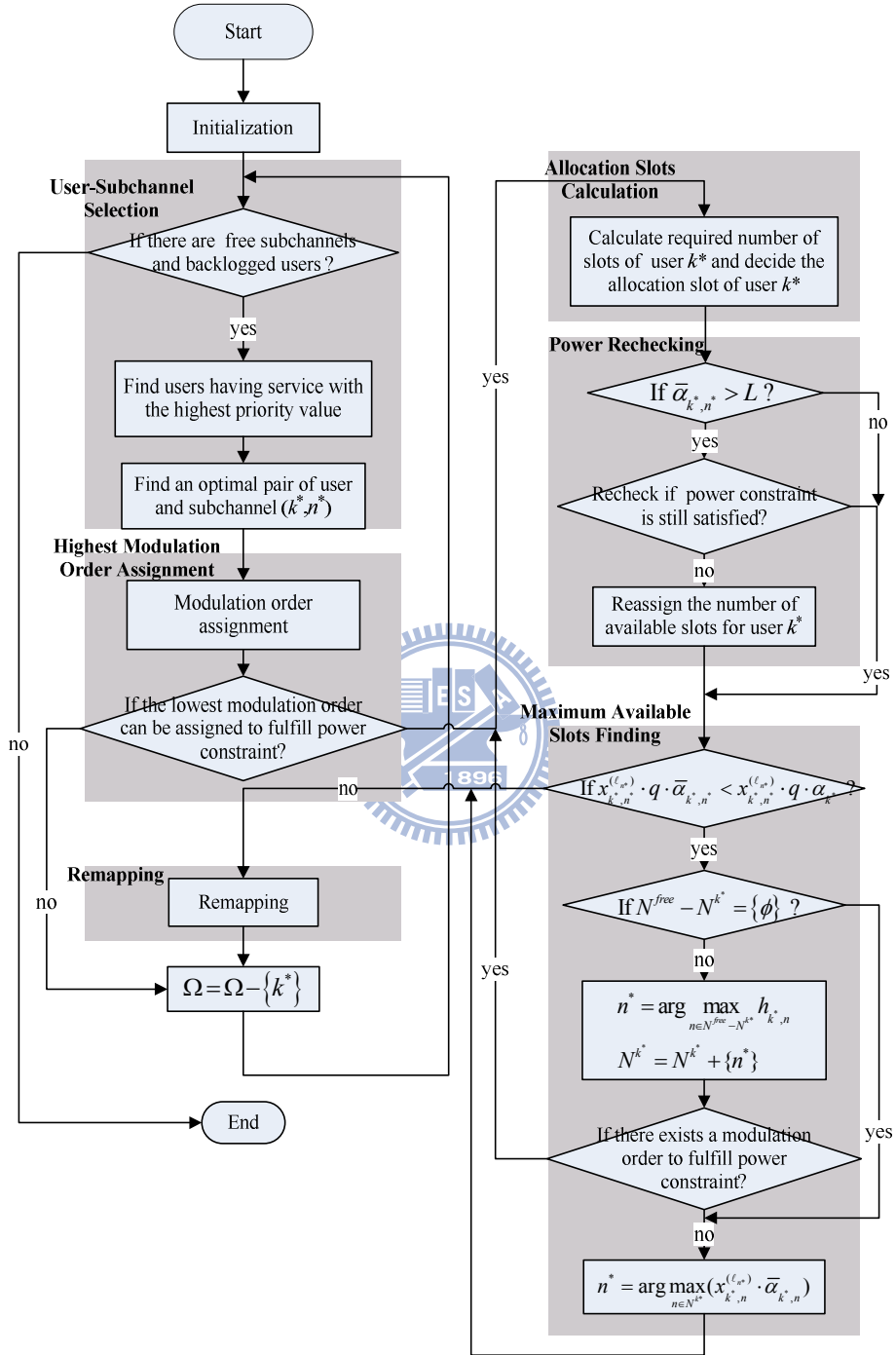


Figure 3.1 Flow chart of the DPRA scheme

## 3.4 Simulation Results

### 3.4.1 Simulation Environment

In the simulations, the system parameters of uplink OFDMA environment are set to be compatible with the IEEE 802.16 standard [42], and the scalable parameters are configured according to the suggested values in [62] and listed in Table 3.1. Also, slow fading and fast fading for wireless channels are considered. The path loss is modeled as  $128.1 + 37.6 \log R$  (dB) [63], where  $R$  is the distance between the BS and SS in unit of

Table 3.1 System-Level Parameters

Parameters	Value
Cell size	1.6 km
Frame duration	5 ms
System bandwidth	5 MHz
FFT size	512
Subcarrier frequency spacing	10.9375 KHz
Number of data subcarriers	384
Number of subchannels	8
Number of data subcarriers per subchannel	48
OFDMA slot duration	102.86 $\mu$ s
Number of slots for uplink transmission per frame	16
Maximum transmission power for each SS	23 dBm
Thermal noise density	-174 dBm/Hz

kilometer. The shadowing model is assumed to be log-normal distributed with zero mean and standard deviation of 8 dB. Six steps of independent Rayleigh distributed path are adopted for the fast fading model and the power delay profile follows the exponential decay rule. The CSI is assumed to be invariant within a frame and varies from frame to frame.

### 3.4.2 Source Model and QoS Requirements

Four types of traffic corresponding to *UGS*, *rtPS*, *nrtPS*, and *BE* are considered in the simulations. The first one is voice traffic for *UGS* and modeled as the ON-OFF



model [64]. The ON (OFF) time is exponentially distributed with mean of 1 (1.35) second. During ON period, one voice packet is generated every 20 ms with fixed size of 28 bytes including payload and header. Thus the mean data rate of voice traffic during the ON period is 11.2 Kbps [68]. The second service type is video streaming traffic for *rtPS*. The video streaming consists of a sequence of video frames generated regularly with an interval of 100ms. Each video frame is composed of eight slices, and each slice corresponds to a single packet. The size of each packet is in truncated Pareto distribution and the inter-arrival time between two consecutive packets is also distributed with truncated Pareto distribution. Parameters of video streaming model are configured according to [63], [64] and the data rate is 64 Kbps.

Table 3.2 QoS Requirements of each service type

Traffic type	Requirement	Value
Voice ( <i>UGS</i> )	Required BER	$10^{-3}$
	Max. delay tolerance	50 ms
	Max. allowable packet dropping rate	1%
Video ( <i>rtPS</i> )	Required BER	$10^{-4}$
	Max. delay tolerance	15 ms
	Max. allowable packet dropping rate	1%
HTTP ( <i>nrtPS</i> )	Required BER	$10^{-6}$
	Min. transmission rate	100 Kbps
FTP ( <i>BE</i> )	Required BER	$10^{-6}$

The third service type is web browsing HTTP traffic for *nrtPS* [63]. It is modeled as a sequence of page downloads, and each page consists of a sequence of packet arrivals. Each page is composed of a main object and some embedded objects, which can be divided into several packets. The maximum transmission unit of each packet is 1500 bytes. The inter-arrival time between two downloaded pages is exponentially distributed with mean 30 seconds. Sizes of both the main object and the embedded object are in truncated lognormal distribution with mean of 10710 bytes and 7758 bytes, respectively. The last service type is FTP traffic for *BE* service, which is modeled as a sequence of file

downloads. The size of each file is in truncated lognormal distribution with mean of 2 Mbytes, standard deviation of 0.722 Mbytes, and a maximum value of 5 Mbytes. The inter-arrival time between two files is exponentially distributed with mean of 180 seconds. Finally, the QoS requirements for each service type configured in the simulation are listed in Table 3.2. Also, values of  $D_{th}$  and  $T_{th}$  given in (3.2) and (3.3) are set to be 2 and 1, respectively; the  $a$  given in (3.3) is set to be 1.

### 3.4.3 Performance Evaluation

For analyzing the performance of the proposed DPRA scheme, the DPRA scheme is compared to the optimal method, which obtains the optimal solution by exhaustively solving the mathematical equations given in (3.15)-(3.18). But when performing the consistent allocation for a user, the optimal method will choose the sort of consistent allocation which achieves the maximum system throughput. The DPRA scheme is also compared to one conventional scheme called the efficient and fair scheduling (EFS) algorithm proposed in [39]. The EFS algorithm allocates slots to each service according the order of *UGS*, *rtPS*, *nrtPS*, and *BE*, where *UGS* (*BE*) has highest (lowest) priority. At each slot time interval, it assigns a slot of the selected subchannel to the user with maximum transmission rate on that subchannel. If all the subchannels of a certain slot interval are exhausted, the EFS algorithm will move to the next slot interval and perform allocation slot by slot iteratively. It is an intuitive algorithm and its performance is close to an optimal solution [39], [41].

For simplicity, based on the allocated subchannel, modulation order, and available slots, the DPRA scheme will perform consistent allocation for each service type of the selected user. Sequential slots on the selected subchannel will be allocated for *UGS*, *rtPS*, *nrtPS*, and *BE* orderly until the available slots is used out. In the simulations, the number

of users is varied from 4 to 40. Each user is assumed to have voice, video, HTTP, and FTP traffic. The maximum system transmission rate per frame which equals to 7.3728 Mbps can be achieved when the highest modulation order is assigned in each slot. We define the traffic intensity as the ratio of the total average arrival rate of all service types of all users over the maximum system transmission rate. Besides, the average arrival rate for voice, video, HTTP, and FTP is 4.8 Kbps, 64 Kbps, 14.5 Kbps, and 88.9 Kbps, respectively.

Figure 3.2 shows the system throughput versus the traffic intensity. It can be seen that the system throughput of proposed DPRA scheme is close to that of the optimal method; the former is just less than the latter by an amount of 3.58% at traffic intensity 0.8. This conforms to the statement in [41]: the result of the greedy method is close to the optimal solution when the number of user is large. It can also be found that the proposed DPRA scheme achieves higher system throughput than the EFS algorithm, especially at higher traffic load. It is because at high traffic intensity, the proposed DPRA scheme can dynamically adjust the priority values, and the more urgent service will be given a higher priority value and allocated more resource to avoid packet dropping. The DPRA scheme can allocate the radio resource more effectively. On the other hand, the EFS algorithm performs slot by slot allocation using fixed priority scheme and can gain the performance close to optimal solution. The EFS algorithm attains the system throughput as large as that of the DPRA scheme until the traffic intensity exceeds 0.6. However, due to the lack of dynamic priority, which reflects the urgency of each service, in EFS algorithm, more packets may be dropped and system throughput degrades at high traffic intensity.

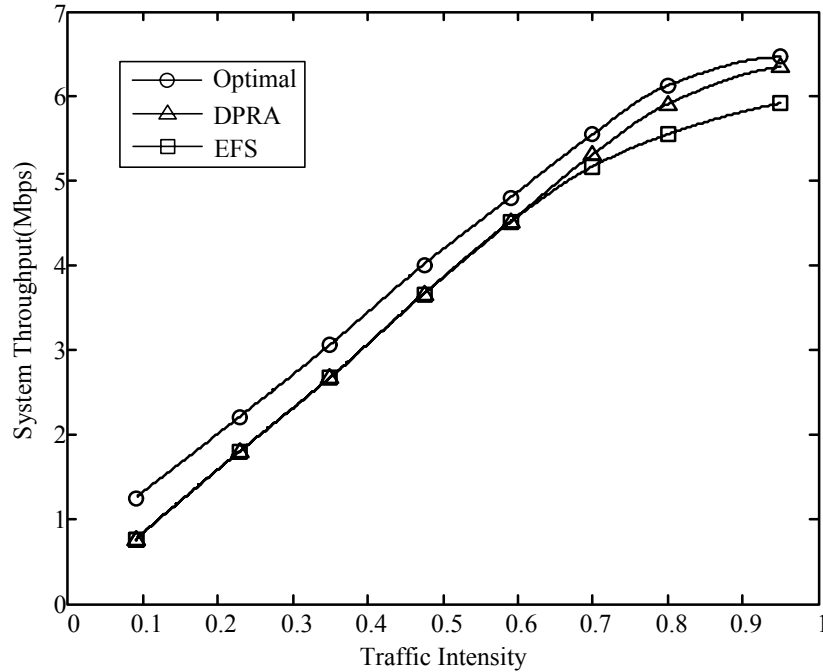


Figure 3.2 System Throughput

Figure 3.3 shows the packet dropping rate of voice traffic. The proposed DPRA scheme, the optimal method, and the EFS algorithm attain the voice packet dropping rates close to zero, which fulfills the voice QoS requirement of 1%. The reason is that these three schemes allocate the voice traffic for *UGS* to a constant amount of bandwidth and the resource allocation is prior to others.

Figure 3.4 shows the packet dropping rate of video traffic. The video packet dropping rate of the DPRA scheme keeps smaller than the QoS requirement of dropping rate (1%) until the traffic intensity is above 0.9. On the other hand, the video packet dropping rate of the optimal method (the EFS algorithm) keeps guaranteed until the traffic intensity is above 0.7 (0.6). The DPRA scheme designs an appropriate dynamic priority value for each service, which is adjusted frame by frame. According to the QoS requirements of video traffic, the priority value of video packets can be promoted and most video packets can be served in time to avoid discarding. But the optimal method is mainly to maximize the system throughput. Thus sometimes, the amount of bandwidth

granted to the video streaming service is not sufficient enough; the EFS algorithm allocates the system resource to video service right after finishing allocation to voice service.

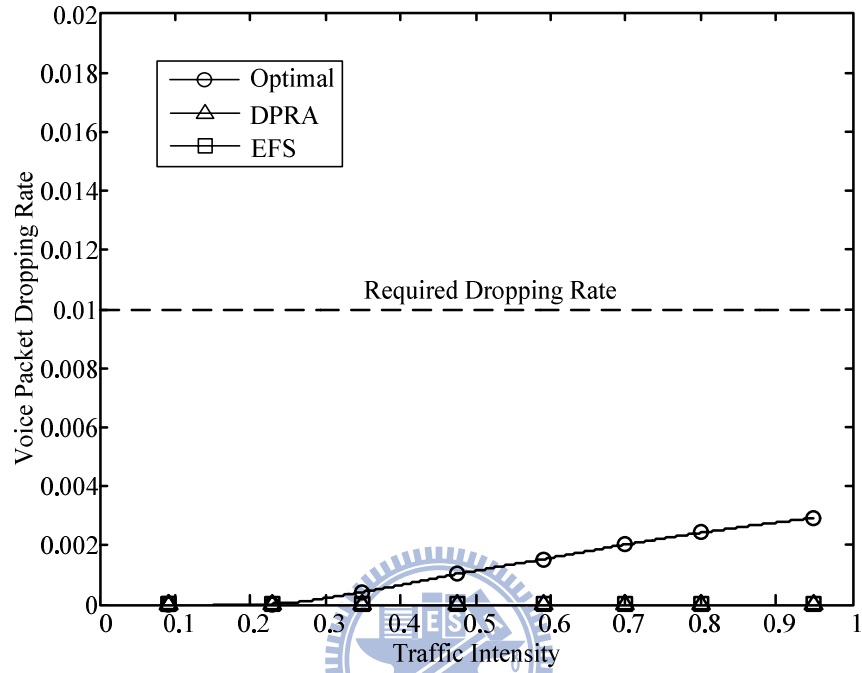


Figure 3.3 Voice Packet Dropping Rate

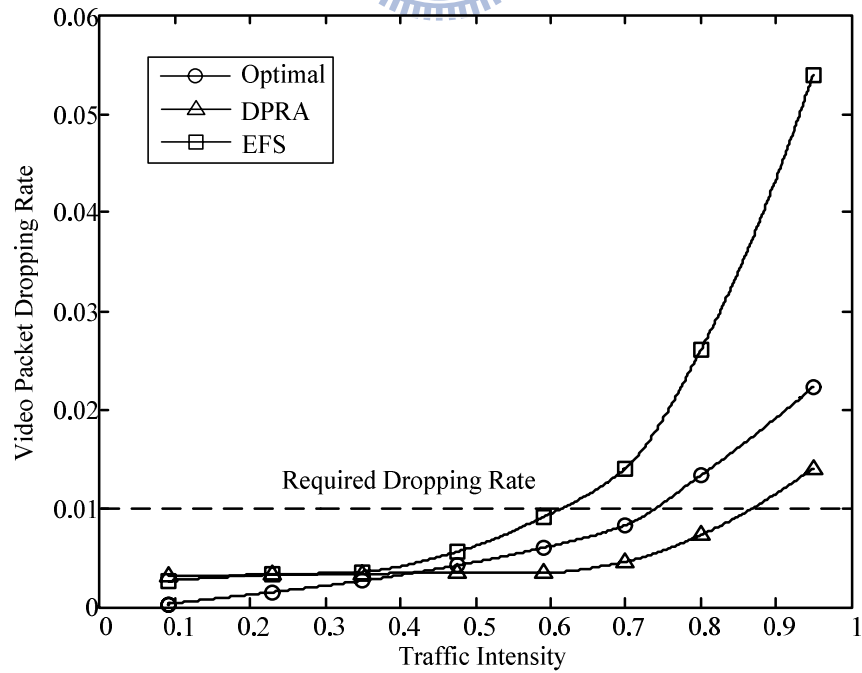


Figure 3.4 Video Packet Dropping Rate

Figure 3.5 shows the ratio of unsatisfied HTTP users, which is defined as the number of HTTP users, whose average transmission rate is less than the minimum required transmission rate, over all HTTP users. For the DPRA scheme, a high priority value will be given to the HTTP user if its average transmission rate is going to be lower than the minimum required transmission rate. Therefore the ratio of unsatisfied HTTP users of DPRA scheme keeps close to zero even when the traffic intensity is high, and the minimum required transmission rate can be guaranteed. On the other hand, the optimal method is mainly to maximize the system throughput. Thus sometimes, the minimum required transmission rate cannot be assured when the traffic intensity becomes high. The EFS algorithm is designed with a fixed priority scheme which initially assigns service traffic with priorities according to their QoS requirements. Thus their ratios of unsatisfied HTTP users will become increasing with traffic load due to lack of enough resource allocated for each HTTP user.

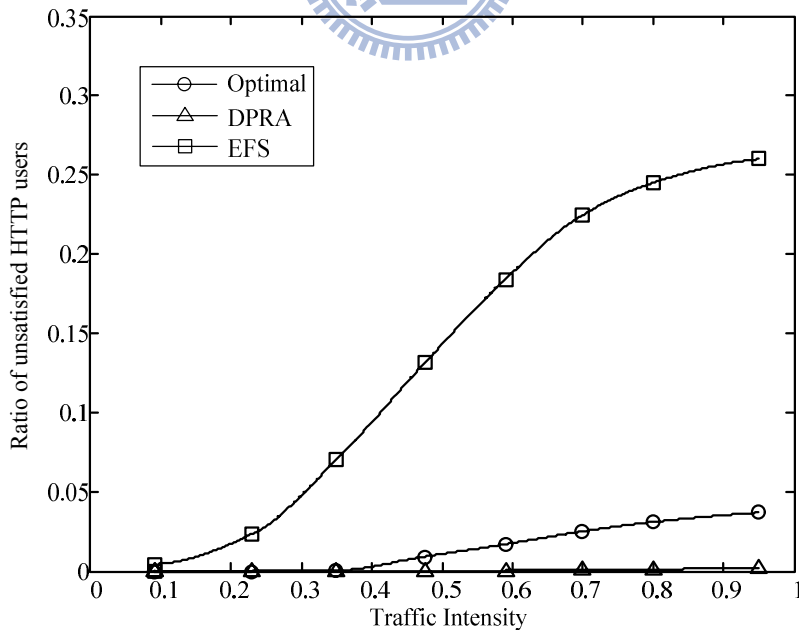


Figure 3.5 Ratio of Unsatisfied HTTP Users

Figure 3.6 shows the average throughput of FTP users. For the EFS algorithm, the

FTP traffic will be transmitted only when voice, video, and HTTP traffic have already been served. Thus its average throughput is the lowest. For the DPRA scheme and the optimal method, the FTP traffic is also specified with the lowest priority value. But the DPRA scheme (the optimal method) achieves the larger (the largest) system throughput, as illustrated in Fig. 3.2, thus comes the result. The optimal method outperforms the DPRA and the EFS scheme by 67.9% and 75.3% in FTP average transmission rate at the traffic intensity 0.8, respectively.

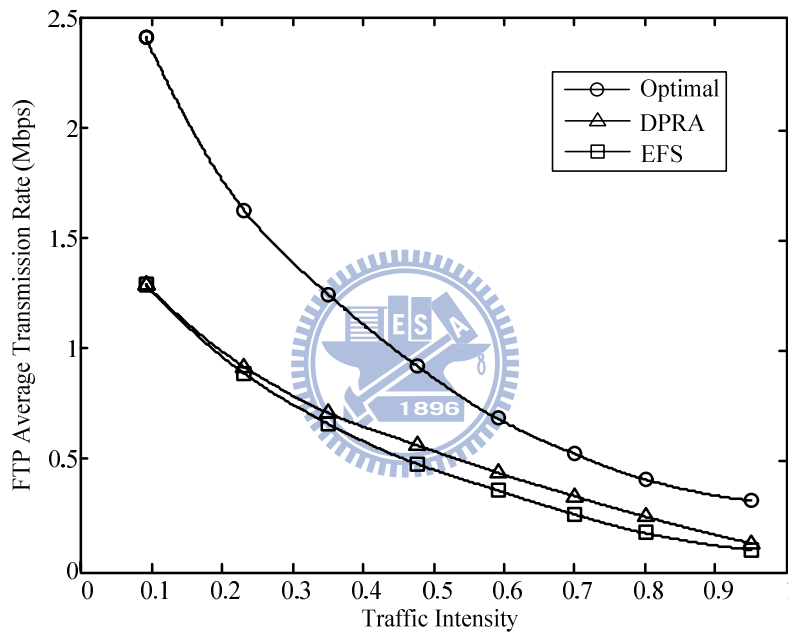


Figure 3.6 FTP Average Transmission Rate

Figure 3.7 shows the average number of iterations per frame for the proposed DPRA scheme, the optimal method, and the EFS algorithm. Here, an iteration is defined as a search for an optimal pair of user and subchannel from  $K$  users and  $N$  free subchannels to be allocated with a slot. The DPRS scheme takes a much smaller number of iterations than the EFS algorithm. It is because the DPRA scheme designs a consistent allocation mechanism, where the iteration computation for allocation to a user in each frame takes only one time and the number of slots allocated to the user could be more

than one. Therefore, the number of iterations by the DPRA scheme is greatly reduced, which is much smaller than the total number of slots,  $N \times L$ , in a frame. On the other hand, the EFS algorithm performs the iteration computation for allocation to a user slot by slot. It searches for an optimal pair of user and subchannel for each symbol. Also, the EFS algorithm may need more than one iterations for each slot allocation to an optimal pair of user and subchannel if there is a power constraint violation. Thus the average number of iterations per frame by the EFS algorithm could be larger than the total number of slots,  $N \times L$ , in a frame. Moreover, in an iteration, the DPRA and the EFS just search for an pair of user and subchannel and check the power constraint. The complexity of an iteration by the DPRA is almost the same as that by the EFS and equal to  $O(KN)$ . It can also be found that the number of iterations by the optimal method is much larger than that by the DPRA scheme. The optimal method takes a number of iterations more than 23,500 times, while the DPRA scheme needs only 23 iterations, at traffic intensity 0.8.

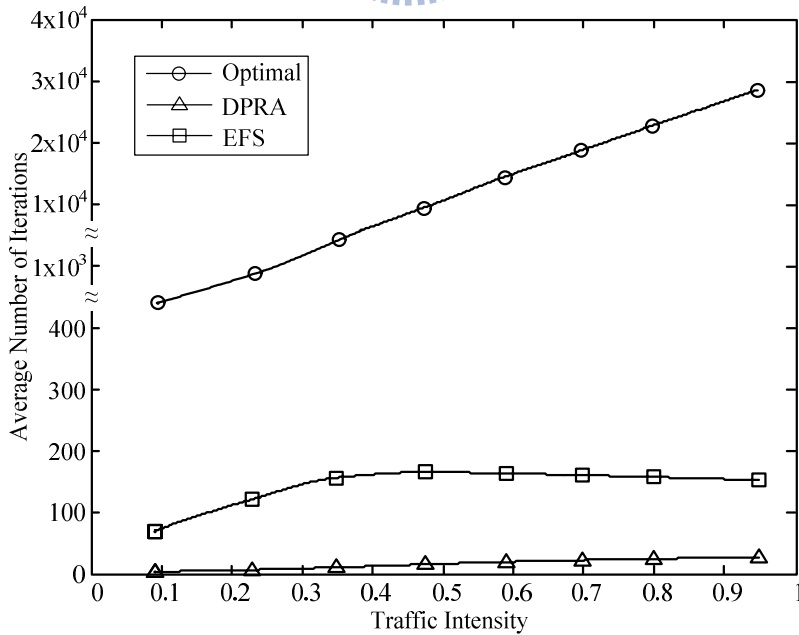


Figure 3.7 Average Number of Iterations

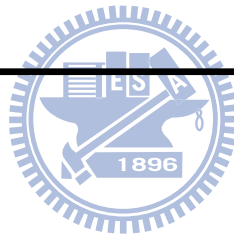


### 3.5 Concluding Remarks

In this chapter, a dynamic priority resource allocation (DPRA) scheme which performs consistent allocation is proposed for IEEE 802.16 uplink system with multimedia traffic. The DPRA scheme intends to maximize the system throughput and fulfill QoS requirements. It originally designs a priority value for each service type according to the urgency and QoS requirements of the traffic and dynamically adjusts it frame by frame. Simulation results show that the performance of the DPRA scheme is better than the conventional EFS algorithm, which performs allocation slot by slot. Besides, benefited from the consistent allocation, the computational complexity of the DPRA scheme is much less than that of the conventional EFS algorithm. Also, the system throughput of proposed DPRA scheme is close to that of the optimal method when traffic intensity is larger, but the computational complexity of the DPRA scheme is much less than that of the optimal solution. Therefore we can conclude that the proposed DPRA scheme can reach throughput maximization and QoS satisfaction with lower computational complexity and transmission overhead. The DPRA scheme would be suitable for real applications.

# Chapter 4

## A Utility-based TMCR Scheduling Scheme for Downlink MIMO OFDMA Systems



### 4.1 Introduction

Multiple-input-multiple-output (MIMO) based orthogonal frequency division multiplexing (OFDM) gives a solution for future wireless communication since it can help to achieve high system capacity and provide transmit/receiver diversity for reliable communication link. Resource management for multiuser OFDM (MU-OFDM) systems has recently been investigated [21]-[25], where topics were focused on transmission power allocation, subcarrier allocation, bit allocation, or adaptive modulation and coding (AMC). The design goal is to maximize system capacity, minimize total transmission power, provide fairness, or guarantee QoS requirements.

Wong et al. proposed an optimal algorithm of subcarrier, bit, and power allocation

for the MU-OFDM system to minimize the total transmission power using Lagrangian multiplier [21]. However, the computational complexity is too high to make it feasible. Yin and Hui studied a so-called low-complexity algorithm, where the total bit rate was maximized, subject to users' rates and total power constraints by decoupling an NP-hard combinatorial problem into two steps [22]. The first step determines the required power and the number of subcarriers for each user; the second step then assigns the subcarrier and the number of bit for each user. However, the computation is still too complicated because of the Hungarian algorithm. Bala and Cimini proposed a low-complexity resource allocation algorithm to minimize the total transmit power using Lagrangian multiplier [23]. This algorithm was based on the idea of linearizing the original problem by transmitting data at the same rate on each subchannel. The computational complexity of this simplification method is still too high. Zhang and Letaief presented a two-step reduced-complexity subcarrier-bit-and-power allocation algorithm, which firstly loosens the rate constraints and allocates subcarriers to maximize the throughput, and afterwards adjusts the subcarrier allocation step by step to satisfy the rate constraints [25]. This iterative algorithm still takes too much time to solve the problem, which is infeasible for real-time applications.

Lee, Choi, and Bahk proposed an optimal opportunistic scheduler to maximize the total utility of a wireless system [53]. Neely presented a theory of utility and delay tradeoffs for general data networks with stochastic channel conditions and randomly arriving traffic [54]. Lei et al. proposed a packet scheduling algorithm in the downlink of OFDMA system for mixed services [55]. They mapped the system resource (bandwidth or power) or performance measures (data rate or delay) to the corresponding utility and optimized the established utility system, where the utility function is either concave or convex. They then used the utility theory to solve the scheduling problem. However,

they did not consider QoS requirements of different service types, and they took only the minimum transmission rate or only the delay time as the QoS requirement. The computational complexity is too high as well.

For multiple-input-multiple-output OFDM (MIMO-OFDM) systems, the computational complexity on radio resource scheduling for downlink multiuser increases exponentially with the number of subcarrier, multiuser, transmitting antenna, and receiving antenna. Fuchs, Galdo, and Haardt [26] proposed a low-complexity space–time–frequency scheduling for MIMO systems with SDMA. Hara, Brunel, and Oshima [27] proposed a spatial scheduling with interference cancellation in multiuser MIMO systems. They focused on maximizing the system throughput but did not consider the QoS requirement. Xu, Wang, and Zhang [28] presented a two-step multiuser scheduling algorithm for system throughput maximization with reduced complexity in a downlink MIMO/OFDMA system with transmitting preprocessing. They decoupled the multiuser scheduling problem into frequency and spatial domains. The preprocessing scheme decomposes the multiuser downlink MIMO channel into multiple parallel independent single-user MIMO channels (like OFDMA). However, the number of simultaneously transmitted users is restricted by the number of transmitting antennas. The computational complexity of the scheduling algorithm is still too high. Hu, Yin, and Yue proposed a low computational complexity scheduling algorithm for a downlink multiuser MIMO-OFDM system in [29]. The more the number of transmitting/receiving antenna and users, the higher the system capacity, which results from space diversity and multiuser diversity. However, the scheduling algorithm did not consider the user demands and QoS requirements; also its symbol by symbol allocation is not suitable for current communication systems.

Yu et al. proposed a QoS guarantee resource allocation algorithm for multiuser

MIMO-OFDMA system, called cross-layer design of packet scheduling (CDPS) [30]. It serves users by considering fixed priority of service traffic. The priority orders of the service traffic are real time service, non-real time service, and best effort service. The real-time traffic can be served in time at low traffic intensity, while the transmission rate of non real-time traffic is too low to fulfill the requirement rate. Tsai et al. proposed a dynamic priority scheduling scheme for downlink OFDMA/SDMA system, called adaptive radio resource allocation (ARRA) scheme [32]. It gives high priority to urgent users and dynamically adjusts the priority of users frame by frame. However, the ARRA scheme adjusts the priority only depending on the time to expiration but not giving the clear differentiation of real-time service from non-real-time one. This may result in that the real-time service may have high possibility to be served after the non real-time traffic at high traffic intensity.

This chapter proposes a utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling scheme for downlink multiuser MIMO-OFDMA systems. The U\_TMCR scheme firstly designs a utility function to adjust the priority of a user. The utility function is composed of a *QoS monitoring function* and a *radio resource function*, where the QoS monitoring function is to indicate the service priority and the degree of urgency of the user, and the radio resource function is to show the spectrum efficiency if it is used by the user. Subsequently, the U\_TMCR scheme formulates the scheduling problem into utility-based optimization equations with an objective to maximize the overall system utility value under two system constraints. Finally, the U\_TMCR scheme proposes a heuristic TMCR algorithm to solve the utility-based scheduling problem. The heuristic TMCR algorithm is a greedy method to maximize the overall system utility value. It contains two functions. The first one is a maximum utility allocation (MUA) function, which finds the most appropriate

subchannel for the user and determines the receiving antenna by a multiple antenna selection (MAS) scheme [56], [57] according to the highest individual utility value. The second one is a consistent allocation (CSA) function, which generalizes the allocation result determined by the MUA function on the following consecutive OFDMA symbols according to the urgency value such that users' QoS requirements can be fulfilled and computational complexity can be reduced.

Simulation results show that the proposed U\_TMCR scheme can achieve the system throughput very close to the optimal solution of the optimization equations by exhaustive search and even higher than those of the ARRA scheme and the CDPS scheme by an amount of about 8% and 21%, respectively. Also, the overall QoS satisfaction degree by the U\_TMCR scheme is higher than those by the ARRA scheme and the CDPS scheme. The packet dropping probability of real-time services can be kept lower than the requirement by the U\_TMCR scheme, but those by the ARRA and the CDPS schemes would violate the QoS requirement at high traffic intensity. Moreover, the U\_TMCR scheme reduces computational complexity. Generally, the total number of allocation trials of the U\_TMCR scheme in a frame is reduced by 6.25% ~ 29.2%, compared with the ARRA scheme.

The chapter is organized as follows. The system model of the considered downlink multiuser MIMO-OFDMA system is described in section 4.2. Section 4.3 presents the U\_TMCR scheduling scheme. Section 4.4 discusses the performance of the U\_TMCR scheduling scheme. Finally, conclusions are given in section 4.5.

## 4.2 System Model

Consider a downlink multiuser MIMO-OFDMA system with the proposed U\_TMCR scheduling scheme. As shown in Fig. 4.1, assume that the system has  $K$  active

mobile users, each mobile user  $k$  has  $Q_k$  receiving antennas, and the base station (BS) has  $N$  subchannels and  $Q$  transmitting antennas. The time is divided into frames and there are  $L$  OFDMA downlink symbols per frame. At the beginning of a frame, the U\_TMCR scheme computes individual utility value for each user with QoS requirements over all subchannels and receiving antenna. It allocates subchannels, antenna sequence, and modulation order to users based on these utility values.

By the allocation order from the U\_TMCR scheduling scheme, the subchannel allocation block chooses the subchannel, the antenna allocation block assigns the antenna sequence, and the adaptive modulation block selects the modulation order to the user. Subsequently, the complex symbols at the output of the adaptive modulation block are transformed into the time domain samples by inverse fast Fourier transform (IFFT). After IFFT and cyclic prefix (CP) addition, the transmission signal is then passed through different frequency selective fading channels to different users. The CP is the guard interval, added to ensure orthogonality between the subchannels. If the length of the CP is longer than the maximum time dispersion, the ISI is mitigated.

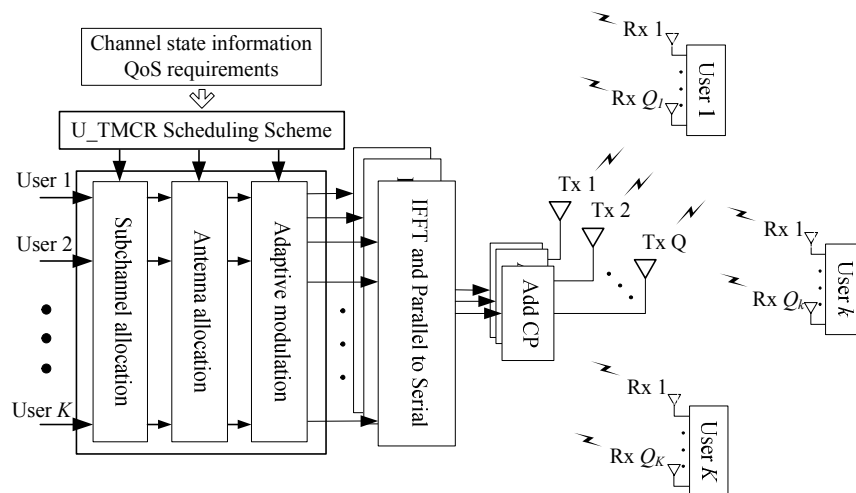


Figure 4.1. System configuration of the downlink MIMO-OFDMA system.

## 4.2.1 System Assumptions

In the multiuser MIMO-OFDMA system, assume that the MIMO channel on each subchannel is regarded as flat fading and correlations among the receiving antennas of mobile users are negligible. The general channel matrix of the MIMO channel on a subchannel, denoted by  $\mathbf{H}$ , is expressed in the form as

$$\mathbf{H}^T = [\mathbf{H}_1, \dots, \mathbf{H}_k, \dots, \mathbf{H}_K], \quad (4.1)$$

where  $\mathbf{H}^T$  is the transpose vector of  $\mathbf{H}$ ,  $\mathbf{H}_k$  is the channel response of user  $k$ ,  $\mathbf{H}_k = [H_{q,r}]_{Q \times Q_k}$ ,  $1 \leq k \leq K$ , and  $H_{q,r}$  is the impulse response between the transmitting antenna  $q$  at BS and the receiving antenna  $r$  at user  $k$ .

In order to achieve the orthogonality among selected users' MIMO channels, we adopt the multiuser antenna selection (MAS) scheme [56], [57]. The MAS scheme first chooses one receiving antenna from mobile users, which can obtain a maximum multiple-input single-output (MISO) channel capacity between the receiving antenna and the transmitting antennas at BS. Then, it selects the next receiving antenna from the remains by the same way. Note that, in order to guarantee the orthogonality, the next selected receiving antenna must not reduce the capacity of the previously selected receiving antennas. Repeat the above process until the number of receiving antennas is equal to  $Q$ , the number of transmitting antennas. Consequently, the virtual MIMO channel matrix between the transmitting antennas at BS and the selected receiving antennas at chosen users on a subchannel can be constructed as

$$\bar{\mathbf{H}}^T = [\bar{\mathbf{H}}_1, \dots, \bar{\mathbf{H}}_{k'}, \dots, \bar{\mathbf{H}}_Q]_{Q \times Q}, \quad (4.2)$$

where  $\bar{\mathbf{H}}_{k'}$  is the  $Q \times 1$  vector of impulse response between the transmitting antennas at BS and the selected receiving antenna at chosen user  $k'$ ,  $1 \leq k' \leq Q$ .

The zero forcing beamforming [56] weight matrix on the subchannel, denoted by



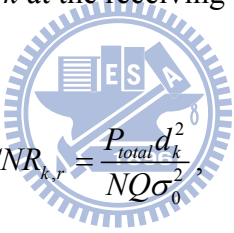
$\mathbf{W}$ , can then be calculated by

$$\mathbf{W} = \overline{\mathbf{H}}^H (\overline{\mathbf{H}}\overline{\mathbf{H}}^H)^{-1} \mathbf{D}, \quad (4.3)$$

where  $\mathbf{D} = \text{diag}(d_1, \dots, d_k, \dots, d_K)$  is the diagonal matrix which keeps the transmitting power unchanged after beamforming and  $\overline{\mathbf{H}}^H$  means the Hermitian transpose of  $\overline{\mathbf{H}}$  [58]. The diagonal elements  $d_k$  is defined as

$$d_k = 1 / \sqrt{[(\overline{\mathbf{H}}\overline{\mathbf{H}}^H)^{-1}]_{k,k}}. \quad (4.4)$$

Because it is concluded that the equal power allocation algorithm over assigned subcarriers for each user can achieve performance similar to the water-filling scheme [59], and in this work the equal power allocation is adopted. Therefore, the signal-to-noise ratio (SNR) of user  $k$  at the receiving antenna  $r$  on the subchannel can be given by

$$SNR_{k,r} = \frac{P_{total} d_k^2}{NQ\sigma_0^2} \quad (4.5)$$


where  $P_{total}$  is the total power of system and  $\sigma_0^2$  is the noise power.

## 4.2.2 Urgency Values

There are three kinds of service classes provided for users: real-time (*RT*) service, non real-time (*NRT*) service, and best-effort (*BE*) service. For *RT* services, requirements of delay bound ( $D^*$ ), bit error rate ( $BER^*$ ), and dropping probability ( $P_D^*$ ) are considered. For *NRT* services, requirements of minimum transmission rate ( $R_{min}^*$ ) and  $BER^*$  are taken into account. For *BE* service, only  $BER^*$  is required. In this work, there are two kinds of *RT* services: voice and video, one kind of *NRT* service: HTTP, and one kind of *BE* service: FTP.

For fulfillment of users' QoS requirement, an urgency value for a user is defined.

The urgency value of a user with  $RT$  or  $NRT$  service is to indicate how much its tolerance time left. The *tolerance time* of a user is defined as the residual time that the head-of-line (HOL) packet of the user at the present frame remains before its violation of QoS requirements. The smaller (larger) the tolerance time, the higher (lower) the urgency value. Denote  $\zeta_{RT}(k)$  and  $\zeta_{NRT}(k)$  the urgency values of user  $k$  with  $RT$  and  $NRT$  services, respectively, and define them as

$$\zeta_{RT}(k) = \begin{cases} 1 + \frac{D^*}{D^* - F(k)}, & \text{if } 0 \leq F(k) < D^*, \\ 0, & \text{if } F(k) \geq D^*, \end{cases} \quad (4.6)$$

and

$$\zeta_{NRT}(k) = \begin{cases} 1 + \frac{D^*}{F^*(k) - F(k)}, & \text{if } 0 \leq F(k) < F^*, \\ 1 + D^*, & \text{if } F(k) \geq F^*, \end{cases} \quad (4.7)$$

where  $F(k)$  is the integer number of frames that the HOL packet of user  $k$  has experienced at the beginning of the present frame, and  $F^*(k)$  is the maximum number of delay (unit of frames) that the HOL packet of user  $k$  with  $NRT$  service can have so as to achieve its minimum rate requirement  $R_{\min}^*$ . Note that  $F^*(k)$  can be regarded as a requirement value in unit of time equivalent to  $R_{\min}^*$  in unit of bit rate. The numerators of fractions in (4.6) and (4.7) are set to be the same, making maximum values of  $\zeta_{RT}(k)$  and  $\zeta_{NRT}(k)$  equal to  $1 + D^*$ ; the denominators in (4.6) and (4.7) are the differences just presenting the tolerance time for the  $RT$  and  $NRT$  services, respectively. For the  $RT$  service, if  $F(k)$  is not less than  $D^*$ , the  $RT$  packet will be dropped, the resource scheduling is not considered, and thus let  $\zeta_{RT}(k) = 0$ . For  $NRT$  service, the  $F^*$  can be expressed as

$$F^*(k) = \left\lceil \frac{L_{NRT}(k)}{R_{\min}^*} \right\rceil + B(k), \quad (4.8)$$

where  $L_{NRT}(k)$  is the payload size of the *NRT* packet of the user  $k$ ,  $B(k)$  is a balance factor of the user  $k$ , and  $\lceil x \rceil$  denotes the smallest integer larger than  $x$ . The balance factor  $B(k)$  is used to achieve a goal of that user  $k$  with *NRT* service could be scheduled with an overall *average* rate close to  $R_{\min}^*$ . The  $B(k)$  is adaptively adjusted packet by packet and is set as

$$B(k) = \max \{ F_p^*(k) - F_p(k), 0 \}, \quad (4.9)$$

where  $F_p^*(k)$  ( $F_p(k)$ ) is the  $F^*(k)$  ( $F(k)$ ) of the last *NRT* packet of user  $k$ . Eq. (4.9) denotes that if the transmission time of the last *NRT* packet of the user  $k$  is less than  $F_p^*(k)$ , then  $B(k)$  will be a positive integer to make the current *NRT* packet have longer time to be transmitted but the minimum rate requirement is still satisfied.

The BS provides one individual FIFO queue for each service class of every active user. The packet of *RT* services will be dropped if the delay of the packet exceeds the  $D^*$ . The packet of *NRT* services or *BE* services are allowed to be queued and delayed without being dropped if its buffer occupancy is not overflowed. Retransmission due to erroneous transmission of packets is here not considered.

### 4.3 Utility-based TMCR Scheduling Scheme

The utility-based throughput maximization and complexity reduction (U\_TMCR) scheme first designs a utility function for each user, then formulates the scheduling problem into optimization equations of overall system utility maximization, and finally solves the optimization equations by a heuristic TMCR algorithm.

#### 4.3.1 Utility Function

Denote  $U(k, n, r)$  the individual utility function of user  $k$  on subchannel  $n$  using receiving antenna  $r$  at OFDMA symbols of a frame, where  $1 \leq k \leq K$ ,  $1 \leq n \leq N$ , and

$1 \leq r \leq Q_k$ . The  $U(k, n, r)$  is defined as a *radio resource function* of user  $k$  on subchannel  $n$  at the  $l$ th OFDMA symbol using receiving antenna  $r$ , denoted by  $R(k, n, r)$ , weighted by its *QoS monitoring function* at the same frame, denoted by  $M(k)$ . It is given by

$$U(k, n, r) = R(k, n, r) \times M(k). \quad (4.10)$$

The radio resource function  $R(k, n, r)$  indicates the number of bits eligibly carried for user  $k$  on the subchannel  $n$  with receiving antenna  $r$  per symbol at a frame. The  $R(k, n, r)$  has been obtained in [60-(20)] by

$$R(k, n, r) = n_s \times \log_2(1 - SNR_{k,r} \times 1.5 / \ln(5 \times BER^*)), \quad (4.11)$$

where  $n_s$  is the number of subcarriers in a subchannel. The  $SNR_{k,r}$  has been given in (4.5) such that the orthogonality among selected users' channels is obtained. A larger  $R(k, n, r)$  denotes user  $k$  having a better channel quality, and more radio resource should be allocated to the user to achieve a higher system throughput. The larger the  $R(k, n, r)$  is, the higher the system throughput can be achieved.

The QoS monitoring function  $M(k)$  is defined as

$$M(k) = \begin{cases} \gamma_{RT} \times \zeta_{RT}(k), & \text{if user } k \text{ is with } RT \text{ service,} \\ \gamma_{NRT} \times \zeta_{NRT}(k), & \text{if user } k \text{ is with } NRT \text{ service,} \\ \gamma_{BE}, & \text{if user } k \text{ is with } BE \text{ service,} \end{cases} \quad (4.12)$$

where  $\gamma_{RT}$ ,  $\gamma_{NRT}$ , and  $\gamma_{BE}$  are default priority constants for *RT*, *NRT*, and *BE* services, respectively, and  $\zeta_{RT}(k)$  ( $\zeta_{NRT}(k)$ ) is the urgency value of user  $k$  with *RT* (*NRT*) service, which has been given in Eq. (4.6) (Eq. (4.7)). It would be obvious that  $\gamma_{RT} > \gamma_{NRT} > \gamma_{BE}$  since the *RT* (*BE*) service is the most (least) time-constraint service. The larger the  $M(k)$  is, the more urgently the user  $k$  will be served. This will make QoS requirement guaranteed.

### 4.3.2 Problem Formulation

Assume that the allocation matrix  $\mathbf{A}$  is given by  $\mathbf{A} = [A_1^T, \dots, A_l^T, \dots, A_L^T]$ . The  $A_l$  is the allocation vector at the  $l$ th OFDMA symbol denoted by  $A_l = [\delta_l(1,1,1), \dots, \delta_l(1,1,Q_1), \dots, \delta_l(k,n,1), \dots, \delta_l(k,n,r), \dots, \delta_l(k,n,Q_k), \dots, \delta_l(1,N,1), \dots, \delta_l(K,N,Q_K)]$ , where  $\delta_l(k,n,r)$ ,  $1 \leq k \leq K$ ,  $1 \leq n \leq N$ ,  $1 \leq r \leq Q_k$ , is an allocation indicator given by

$$\delta_l(k,n,r) = \begin{cases} 1, & \text{if subchannel } n \text{ is allocated to user } k \text{ with receiving antenna } r, \\ 0, & \text{otherwise.} \end{cases} \quad (4.13)$$

The U\_TMCR scheme formulates the scheduling problem of the utility-based radio resource allocation in a frame for downlink MIMO-OFDMA system as

$$\mathbf{A}^* = \arg \max_{\mathbf{A}} \left\{ \sum_{k=1}^K \sum_{n=1}^N \sum_{l=1}^L \sum_{r=1}^{Q_k} [U(k,n,r) \times \delta_l(k,n,r)] \right\}, \quad (4.14)$$

subject to system constraints

$$\sum_{k=1}^K \sum_{r=1}^{Q_k} \delta_l(k,n,r) \leq Q, \forall n, \quad (4.15)$$

$$\sum_{l=1}^L \delta_l(k,n,r) \leq \left[ \frac{\xi_{HOL}^k}{\sum_{n=1}^N \sum_{r=1}^{Q_k} R(k,n,r) \times \delta_l(k,n,r)} \right], \forall k, \quad (4.16)$$

where the term inside the bracket of (4.14) is the overall system utility, and  $\xi_{HOL}^k$  in (4.16) is the residual bits of the HOL packet of user  $k$ . The constraint of (4.15) means that each subchannel can be allocated to  $Q$  users at most for the same symbol time; and the constraint of (4.16) implies the bits allocated to each user cannot be larger than the residual bits of its HOL packet to avoid wasting resource. Notice that maximizing the overall system utility value in (4.14) can achieve the high system throughput and QoS

guaranteed.

### 4.3.3 Heuristic TMCR Algorithm

Finally, the U\_TMCR scheme contains a heuristic TMCR algorithm to solve the radio resource scheduling problem given in (4.14)-(4.16). Because the optimal method by exhaustive search [41] to find an optimal solution takes too much time, which is infeasible to realize, the heuristic TMCR algorithm is proposed to maximize throughput and to reduce complexity. It can also conform to the allocation map structure defined in the standard [61]. Figure 4.2 shows the flow chart of the heuristic TMCR algorithm. The heuristic TMCR algorithm mainly contains two functions: maximum utility allocation (MUA) and consistent allocation (CSA). It finds an allocation map of a frame, allocates the radio resource from the first symbol of each subchannel, and schedules the allocation until there is no unallocated OFDMA symbol in the frame. At the initialization, the heuristic TMCR algorithm calculates the individual utility value  $U(k, n, r)$  given in (4.10) by using the MAS scheme [56], [57], where  $1 \leq k \leq K$ ,  $1 \leq n \leq N$ , and  $1 \leq r \leq Q_k$ . Then, it uses the MUA function to schedule the allocation according to the individual utility value. It also uses the CSA function to perform continuous allocation to the user, determined by the MUA function, by an appropriate number of symbols for QoS fulfillment. In the heuristic TMCR algorithm, a possible allocation of a subchannel and a receiving antenna to a user is named as an allocation trial; the process to find the right channel and the right receiving antenna to the right user is called an allocation iteration. Notice that the heuristic TMCR algorithm is a greedy method, and the result of the greedy method is close to the optimal solution when the number of user is large [41]. Furthermore, its computational complexity is much smaller than that of the exhaustive search.

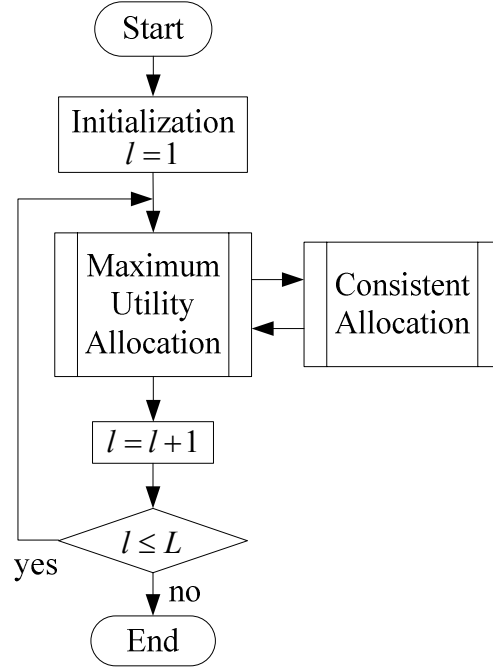


Figure 4.2. Flow chart of the heuristic TMCR algorithm

### [MUA Function]

The maximum utility allocation (MUA) function is to find an optimal allocation indicator  $\delta_l(k, n, r)$ , which can have the largest individual utility value for the  $l$ th OFDMA symbol. The MUA function finds the highest individual utility value, denoted by  $U(k^*, n^*, r^*)$ , and then assigns to the user  $k^*$  with the receiving antenna  $r^*$  on the subchannel  $n^*$  and checks whether the number of allocated users equal to  $Q$  (comply with the constraint of (4.15)). When the subchannel is assigned, the next allocated user cannot interfere the previously assigned user. The MUA adopts the MAS scheme to guarantee the orthogonality among mobile users and to recalculate the radio resource function  $R(k, n^*, r)$  of the subchannel. Afterwards, the MUA function will call the CSA function to accomplish the resource allocation for the user  $k^*$ . Repeat the next highest individual utility value user allocation and QoS requirement fulfillment until there is no unallocated user or no free subchannel in the  $l$ th OFDMA symbol. The

pseudocode of the MUA function is given below.

**Function: [MUA]**

Step 1: Find  $(k^*, n^*, r^*) = \arg \max_{k \in \Omega, n \in S_{free}^l, r \in \Psi_{n,k}^l} (U(k, n, r))$

Step 2: Set  $\delta_l(k^*, n^*, r^*) = 1$ ,  $S_{k^*} = S_{k^*} \cup \{n^*\}$ ,  $\Psi_{n,k}^l = \Psi_{n,k}^l \setminus \{r^*\}$

if  $\sum_{k=1}^K \sum_{r=1}^{Q_k} \delta_l(k, n^*, r) = Q$

$S_{free}^l = S_{free}^l \setminus \{n^*\}$

else recalculate  $R(k, n^*, r)$  for all  $k \in \Omega, r \in \Psi_{n^*,k}^l$

Step 3: Call the CSA function

Step 4: Find  $(k^*, n^*, r^*) = \arg \max_{k=k^*, n \in S_{free}^l, r \in \Psi_{n,k^*}^l} (U(k, n, r))$ , and go to step 2

Step 5: Repeat Step 1 - Step 3, if  $\{S_{free}^l\} \neq 0$  and  $\{\Omega\} \neq 0$  ■

Note that, in the above pseudocode,  $S_{free}^l$  is the set of available subchannels which can be allocated at the  $l$ th OFDMA symbol,  $\Omega$  is the set of unallocated users,  $\Psi_{n,k}^l$  is the unallocated antenna of user  $k$  on subchannel  $n$  at the  $l$ th OFDMA symbol, and  $S_k$  is the set of subchannels that are allocated to user  $k$ .

**[CSA Function]**

The consistent allocation (CSA) function is to perform continuous allocation to fulfill the QoS requirement and to reduce the computational complexity, besides the allocation map structure defined by the specification [61] conformed. In order to satisfy the user's QoS requirement, the CSA function first determines an appropriate number of transmission symbols required for the allocated user selected by the MUA function and then performs the consistent allocation for the user. Assume that the selected user is user  $k$ . The required number of transmission symbols for user  $k$ , denoted by  $\tau_k$ , is according to its degree of urgency  $\zeta(k)$ , which was given in (4.6) ((4.7)) if the user  $k$  is with  $RT$  ( $NRT$ ) service. The more urgent user would be allocated more transmission symbols to



make service fulfill its QoS requirement as much as possible. The  $\tau_k$  is designed as

$$\tau_k = \begin{cases} 1 & \text{if } 1 < \zeta(k) \leq \frac{\zeta_{\max}(k)}{4}, \\ \left\lfloor \frac{\tau_{k, \max}}{4} \right\rfloor & \text{if } \frac{\zeta_{\max}(k)}{4} < \zeta(k) \leq \frac{\zeta_{\max}(k)}{2}, \\ \left\lfloor \frac{2\tau_{k, \max}}{3} \right\rfloor & \text{if } \frac{\zeta_{\max}(k)}{2} < \zeta(k) \leq \frac{3\zeta_{\max}(k)}{4}, \\ \tau_{k, \max} & \text{if } \frac{3\zeta_{\max}(k)}{4} < \zeta(k) \leq \zeta_{\max}(k), \end{cases} \quad (4.17)$$

where  $\zeta_{\max}(k)$  is the maximum value of  $\zeta(k)$ , and  $\tau_{k, \max}$  is the maximum transmission symbols required by the current packet of user  $k$ . The  $\zeta_{\max}(k) = 1 + D^*$  and  $\tau_{k, \max}$  can be expressed as

$$\tau_{k, \max} = \left\lceil \frac{\zeta_{HOL}^k}{\sum_{n \in S_k} \sum_{r=1}^{Q_k} R(k, n, r) \times \delta(k, n, r)} \right\rceil. \quad (4.18)$$

If the unallocated OFDMA symbols of the allocated subchannel is smaller than the  $\tau_k$ , the CSA function will go to Step 4 of the MUA function. The MUA function will assign another new subchannel to the user according to the utility value, add the subchannel to the  $S_k$ , and recalculate  $\tau_k$  based on (4.17) and (4.18). Repeat above steps until  $\tau_k$  is smaller than unallocated OFDMA symbols of the allocated subchannel or there is no any free subchannel. Then, the CSA function generalizes the allocation to the same user in the following consecutive  $\tau_k$  (remaining) unallocated OFDMA symbols, and removes the user  $k$  from the unallocated users. Thus the constraint of (4.16) can be observed and the computation complexity of the U\_TMCR can be reduced. The pseudocode of the CSA function is given below.

**Function: [CSA]**Step 1: Calculate  $\tau_{k^*}$ Step 2: if  $\tau_{k^*} > L - l + 1$ if  $\{S_{free}^l\} \neq 0$ 

Go to Step 4 of the MUA function

else

Set  $\delta_{l+1}(k^*, n, r^*) = \dots = \delta_L(k^*, n, r^*) = 1, \forall n \in S_{k^*}$  $\Omega = \Omega \setminus \{k^*\}$ 

Go to Step 5 of the MUA function

else

Set  $\delta_{l+1}(k^*, n, r^*) = \dots = \delta_{l+\tau_{k^*}-1}(k^*, n, r^*) = 1, \forall n \in S_{k^*}$  $\Omega = \Omega \setminus \{k^*\}$ 

Go to Step 5 of the MUA function ■

Note that the CSA function extends the allocation result spreading over following consecutive OFDMA symbols, and this can reduce the system overhead for the U\_TMCR scheme. The reason is that the CSA function makes the user using the same subchannel with the same modulation order on several continuous OFDMA symbols in a frame. Therefore, the allocation map just needs to record information which includes subchannel number and transmission period. For conventional allocation schemes, a user may be allocated on different subchannels in adjacency OFDMA symbols and this needs a large number of overhead bits to record it. Note that the TMCR algorithm immediately uses the CSA function to generalize the allocation result after the MUA function finds an allocation indicator. Therefore, the urgent user has high probability to have long length for generalization and easily fulfill the QoS requirement.

## 4.4 Simulation Results

In the simulations, the system level parameters of the downlink MIMO-OFDMA

system are set to be compatible with the IEEE 802.16 standard [42], and the scalable parameters are configured according to the suggested values in [62] and listed in Table 4.1. The path loss model is modeled as  $128.1+37.6 \log(R)$  dB, where  $R$  is the distance between the base station and the user in kilometers [63]. The log-normal shadowing is assumed with zero mean and standard deviation of 8 dB. The QoS requirements of each service are listed in Table 4.2.

Table 4.1 System-Level Parameters

Parameters	Value
Cell size	1.6 km
Frame duration	5 ms
System bandwidth	5 MHz
FFT size	512
Subcarrier frequency spacing	10.9375 KHz
Number of data subcarriers	384
Number of subchannels	8
Number of receiving antennas	2
Number of transmitting antennas	2
Number of data subchannel	8
Number of data subcarriers per subchannel	48
Number of slots for downlink transmission per frame	24
Maximum transmission power BS	43 dBm
Thermal noise density	-174 dBm/Hz

Table 4.2 QoS requirements of each service

Requirement \ Service	voice ( $RT$ )	Video ( $RT$ )	HTTP ( $NRT$ )	FTP ( $BE$ )
Required BER ( $BER^*$ )	$10^{-3}$	$10^{-4}$	$10^{-6}$	$10^{-6}$
Maximum Packet Delay Tolerance ( $D^*$ )	40ms	10ms	N/A	N/A
Maximum Packet Dropping Ratio ( $P_p^*$ )	1%	1%	N/A	N/A
Minimum Required Transmission Rate ( $R_{min}^*$ )	N/A	N/A	100kbps	N/A

N/A: Not Applicable

In the simulations, it is assumed that each user is assumed to have one service type and the traffic intensity is defined as the ratio of the total average arrival data rate of all service types of all users over the maximum system transmission rate. Besides, the average arrival data rates of voice, video, HTTP, and FTP are 4.8 Kbps, 64 Kbps, 14.5 Kbps, and 88.9 Kbps, respectively. Thus, the traffic intensity varies from 0.15 to 0.90 as the number of users varies from 80 to 480.

The voice service is modeled as the ON-OFF model, in which lengths of ON (OFF) period follow an exponential distribution with means 1.0 (1.5) seconds [63]. The video data is assumed to arrive at a regular interval of 100ms, each frame is decomposed into eight slices (packets), and the size of a packet is distributed in a truncated Pareto distribution [64]. In which, there are delay intervals between two consecutive packets of a frame which denote the encoding delay at the video encoder. These intervals are modeled by a truncated Pareto distribution. The HTTP of NRT service is modeled as the behavior of web browsing. Thus, the HTTP traffic is modeled as a sequence of page downloads, and each page download is modeled as a sequence of packet arrivals. The interval between two consecutive page downloads, representing the reading time in web browsing, is distributed in an exponential distribution. For detailed parameters of video and HTTP traffic models please refer to [64]. The FTP traffic of BE user is modeled as a sequence of file downloads. The size of a file is distributed in a truncated lognormal distribution with mean 2M bytes, standard deviation 0.722M bytes, and a maximum value 5M bytes. In addition, the interval between files is distributed in an exponential distribution with mean 180 seconds.

In the following, the proposed U\_TMCR scheme is compared to the exhaustive search method [41], the CDPS [30] scheme and the ARRA [32] scheme for performance analysis. The exhaustive search method is the one which exhaustively search the optimal

solution of the optimization equations given in (4.14)-(4.16). The CDPS scheme and the ARRA scheme are modified to fulfill the MIMO OFDMA system.

Figure 4.3 shows the system throughput versus the traffic intensity for the proposed U\_TMCR scheme, the exhaustive search method, the CDPS scheme in [30], and the ARRA scheme in [32]. It can be seen that the system throughput of the proposed U\_TMCR scheme is very close to that of the exhaustive search method; the former is less than the latter by only an amount of 2.12% at traffic intensity 0.75. The reason is that the heuristic U\_TMCR is a kind of greedy method [41]. Also, at high traffic intensity, the proposed U\_TMCR scheme can achieve system throughput higher than the ARRA scheme and the CDPS scheme by an amount of 8% and 21% at traffic intensity 0.75, respectively. It is because the U\_TMCR scheme allocates the resource according to the utility function given in (4.10), where a user with a high individual utility value is the one who is urgent and with the good channel quality; whereas, the services order of users in the ARRA scheme is only based on the urgency which denotes the residual time to expiration. This would make the U\_TMCR scheme attain higher system throughput than the ARRA scheme. Also, the U\_TMCR scheme designs the maximum urgency values of *RT* and *NRT* packets to be the same but set the priority constants in the QoS monitoring function to further differentiate the services so that the urgent *RT* packet will have higher probability to be served than the urgent *NRT* packet. The ARRA scheme adaptively adjusts the priority of users according to just the time to expiration. There are possibilities that the priority value of *NRT* packet may be larger than that of the *RT* packet. Thus, some packets of *RT* traffic will be squeezed and not be served in time at high traffic intensity. These *RT* packets will be dropped, resulting in the system throughput decrement for the ARRA scheme. On the other hand, the CDPS scheme serves users by the fixed priority, and thus the multiuser gain cannot be obviously

appeared. Additionally, when the traffic intensity increases, the probability of serving video packets would become small and the video packet dropping rate intuitively increases. This makes the system throughput of the CDPS scheme be the smallest.

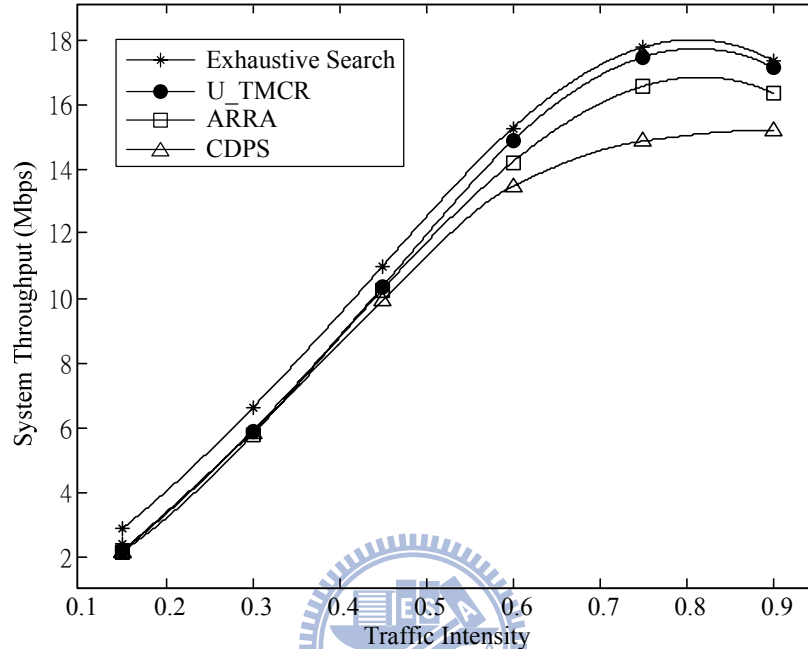


Figure 4.3 System Throughput

Figures 4.4(a) and 4.4(b) show voice and video packet dropping rates, respectively, where the dropping rate requirements (1%) are also given. It can be seen that the U\_TMCR performs quite close to the exhaustive search solution in the performance measures of the two *RT* packet dropping rates. The U\_TMCR (ARRA) scheme keeps the voice packet dropping rate lower than the voice packet dropping probability requirement of 1% until the traffic intensity is above 0.9 (0.8), while the CDPS scheme always attains the voice packet dropping rates close to zero. Also, the U\_TMCR scheme keeps the video packet dropping rates close to zero. Also, the U\_TMCR scheme keeps the video packet dropping rate lower than the video packet dropping probability requirement of 1% until the traffic intensity is near 1.0, while that of the CDPS (ARRA) scheme violates the requirement at traffic intensity 0.56 (0.76). The reasons are quite similar to what we give for the system throughput comparison among the U\_TMCR, the ARRA,

and CDPS schemes in Fig. 4.3.

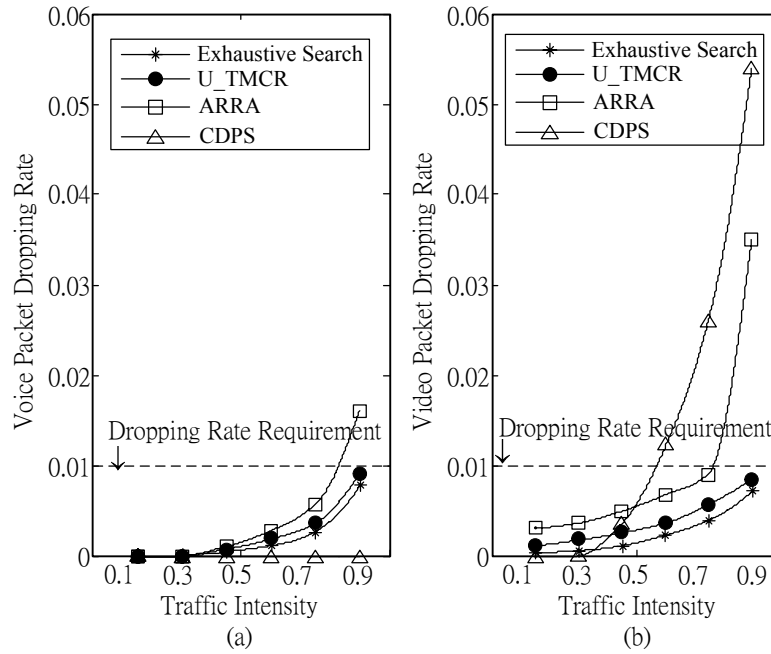


Figure 4.4 (a) Voice packet dropping rate; (b) Video packet dropping rate

Figs. 4.5(a) and 5(b) show the mean voice and video packet delays, respectively, where the maximum packet delay requirements ( $D^*$ ) are also included. It can be seen that the mean  $RT$  packet delays by the U\_TMCR scheme are larger than that by the exhaustive search method. The CDPS scheme employs the fixed priority, thus it has the smallest voice packet delay but the largest video packet delay. Both of the U\_TMCR scheme (exhaustive search method) and the ARRA scheme exploit the dynamic priority and make full use of the voice packet delay tolerance. The three schemes delay the voice packets to serve more video packets in time. Thus, the delays of video packet of these three schemes are smaller than that of the CDPS scheme. It can also be found that the voice and video packet delay by the U\_TMCR scheme are smaller than by the ARRA scheme at traffic intensity over 0.5. The reason is that the *urgent* packet of  $RT$  user in the U\_TMCR scheme has the highest priority while the priority of *urgent* packet of  $RT$  user in the ARRA scheme may be smaller than that of  $NRT$  packet; the *urgent*  $RT$  users in the

U\_TMCR scheme have higher probability to be served in time.

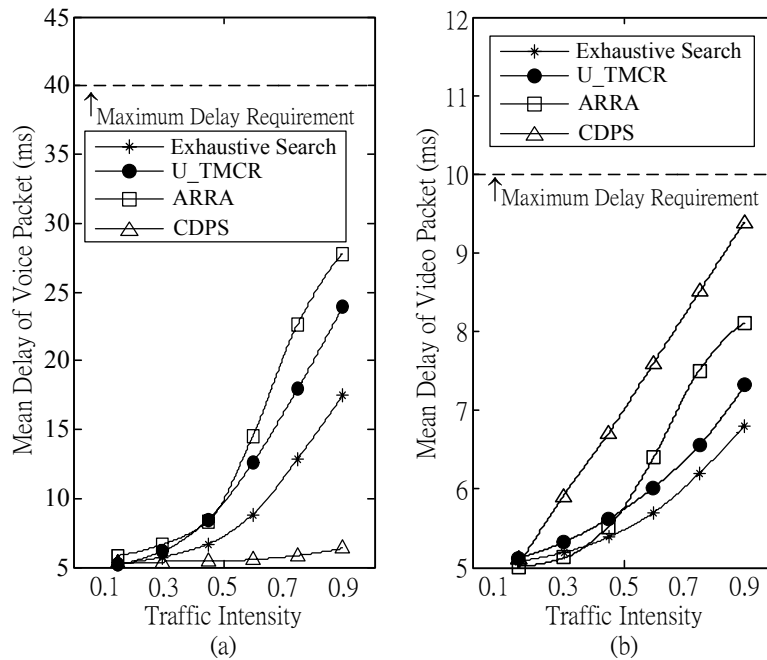


Figure 4.5 (a) Mean delay of voice packet; (b) Mean delay of video packet

Figure 4.6 illustrates the guaranteed ratio for HTTP packets. Unlike the *RT* traffic, packets of the HTTP traffic will not be dropped but still wait for service when the requirement of minimum transmission rate  $R_{\min}^*$  cannot be kept. It can be observed that the guaranteed ratio of the U\_TMCR scheme is close to that of the exhaustive search method, and it is much higher than that of the CDPS scheme but lower than that of the ARRA scheme. As the same reason given above, by the CDPS scheme, the priority of HTTP packets is the third one which means that HTTP packets have to wait until all real time packets are served. By the U\_TMCR scheme (exhaustive search method) and the ARRA scheme, since the priority of users are dynamically adjusted frame by frame, it can avoid more resource being always occupied by *RT* traffic. Thus more HTTP packets can be guaranteed to be served with the minimum transmission rate. Additionally, the ARRA scheme further gives the *NRT* service to override the *RT* service to be first served more possibly than the U\_TMCR scheme when the requirement of minimum



transmission rate  $R_{\min}^*$  is going to be violated. Thus, the ARRA scheme has the largest guaranteed ratio for HTTP packets.

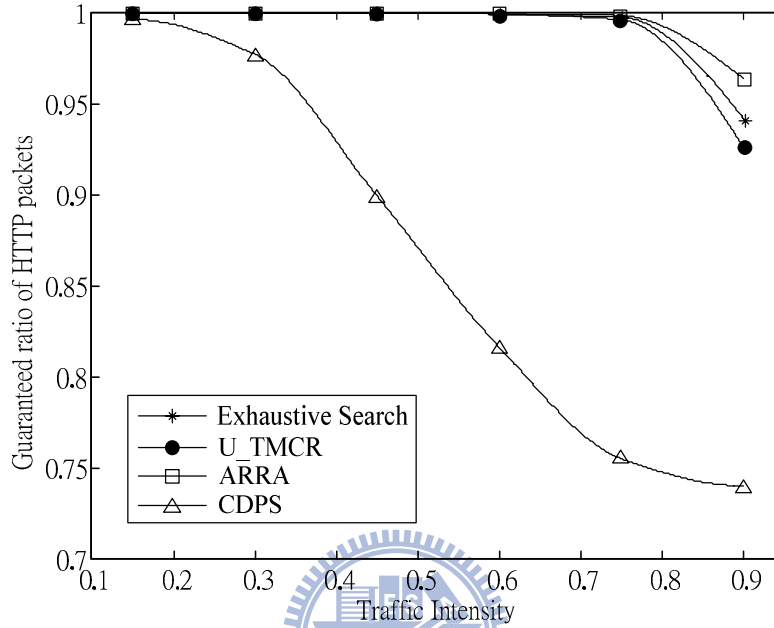


Figure 4.6 Guaranteed ratio of HTTP packets

Figure 4.7 shows the  $BE$  throughput versus the traffic intensity. It can be found that the U\_TMCR scheme has  $BE$  throughput lower than the exhaustive search method, as the system throughput shown in Fig.3.3; the U\_TMCR scheme (exhaustive search method) and the ARRA scheme outperform the CDPS scheme in the  $BE$  throughput. The  $BE$  service in the U\_TMCR scheme (exhaustive search method) still has more chances to transmit. The reason is that the design of utility function in (4.10) can achieve higher multiuser gain and make the resource allocation more efficiently. Furthermore, it could make the  $RT$  users be delayed near the delay bound and the  $NRT$  users with good channel quality be served with high probability. It can also be found that the  $BE$  throughput of the U\_TMCR scheme (exhaustive search method) increases until traffic intensity is 0.6 and then decreases. It is because the U\_TMCR scheme (exhaustive

search method) will give the *RT* services more resource in order to satisfy their QoS requirements when traffic intensity is higher, which can be seen in Fig. 4.4. For the CDPS scheme, the FTP traffic will be transmitted only when voice, video, and HTTP traffic have already been served. Thus its average throughput of *BE* is the lowest.

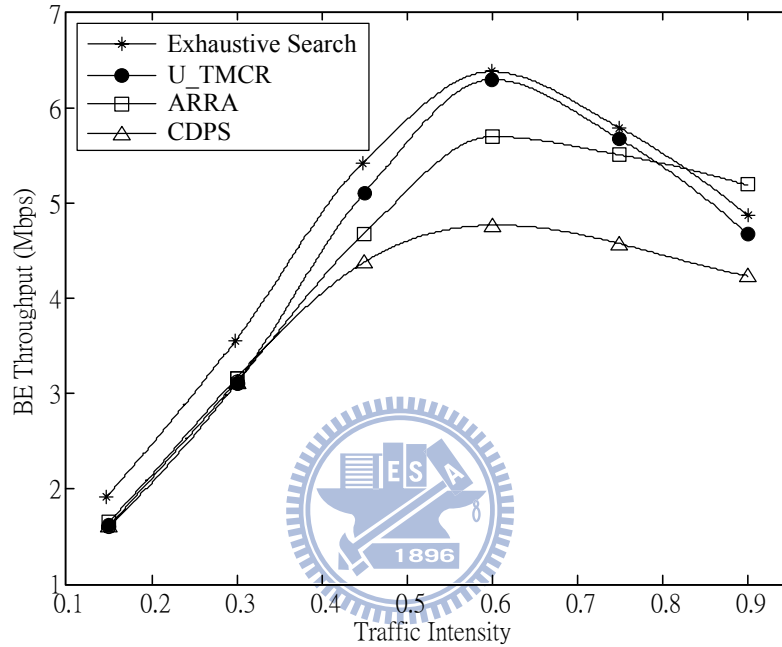


Figure 4.7 *BE* throughput

Finally, the computational complexities among the exhaustive search method, the U\_TMCR, ARRA, and CDPS schemes are compared. The computation complexity of a scheme depends on the number of allocation trials in an allocation iteration and the total number of allocation iterations needed for the scheme. Theoretically, the worst-case computational complexity for the four schemes would be  $O(LKN^2Q_M)$ , where  $Q_M$  is maximal number of receiving antenna  $Q_k$ . However, either the consistent allocation (CSA) function in the U\_TMCR scheme or the generalization function in the ARRA scheme continue the allocation to the same user in several following OFDMA symbols. Thus the complexities of the U\_TMCR and ARRA schemes will be greatly reduced by at

least  $L$  times over the exhaustive search method and the CDPS scheme.

However, the U\_TMCR scheme has lower computational complexity than the ARRA scheme in realistic operations. The U\_TMCR scheme uses the CSA function to generalize the allocation result immediately after the MUA function has allocated a subchannel to a user. Thus, if a user needs more than one subchannels to transmit, the U\_TMCR scheme performs the allocation for the user in just one iteration and removes the user from the unallocated user set. On the other hand, the ARRA scheme extends the results after all subchannels of an OFDMA symbol are allocated so that the ARRA scheme needs the number of allocation iteration equal to the number of required subchannels of the user and it removes the user until the QoS requirements of the user are satisfied. Therefore, the U\_TMCR scheme needs a smaller number of allocation iteration to allocate the resource and a smaller number of allocation trials in the following allocation iteration than the ARRA scheme. Take one example. Assume that users are with packet length uniformly distributed between 72 bytes and 576 bytes. The number of allocation trials in a frame by the U\_TMCR scheme is reduced by 6.25% ~ 29.2%, compared with the ARRA scheme. The larger the packet length is, the reduction of computation complexity by the U\_TMCR scheme would be.

## 4.5 Concluding Remarks

In this chapter, a utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling scheme is proposed for downlink multiuser MIMO-OFDMA systems, where the radio resource allocation to multimedia users includes subchannel allocation, modulation order assignment and receiving antenna. The goals of the U\_TMCR scheme are to maximize system throughput, fulfill QoS requirements, and lessen the computational complexity, while considering multiple service classes, such as

*RT*, *NRT*, and *BE* services. The proposed U\_TMCR scheme designs a utility function, formulates the utility-based scheduling problem, and solves the problem by a heuristic TMCR algorithm. For *RT* (*NRT*) service, the value of the utility function is dynamically adjusted to maximize the spectrum efficiency and to fulfill the delay requirement (minimum transmission rate) and the *BER* requirement. The heuristic TMCR algorithm includes a maximum utility allocation (MUA) function to maximize the overall utility value and a consistent allocation (CSA) function to fulfill QoS requirements of users and reduce the computational complexity.

Simulation results show that the performance of the proposed U\_TMCR scheme is very close to that of the exhaustive search method, and the proposed U\_TMCR scheme outperforms the conventional ARRA scheme [32] and the CDPS scheme [30] in system throughput by an amount of 8% and 21.5%, respectively. The U\_TMCR scheme can sustain users' QoS requirement up to the traffic intensity 0.9, while the ARRA (CDPS) scheme can not guarantee QoS requirements after a traffic intensity of 0.8 (0.55). The overall QoS satisfaction by the U\_TMCR scheme is higher than that by the ARRA and the CDPS schemes. This is because the U\_TMCR scheme adjusts the individual utility value of every user by considering users' service priority, urgency degree, and channel quality, frame by frame. Also, the U\_TMCR scheme gives *RT* (*NRT*) users with larger packet delay (lower average transmission rate) a higher priority to quickly obtain the resource and clearly differentiates the *RT* and *NRT* users. Moreover, the U\_TMCR scheme can reduce the number of allocation iterations. The total number of the allocation trials in a frame by the U\_TMCR scheme is reduced by 6.25% ~ 29.2%, compared to the ARRA scheme.

# Chapter 5

## Conclusions and Future Works

---

In this dissertation, the radio resource allocation schemes in future wireless network are studied, including an adaptive  $P$ -persistent MAC scheme for multimedia service in PAN to provide delay fairness, a utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling scheme for downlink WMAN to maximize system throughput under QoS guarantee and reduce computational complexity, and a dynamic priority resource allocation (DPRA) scheme for uplink WMAN to maximize the system throughput and satisfy differentiated QoS requirements.

In Chapter 2, an adaptive p-persistent (APP) MAC scheme for IEEE 802.11 WLAN achieving fairness in the sense of low delay variance is proposed and analyzed. By adaptively determining the permission probability of station according to the state of packet transmission of the station, the APP MAC scheme resolves the fairness problem at each access of stations. It differentiates the permission probabilities of stations with various waiting delays, and assigns a higher probability to stations with larger packet delay. The collision probability, the system throughput, and the mean delay of the APP MAC scheme are successfully obtained by Markov-chain model. Results show that the analyses are quite correct and the discrepancy between the numerical results and the

simulation results is very small.

For multimedia environment, by setting different initial permission probabilities, the APP MAC scheme can differentiate stations with various AC of services in multimedia WLAN. According to its transmission state, the APP MAC scheme dynamically determines the permission probability of station in the same AC to reduce the delay variance of station. Simulation results show that the APP MAC scheme can enhance the performance of multimedia WLAN; it effectively improves the capacity of high priority stations, reduces the mean delay, enhances the mean throughput, and achieves lower delay variance, compared to conventional algorithms.

The initial permission probability  $P_0$  is an important design parameter in the APP MAC scheme. It can be determined by considering the system design objective which is to reduce the delay variance or enhance the system throughput. Besides, the initial permission probability  $P_0$  can be adaptively determined according to the system load. For example,  $P_0$  could be set to be  $1/16$  ( $1/2$ ) when the system is in heavy (light) load. In realistic implementation, the number of rebackoffs and the number of retransmissions are statistical data recorded by station. The current  $CW$  of a station can indicate retransmission, thus only a register is needed in the station to store the value of rebackoff. Also, the value of  $P_0$  ( $RB_{max}$ ) for an AC would be set larger (smaller) if the AC is with more delay sensitive service, for the configuration of WLAN MAC.

In Chapter 3, the radio resource allocation to multimedia users for downlink multiuser MIMO-OFDMA systems is investigated and a utility-based throughput maximization and complexity reduction (U\_TMCR) scheduling scheme is proposed, where the U\_TMCR scheme includes subchannel allocation, modulation order assignment and receiving antenna selection. While considering multiple service classes,

the goals of the U\_TMCR scheme are to maximize system throughput, fulfill QoS requirements, and lessen the computational complexity. The proposed U\_TMCR scheme designs a utility function, formulates the utility-based scheduling problem, and solves the problem by a heuristic TMCR algorithm. For *RT* (*NRT*) service, the value of the utility function is dynamically adjusted to maximize the spectrum efficiency and to fulfill the delay requirement (minimum transmission rate) and the *BER* requirement. The heuristic TMCR algorithm includes a maximum utility allocation (MUA) function to maximize the overall utility value and a consistent allocation (CSA) function to fulfill QoS requirements of users and reduce the computational complexity.

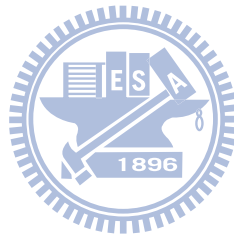
Simulation results show that the performance of the proposed U\_TMCR scheme is very close to that of the exhaustive search method, and the proposed U\_TMCR scheme outperforms the conventional ARRA scheme [32] and the CDPS scheme [30] in system throughput by an amount of 8% and 21.5%, respectively. The U\_TMCR scheme can sustain users' QoS requirements up to the traffic intensity 0.9, while the ARRA (CDPS) scheme can not guarantee QoS requirements after a traffic intensity of 0.8 (0.55). The overall QoS satisfaction by the U\_TMCR scheme is higher than that by the ARRA and the CDPS schemes. This is because the U\_TMCR scheme adjusts the individual utility value of every user by considering users' service priority, urgency degree, and channel quality, frame by frame. Also, the U\_TMCR scheme gives *RT* (*NRT*) users with larger packet delay (lower average transmission rate) a higher priority to quickly obtain the resource and clearly differentiates the *RT* and *NRT* users. Moreover, the U\_TMCR scheme can reduce the number of allocation iterations. The total number of the allocation trials in a frame by the U\_TMCR scheme is reduced by 6.25% ~ 29.2%, compared to the ARRA scheme.

In Chapter 4, the consistent allocation, performed by a dynamic priority resource allocation (DPRA) scheme, for IEEE 802.16 uplink system with multimedia traffic is proposed. The intention of the DPRA scheme is to fulfill QoS requirements and maximize the system throughput. Originally, a priority value is designed for each service type according to the urgency and QoS requirements of the traffic and dynamically adjusted frame by frame. Simulation results reveal that the performance of the DPRA scheme is better than that of the conventional EFS algorithm, which performs allocation slot by slot. Besides, benefited from the consistent allocation, the computational complexity of the DPRA scheme is much less than that of the conventional EFS algorithm. Likewise, when traffic intensity is larger, the system throughput of proposed DPRA scheme is close to that of the optimal method. However, the computational complexity of the DPRA scheme is much less than that of the optimal solution. As a result, we can infer that the proposed DPRA scheme can accomplish throughput maximization and QoS satisfaction with lower computational complexity and transmission overhead.

The work called LTE Advanced (LTE-A) is on the next evolution of Long Term Evolution (LTE). Not only is the extended bandwidth support of up to 100 MHz included but also the peak data rates in excess of 1 Gbps in the downlink for low mobility [70] is offered by LTE-Advanced. Carrier aggregation, in which several LTE Rel'8 compatible component carriers are placed adjacent to each other on the same subcarrier grid [71] is the agreed method for achieving bandwidth extension in LTE-Advanced. The carrier aggregation capability and bandwidth, in different cellars, can be controlled, and this creates a new resource dimension for scheduling. By using this dimension, the scheduler can maximize the capacity of the system and reduce the



inter-cell interference. For instance, considering the multiple cell environments, every cell can choose frequency band according to the interference level and choose how many frequency bands according to the traffic load.



# Bibliography

- [1] N. Abramson, "The ALOHA system—Another alternative for computer communication," in *Proc. Fall Joint Comput. Conf., AFIPS Conf.*, 1970, pp. 281–285.
- [2] F. A. Tobagi and L. Kleinrock, "Packet switching in radio channels: Part I—Carrier sense multiple-access modes and their throughputdelay characteristics," *IEEE Trans. Commun.*, vol. COM-23, no. 12, Dec. 1975, pp. 1400–1415.
- [3] "IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band," IEEE Std 802.11b-1999, Sep. 1999.
- [4] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1992, p. 177.
- [5] Y. Kwon, Y. Fang, and H. Latchman, "A novel MAC protocol with fast collision resolution for wireless LANs," in *Proc. IEEE INCOFOM*, 2003, pp. 853–862.
- [6] G. Bianchi, L. Frata, and M. Oliveri, "Performance evaluation and enhancement of the CSMA/CA MAC protocol for 802.11 wireless LANs," in *Proc. IEEE Symp. Pers., Indoor, Mobile Radio Commun.*, 1996, pp. 392–396.
- [7] F. Cali, M. Conti, and E. Gregori, "IEEE 802.11 protocol: Design and performance evaluation of an adaptive backoff mechanism," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 9, Sep. 2000, pp. 1774–1786.
- [8] M. Natkaniec and A. R. Pach, "An analysis of the backoff mechanism used in IEEE 802.11 networks," in *Proc. IEEE Symp. Comput. and Commun.*, 2000, pp. 444–449.
- [9] Y. Kwon, Y. Fang, and H. Latchman, "Design of MAC protocols with fast collision resolution for wireless local area networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 3, May 2004, pp. 793–807.
- [10] D. Raychaudhuri and K. Joseph, "Analysis of the stability and performance of exponential backoff," in *Proc. IEEE WCNC*, 2003, vol. 3, pp. 1754–1759.
- [11] N.-O. Song, B.-J. Kwak, J. Song, and L. E. Miller, "Enhancement of IEEE 802.11

- distributed coordination function with exponential increase exponential decrease backoff algorithm,” in *Proc. IEEE VTC—Sprin*, 2003, vol. 4, pp. 2775–2778.
- [12] G. Bianchi and I. Tinnirello, “Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network,” in *Proc. IEEE INFOCOM*, 2003, pp. 844–852.
- [13] S. R. Ye and Y. C. Tseng, “A Multichain Backoff Mechanism for IEEE 802.11 WLANs,” *IEEE Trans. Veh. Technol.*, vol. 55, no. 5, Sep 2006, pp. 1613–1620.
- [14] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, “MACAW: a media access protocol for wireless LANs,” in *Proc. ACM SIGCOMM’94*, 1994, pp. 212–225.
- [15] H. Yamada, H. Morikawa, T. Aoyama, “Decentralized control mechanism suppressing delay fluctuation in wireless LANs,” *Proc. VTC 2003, Fall*, vol. 2, pp. 801–805.
- [16] F. Cali, M. Conti, and E. Gregori, “Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit,” *IEEE/ACM Trans. on Networking*, vol. 8, no. 6, Dec. 2000, pp. 785–799.
- [17] S. Yan, Y. Zhuo, S. Wu, and W. Guo, “Priority backoff algorithm for IEEE 802.11 DCF,” *International Conference on Communications, Circuits and Systems (ICCCAS)*, vol. 1, June 2004, pp. 423–427.
- [18] I. Aad and C. Castelluccia, “Differentiation mechanisms for IEEE 802.11,” *IEEE INFOCOM*, vol. 1, April 2001, pp. 209–218.
- [19] S. Choi, J. Del Prado, S. Mangold, and S. Shankar, “IEEE 802.11e contention-based channel access (EDCF) performance evaluation,” *IEEE ICC*, vol. 2, May 2003, pp. 1151–1156.
- [20] R. G. Cheng, C. J. Chang, C. Y. Shih, and Y. S. Chen, “A new scheme to achieve weighted fairness for WLAN supporting multimedia services,” *IEEE Trans. Wireless Communications*, vol. 5, no. 5, May 2006, pp. 1095–1102.
- [21] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, “Multiuser OFDM with adaptive subcarrier, bit, and power allocation,” *IEEE J. Select. Areas Commun.*, vol. 17, Oct. 1999, pp. 1747–1758.
- [22] H. Yin and L. Hui, “An efficient multiuser loading algorithm for OFDM-based broadband wireless system,” *Proc. IEEE CLOBECOM*, vol. 1, 2000, pp. 103–107.

- [23] E. Bala and L. J. Cimini, Jr. "Low-complexity and robust resource allocation strategies for adaptive OFDMA," *Proc. Vehicular Technology Conf.* vol. 1 Sept. 2005, pp. 176-180,.
- [24] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access system," *IEEE Trans. Broadcasting*, vol. 49, Dec. 2003, pp. 367-370.
- [25] Y. J. Zhang, and K. B. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for OFDM systems," *IEEE Trans. on Wireless Communications*, vol. 3, Sept. 2004, pp.1566-1575.
- [26] M. Fuchs, G. D. Galdo, and M. Haardt, "Low-complexity space-time-frequency scheduling for MIMO systems with SDMA," *IEEE Trans. on Veh. Technol.*, vol. 56, no. 5, Sept. 2007, pp. 2775-2784.
- [27] Y. Hara, L. Brunel, and K. Oshima, "Spatial scheduling with interference cancellation in multiuser MIMO systems," *IEEE Trans. on Veh. Technol.*, vol. 57, no. 2, Mar. 2008, pp. 893-905.
- [28] N. Xu, Y. Wang, and P. Zhang, "Multiuser scheduling in downlink MIMO/OFDMA System with Transmit Preprocessing," *Proc. Asia-Pacific Conf. on Commun.*, Aug. 2006, pp.1-5.
- [29] Y. Hu, C. Yin and G. Yue, "Multiuser MIMO-OFDM with adaptive antenna and subcarrier allocation," *Proc. Vehicular Technology Conf.*, vol. 6, May 2006, pp. 2873 – 2877.
- [30] J. Yu, Y. Cai, Y. Ma, D. Zhang, and Y. Xu, "A cross-layer design of packet scheduling and resource allocation for multiuser MIMO-OFDM system," *Proc. Information, Communications, and Signal processing Conf.*, Dec. 2007, pp. 1-5.
- [31] C. F. Tsai, C. J. Chang, F. C. Ren, and C. M. Yen, "Adaptive Radio Resource Allocation for Downlink OFDMA/SDMA Systems", *Proc. of IEEE ICC 2007*, pp. 5683 – 5688.
- [32] C. F. Tsai, C. J. Chang, F. C. Ren, and C. M. Yen, "Adaptive radio resource allocation for downlink OFDMA/SDMA systems with multimedia traffic," *IEEE Trans. on Wireless Communications*, Vol. 7, No. 5, May 2008, pp. 1734-1743.
- [33] C. M. Yen, C. J. Chang, F. C. Ren, and J. A. Lai, "Dynamic Priority Resource

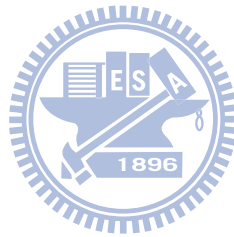
- Allocation for Uplinks in IEEE 802.16 Wireless Communication Systems,” *IEEE Transactions on Vehicular Technology*, Vol. 58, No. 8, Oct. 2009, pp. 4587-4597.
- [34] C. J. Chang, C. M. Yen, F. C. Ren, and J. A. Lai, “Dynamic Priority Resource Allocation for Uplinks in IEEE 802.16 Wireless Communication Systems,” *Proc. of IEEE VTC 2008 fall*, pp. 1 - 5.
- [35] S. Das and G. D. Mandyam, “An efficient sub-carrier and rate allocation scheme for M-QAM modulated uplink OFDMA transmission,” *Thirty seventh Asilomar Conference on Signals, Systems and Computers*, 2003, vol. 1, pp.136-140.
- [36] K. Kim, Y. Han, and S. L. Kim, “Joint subcarrier and power allocation in uplink OFDMA systems,” *IEEE Communication Letters*, vol. 9, no. 6, June 2005, pp.526-528.
- [37] P. Hosein, “Adaptive subchannel allocation in an OFDMA-based wireless network,” *2006 IFIP International Conference on Wireless and Optical Communication Networks*, pp. 1-5.
- [38] J. P. Yoon, W. J. Kim, J. Y. Baek and Y. J. Suh, "Efficient Uplink Resource Allocation for Power Saving in IEEE 802.16 OFDMA Systems" *Proc. VTC Spring 2008*, pp. 2167-2171.
- [39] V. Singh and V. Sharma, “Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks,” *Proc. IEEE Wireless Communication and Networking Conference (WCNC2006)*, vol. 2, pp. 984-990.
- [40] D. Niyato and E. Hossain, “Queue-aware uplink bandwidth allocation for polling services in 802.16 broadband wireless networks,” *Proc. IEEE GLOBECOM'05*, vol. 6, pp. 3702-3706.
- [41] K. G. Murty, *Operations Research*, Prentice Hall, 1995.
- [42] IEEE Std. 802.16-2004, “IEEE standard for local and metropolitan area networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems,” Oct. 2004.
- [43] C. J. Chang, C. M. Yen, “A Utility-based Throughput Maximization (UTM) Scheduling Scheme for Downlink MIMO-OFDMA Systems,” *Proc. of IEEE VTC 2009 Fall*.
- [44] C. M. Yen and C. J. Chang, “A Utility-based TMCR Scheduling Scheme for Downlink MIMO-OFDMA Systems,” submitted to *IEEE Transactions on Vehicular*

*Technology* on 2009.3, revised on 2009.10 and 2010.2.

- [45] Z. J. Haas, J. Deng, “On optimizing the backoff interval for random access schemes,” *IEEE Trans. Commun.*, vol. 51, Dec. 2003, pp. 2081-2090.
- [46] C. M. Yen, C. J. Chang, and Y. S. Chen, “An Adaptive P-Persistent MAC Scheme for Multimedia WLAN,” *IEEE Communications Letters*, Vol. 10, No.11, Nov. 2006, pp. 737–739.
- [47] C. J. Chang, F. Z. M. Yen, Y. S. Chen, C. Y. Huang, “A novel adaptive p-persistent MAC scheme for WLAN providing low delay variance,” *ICC 2006*, pp. 3639-3644.
- [48] C. M. Yen, C. J. Chang, Y. S. Chen, and C. Y. Huang, “Analysis of an Adaptive P-Persistent MAC Scheme for WLAN Providing Delay Fairness,” *IEICE*, Vol. E93-B, No.2, Feb. 2010.
- [49] G. Bianchi, “Performance analysis of IEEE 802.11 distributed coordination function,” *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, March 2000, pp. 535-547
- [50] Z. Ziouva and P. Bahl, and S. Gupta, “CSMA/CA performance under high traffic conditions: throughput and delay analysis,” *Computer Communications*, vol.25, no. 3, Feb. 2002, pp.313-321.
- [51] Y. Xiao, “A simple and effective priority scheme for IEEE 802.11,” *IEEE Communication Letters*, vol. 7, no.2, Feb. 2003, pp. 70-72.
- [52] “IEEE Standard for Information technology – Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements,” *IEEE Std 802.11e–2005*, 2005.
- [53] N. H. Lee , J. G. Choi, and S. Bahk, “Opportunistic scheduling for utility maximization under QoS constraints,” *Proc. of PIMRC*, 2005, pp. 1818-1822.
- [54] M. J. Neely, “Super-fast delay tradeoffs for utility optimal fair scheduling in wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, Aug. 2006, pp. 1489-1501.
- [55] H. Lei, L. Zhang, X. Zhang, and D. Yang, “A packet scheduling algorithm using utility function for mixed services in the downlink of OFDMA systems,” *Proc. of VTC2007 Fall*, pp. 1664-1668.

- [56] J. Kim, S. Park, J. H. Lee, J. Lee, and H. Jung, "A scheduling algorithm combined with zero-forcing beamforming for a multiuser MIMO wireless system," *Proc. IEEE VTC 2005 fall*, pp. 211-215.
- [57] J. Zhang, G. Liu, and W. Wang, "Multiuser antenna selection for zero forcing beamforming based MIMO OFDMA," *Proc. IEEE APCC*, pp. 107-110, 2007.
- [58] B. Noble and J. W. Danlel, *Applied Linear Algebra 3rd*, Prentice-Hall, 1988.
- [59] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, issue 2, Feb. 2003, pp.171-178.
- [60] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. on Communications*, vol. 45, Oct. 1997, pp. 1218-1230.
- [61] WiMAX forum, "WiMAX system evaluation methodology," V.1.0, Tech. Rep., Jan. 2007.
- [62] H. Yaghoobi, "Scalable OFDMA physical layer in IEEE 802.16 WirelessMAN," *Intel Technology Journal*, Volume 8, Issue 3, 2004.
- [63] 3GPP TR 25.892, "Feasibility study for OFDM for UTRAN enhancement," 3rd Generation Partnership Project, Tech. Rep., 2004-06.
- [64] Universal Mobile Telecommunication System, Selection procedures for the choice of radio transmission technologies of the UMTS, UMTS Std. 30.03, 1998.
- [65] J. Chen and C. W. Chang, "A Signal-Aware Uplink Resource Allocation Strategy in IEEE 802.16 Systems" *Proc. WiMOB 2007*, pp. 15-20.
- [66] Louay M. A. Jalloul and Sam P. Alex, "Coverage Analysis for IEEE 802.16e/WiMAX Systems," *IEEE Trans. Wireless Communications*, vol. 7, pp.4627-4634, Nov. 2008.
- [67] H. Yaghoobi, "Scalable OFDMA physical layer in IEEE 802.16 WirelessMAN," *Intel Technology Journal*, Volume 8, Issue 3, 2004.
- [68] CISCO Tech Notes, "Voice over IP – per call bandwidth consumption," Document ID 7934.
- [69] IEEE 802.16m-08/004r2, IEEE 802.16m Evaluation Methodology Document (EMD), 2008-07-03.
- [70] 3GPP Technical Report 36.913 "Requirements for further advancements for E-UTRA (LTE-Advanced)", June 2008.

[71] 3GPP TR 36.814 V1.0.0. “Further Advancements for E-UTRA Physical Layer Aspects (Release 9)”, Feb. 2009.





# Vita

Chih-Ming Yen was born in Tainan, Taiwan, R.O.C., on March 18, 1980. He received B.E. degree from the Department of Mathematics, National Cheng Kung University, Taiwan, in 2003. Currently, he is a candidate for the Ph. D. in the Institute of communication engineering in National Chiao-Tung University, Taiwan. His research interests include performance analysis, protocol design, wireless communication networks.

