Chapter 1

Introduction

The standard 802.16e [1] has been standardized in 2006. This version provides the possibility of MSs with mobility which is really different from 802.16d [2] only provides fixed broadband wireless access. With the mobility, the role of the system is not only the last mile solution for the end users. The system will be more like cellular system but with higher transmission rate and bandwidth than 3G or even 3.5G system. With the mobility capability and higher bandwidth, users can access the wireless resource while they are moving. Higher bandwidth can also support services which must take more resource to transmit.

Obviously, 802.16e system will be one of the future wireless communication system solutions. In the future wireless communication system, the system must support mobility, higher bandwidth, and higher transmission rate to support more users and services which would cost lots of resources in the system. As the increasing of users and different kinds of services, try to maintain QoS requirement for different types of service will be more and more important. Therefore, using a mechanism schedules and distributes resource to each users and service to maintain their QoS requirement is essential. However, only maintaining the QoS requirement of services is not enough, due to the precious radio resource, system must utilize the radio resource efficiently. In this way, system can transmit with higher throughput and might accommodate more services.

Therefore, having a scheduling algorithm can maintain the services' required QoS and enhance the system performance is really crucial in the future wireless communication system. In the thesis, we will discuss this issue and propose a scheduling algorithm which can achieve the goal mentioned before. The algorithm will be implemented in the 802.16e system which is highly promoted by Intel and will be widely used in the future.

The rest of this thesis is organized as follows: In chapter 2, the overview of PHY and MAC layers of 802.16e standard, and the overview of some existing scheduling algorithm are briefly introduced. In chapter 3, the proposed algorithm will be discussed detailed. In chapter 4, the setting of simulation platform is addressed. In chapter 5 shows the simulation results. Finally in chapter 6, the conclusion and future works will be provided.

Chapter 2

Overview of 802.16e scheduling mechanism and algorithm

In 802.16e standards, system has ability to support mobility and many types of services. For different services, they have different bandwidth request mechanism and different QoS definition according to the service characteristics. System must base on the information to do scheduling to meet the requirement. In this chapter, first of all will introduce the PHY applied in this thesis. Then introduce MAC operation which will be involved in scheduling decision. Then will briefly introduce the scheduling algorithms which have been proposed by others.

2.1 Brief introduction of 802.16e PHY

The operation frequency of 802.16e standard [1] is distributed from 10-66GHz, which is line of sight transmission, or 2-11GHz, which is non line of sight transmission. In the urban area, line of sight transmission is nearly impossible due to the obstruction of buildings. So the research will focus on 2-11GHz PHY. There are three air interface defined in spec:

- WirelessMAN-SCa

  This uses a single-carrier modulated air interface.

- WirelessMAN-OFDM

  It is a 256 carrier orthogonal frequency division multiplexing scheme. Multiple access mechanism is time division multiple access based.

- WirelessMAN-OFDMA

It uses orthogonal frequency division multiple access. Multiple access is provided by assigning a subset of the carriers to specific user. In 802.16e standard, the number of subcarriers could be 128, 512, 1024, and 2048.

The standard supports two duplex methods: time division duplex (TDD) and frequency division duplex (FDD).

In this thesis, the research adopts OFDMA PHY and TDD mode. So in the latter part, the introduction of this system will be focused on the OFDMA PHY.

2.1.1 PHY Frame Structure

The OFDMA physical layer [3] [4] is different from other options. It can use the resource more flexible and efficient than SC and OFDM physical layer because of the two-dimension resource allocation. Besides, it can overcome the multi-path effect by guard interval and with the orthogonal subcarriers and flat narrow band channels, it is much easier to design equalizer and receiver.

In TDD frame structure, makes system separate downlink and uplink from time index. A frame is divided into two parts. First part is downlink subframe, it contains preamble, Frame Control Header (FCH), DL_MAP, UL_MAP and DL burst which could be used to transmit data. BS will transmit common control information and data to MSs. Second part is uplink subframe, MS may request bandwidth for uplink transmission or transmit control information and data to BS. Uplink subframe also contains some other control channels, like ranging channel, fast feedback channel, sounding zone, etc…

In figure 2-1 shows the frame structure of the system.

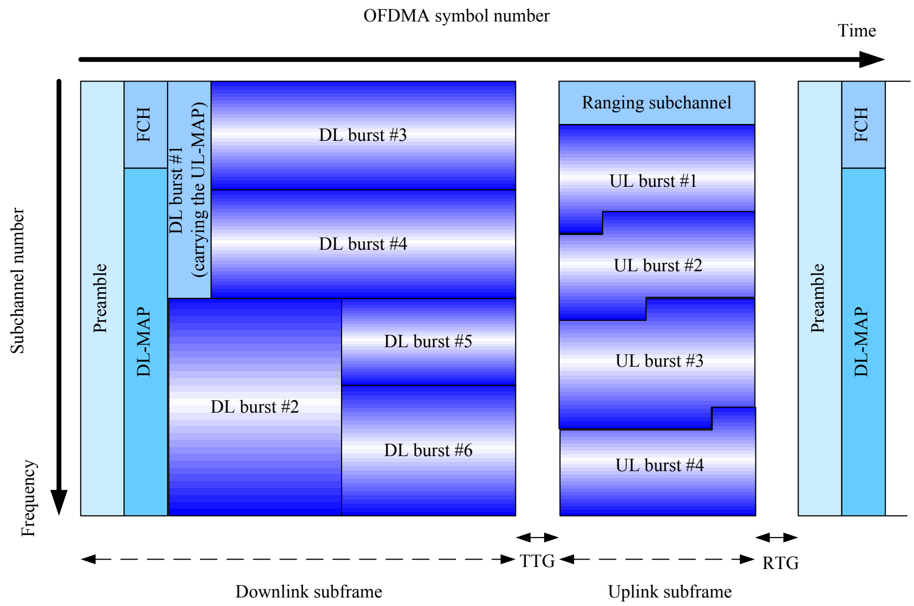

Figure 2-1 Example of an OFDMA frame (with only mandatory zone) in TDD mode

In the beginning of frame, it is consisted of preamble and some control information. The preamble is used for synchronization. The FCH contains DL_Frame Prefix, and specifies the length of the DL-MAP message that immediately follows the DL_Frame_Prefix and the repetition coding used for the DL-MAP message. The DL-MAP and UL-MAP completely describe the contents of the DL and UL subframes. The DL-MAP and UL-MAP are used for resource allocation of DL and UL data bursts, include what bursts belong to MSs, modulation and coding schemes per burst.

The UL-MAP also grants bandwidth to specific MSs. Therefore, the MSs can transmit data in the assigned uplink allocation. The DL bursts and UL bursts are used for data transmission of different users. The ranging subchannel is used for initial ranging, periodic ranging, and bandwidth requests. The initial ranging transmission shall be used by and MS wants to synchronize to the system channel for the first time. Periodic ranging transmissions are sent periodically for system time and power update. Bandwidth requests transmissions are for requesting uplink allocation from the BS. Ranging subchannels are dynamically allocated by the MAC layer and indicated in the UL-MAP. The transmit/receive transition gap (TTG) is a gap between the downlink burst and the subsequent uplink burst. It allows time for the BS to switch from transmission to receiving mode and MSs to switch from receive to transmit mode. The receive/transmit transition gap (RTG) is a gap between the uplink burst and the subsequent downlink burst. It allows time for the BS to switch from receiving to transmission mode and MSs to switch from transmit to receive mode.

2.1.2 PHY Slot and Data Mapping

The OFDMA slot is a minimum unit for data transmissions. One OFDMA slot occupies one subchannel and several OFDMA sysmols. For downlink Full Usage of Subcarriers (FUSC) using the distributed subcarrier permutation, one slot is one subchannel by one OFDMA symbol. For downlink Partial Usage of Subcarriers (PUSC) using the distributed subcarrier permutation, one slot is one subchannel by two OFDMA symbols. For uplink PUSC using either of the distributed subcarrier permutations, one slot is one subchannel by three OFDMA symbols. For uplink and downlink Band Adaptive modulation and coding (Band AMC) using the adjacent subcarrier permutation. One slot is one subchannel by one, two, three, or six OFDMA symbols.

A Data Region is a two-dimensional allocation which contents a group of contiguous subchannels and OFDMA symbols. All the allocation refers to logical subchannels. The minimum unit of data mapping is an OFDMA slot.

Based on standard, how many and which resource units would be assigned to a transmission is decided by BS and the mechanism is different in downlink and uplink transmission. In downlink resource allocation, system will consider the data size and try to fulfill the resource units first in frequency domain. After the frequency domain is fulfilled, then it goes to the next time domain to fulfill resource units. Which are really different in uplink resource allocation, when uplink resource allocation will try to fulfill resource units in time domain, after the time domain resource units are full, then go to another frequency domain and repeat the procedure.

In figure 2-2 and figure 2-3, show downlink resource allocations and uplink resource allocation mechanism.

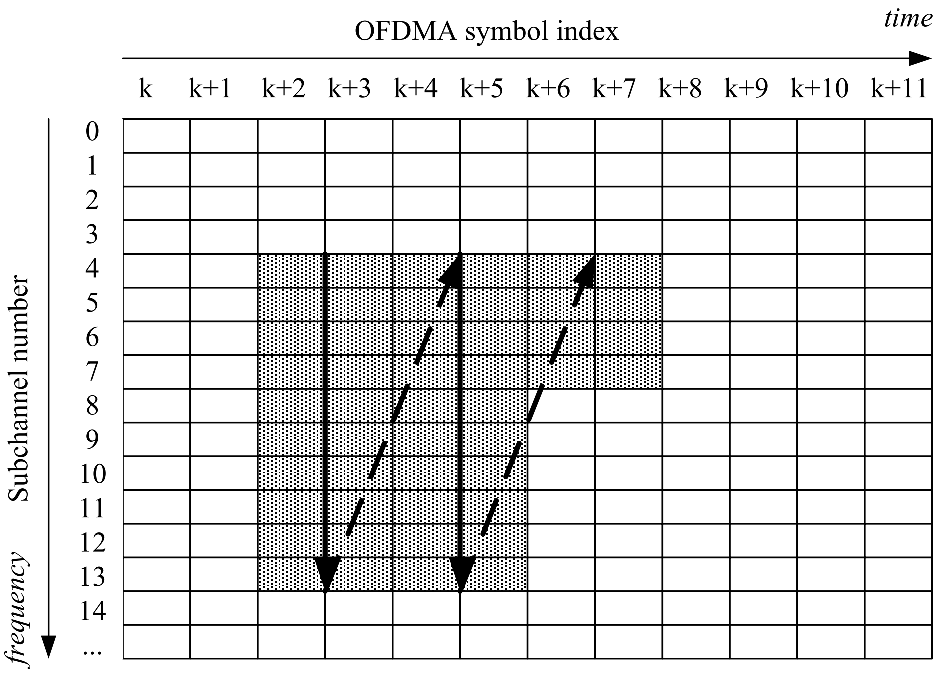

Figure 2-2 Example of mapping OFDMA slots to subchannels and symbols in the downlink PUSC mode

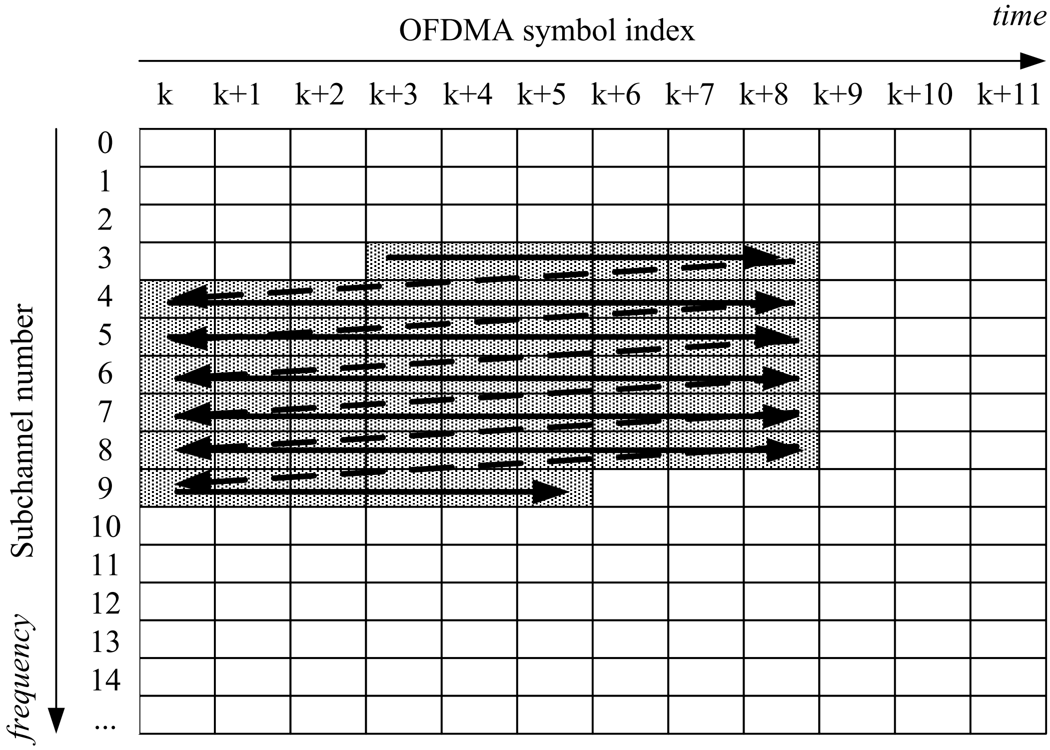

Figure 2-3 Example of mapping OFDMA slots to subchannels and symbols in the uplink PUSC mode

After decided which resource will be allocated to which users, BS forms DL-MAP and UL-MAP message and includes the results.

2.1.3 Subcarrier Permutation

Subcarrier permutation is a method to assign frequency subcarriers into subchannels. The allocation of subcarriers to subchannels is accomplished via permutation rule.

There are two kinds of permutation modes: distributed subcarrier permutation and adjacent subcarrier permutation. Distributed permutation means the subcarriers belonged to a subchannel are selected pseudo randomly from all subcarriers. It can average intercell interference and avoid fading effect. The adjacent subcarrier permutation will form the subchannel whose subcarriers coming from adjacent subcarriers. In this method, system can take advantage of frequency select fading and get multiuser diversity form frequency domain.

802.16 standard provides three ways to group subcarriers into subchannels:

- Full Usage of Subchannels (FUSC)

  This method is used in downlink only and can use all subcarriers to do permutation for one subchannel. It can achieve the best frequency diversity by spreading subcarriers over entire band. It will use distributed permutation mode.

- Partial Usage of Subchannels (PUSC)

  This method can be used both in downlink and uplink. It will group subcarriers first then choose subcarriers per group and each group only provides one subcarriers to form a subchannel. It also use distributed permutation mode.

- Band Adaptive Modulation and Coding (BandAMC)

  This method also can be used in downlink and uplink. It uses adjacent permutation mode. The bandwidth is divided into sub-band and tries to utilize the frequency select fading to enhance system performance.

2.1.4 Adaptive Modulation and Coding (AMC)

802.16e system is able to adjust modulation coding scheme depends on the carrier to interference plus noise ration (CINR) condition of the radio link. 802.16e OFDMA architecture supports multiple modulation methods: Quadrature Phase Shift Keying (QPSK), 16-state Quadrature Amplitude modulation (16-QAM), and 16-state Quadrature Amplitude modulation (64-QAM). It can be adjusted dynamically to trade off between efficiency and robustness. The OFDMA architecture also supports several coding schemes, like Convolution Code (CC), Low Density Parity Check Code (LDPC), Block Turbo Code (BTC), and Convolution Turbo Code (CTC).

2.2 Brief introduction of 802.16e MAC

The MAC layer of the system will process several tasks which is really important in the communication system. It must manage the radio resource to decide how to let users access the resource, handoff users if they are going to leave the cell coverage, decides whether to let users register into the system, do power control or rate control, etc… Besides, radio resource management, an important issue should be done to maintain the QoS requirement is also implemented in this layer [5] [6] [7]. System must integrate the information from upper layer and PHY layer. It should know the traffic QoS requirement to allocate suitable resource to users. It will do retransmission if PHY tells there is error occurs in the transmission before. It should fragment or pack the SDU into a MAC PDU to avoid transmission error occurs or transmission inefficient due to too many overhead.

In this section will be the overview of 802.16e MAC structure and introduce some of the MAC function which will take a part in the scheduling, which will be the main topic of this thesis. Some in the rest of this section, we will focus on the QoS classes defined in the standard, request-and-grant mechanism, and channel condition feedback mechanism which are more involved in the scheduling decision. Other function of MAC will beyond the scope of this thesis.

2.2.1 802.16e MAC structure

MAC layer of 802.16e will be divided into three sublayers as shown in figure

2-4.

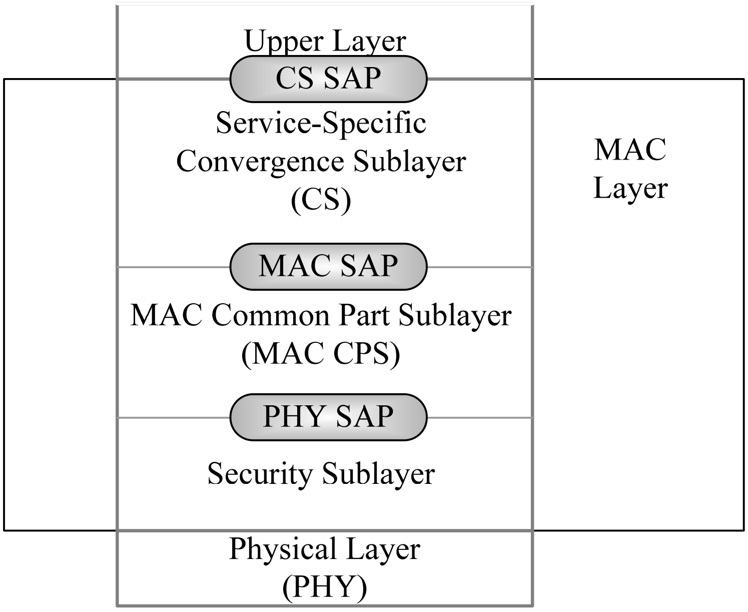

Figure 2-4 802.16 protocol layer

There are three sublayers in the MAC layer. All of them have their own functionality.

- **Service-Specific Convergence Sublayer (CS)**

  This sublayer is an interface between upper layer and MAC layer. It will identify different traffic from upper layer and assign connection ID (CID) to each connection. It also provides payload header suppression and reconstruction to enhance the airlink efficiency.

- **Common Part Sublayer (CPS)**

  The CPS sublayer manages the main function of controlling the whole radio resource. QoS control, fragmentation and packing, scheduling, request-and-grant, admission control, handover, and ARQ will be controlled in this sublayer. Besides, segmentation of SDU into MAC PDU is also implemented in this sublayer.

- **Security Sublayer**

This sublayer performs the authentication of network access, registration, key exchange, and encryption of PDUs.

In the rest of the sector will focus the functionality of CPS sublayer.

2.2.2 MAC PDU Formats

The MAC PDU is a data exchanged unit between the MAC layer of the BS and MSs. A MAC PDU consists of a 48bit MAC header, a variable length data payload, and an optional 32 bits Cyclic Redundancy Check (CRC). Sometimes some MAC PDU will not include payload and CRC bits. These kinds of PDUs are used only in the uplink to transmit control message. Those MAC signaling headers include bandwidth request, uplink transmit power report, CINR report, CQICH allocation request, PHY channel report, uplink sleep control, SN report, and feedback functionalities. MAC PDUs also include some subheaders. Those subheaders will be inserted in MAC PDUs following the generic MAC header. Those subheaders help system perform grant management, packing, ARQ feedback, etc…

2.2.3 Fragmentation and Packing

In 802.16e system, MAC SDUs coming from CS will be formatted according to the MAC PDU format in the CPS, possibly with fragmentation and packing. That's due to the precious radio resources and system hopes to utilize the resources efficiently.

Fragmentation process means to divide a SDU into different PDUs payload areas. That is because the MAC PDU size is limited by standard, 2048 bytes. Besides, larger PDU size may causes error occurs more easily. Therefore, divide SDU properly according to the channel condition will avoid transmission errors and save the resource used for retransmission. Figure 2-5 shows the process of fragmentation.

Packing process is to pack several SDUs into a single PDU payload. In this way, system may avoid resource waste due to the overhead caused by MAC header and CRC. Figure 2-6 shows the process of packing.

Both processes may be initiated by either a BS for a downlink connection or a MS for and uplink connection.

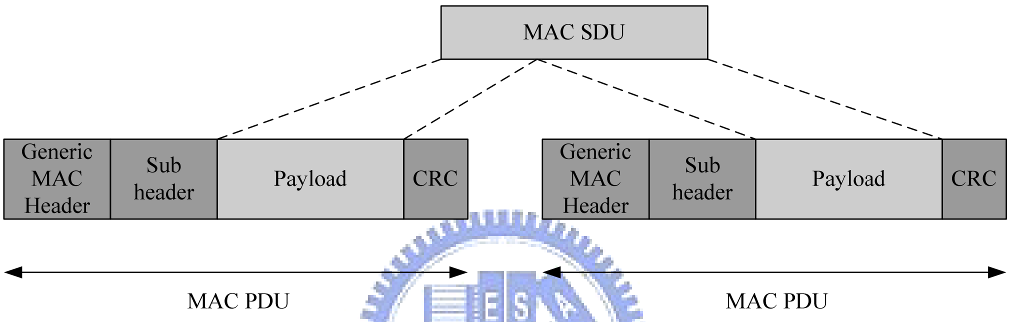

Figure 2-5 Fragmentation

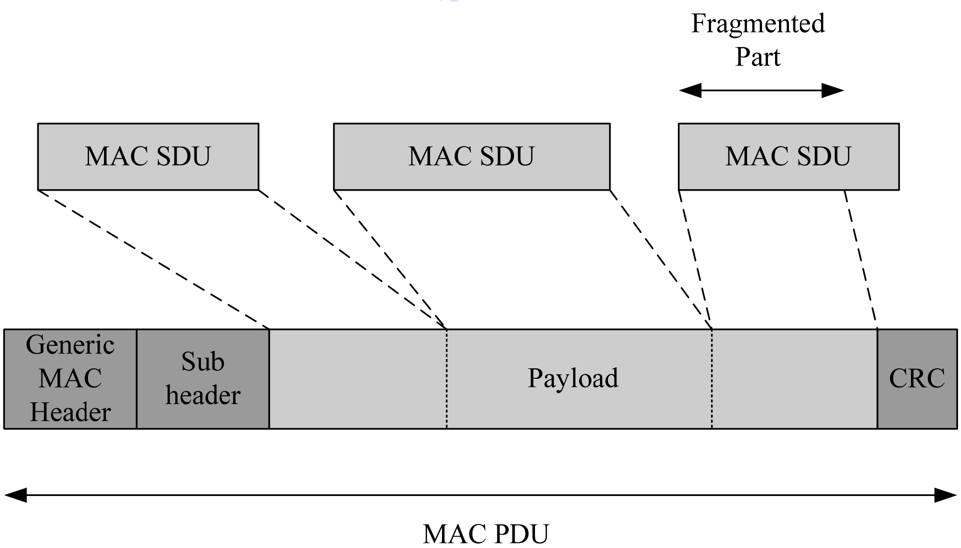

Figure 2-6 Packinig

2.2.4 QoS based service classes

The WiMAX 802.16e standard provides several QoS classes for system supports different kinds of service. For different classes, system sets different parameters and transmission/request methods to let system maintain the QoS requirement for different kinds of service. Here will introduce those classes:

In downlink, it defines four kinds of QoS classes.

a. Real-time CBR data streams

These kinds of service are designed to support real-time service flows that generate fixed-size data packets on a periodic basis, such as T1/E1 and VoIP without silence suppression. There are several parameters used for maintaining the QoS: Maximum Sustained Traffic Rate, Maximum Latency, Tolerated Jitter, and the Request/Transmission Policy.

b. Real-time VBR data streams

These kinds of service are designed to support real-time service flows that generate variable size data packets on a periodic basis, such as moving pictures experts group (MPEG) video. There are several parameters used for this service: Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Maximum Latency, and the Request/Transmission Policy.

c. Delay-tolerant VBR data streams

These kinds of service are designed to support delay-tolerant data streams consisting of variable-sized data packets for which a minimum data rate is required, such as FTP. There are several parameters used for this service: Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Traffic

Priority, and the Request/Transmission Policy.

d. Best effort data streams

These kinds of service are designed to support data streams for which no minimum service level is required and therefore may be handled on a space-available basis. The mandatory QoS service flow parameters for this scheduling service are Maximum Sustained Traffic Rate, and Request/Transmission Policy.

In uplink, it defines five kinds of QoS classes.

a. Unsolicited Grant Service (UGS)

The UGS is designed to support real-time service flows that generate fixed-size data packets on a periodic basis, such as T1/E1 and VoIP without silence suppression. There are several parameters used for this service: Maximum Sustained Traffic Rate, Maximum Latency, Tolerated Jitter, Uplink Grant Scheduling Type and the Request/Transmission Policy.

b. Real-time Polling Service (rtPS)

The rtPS is designed to support real-time service flows that generate variable size data packets on a periodic basis, such as moving pictures experts group (MPEG) video. It is used for VBR service. There are several parameters used for this service: Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Maximum Latency, Uplink Grant Scheduling Type, and the Request/Transmission Policy.

c. Extended Real-time Polling Service (ertPS)

Extended rtPS is a scheduling mechanism which builds on the efficiency of both UGS and rtPS. The Extended rtPS is designed to support real-time service flows that generate variable size data packets on a periodic basis, such as Voice over IP services with silence suppression. The parameters are Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Maximum Latency, and the Request/Transmission Policy.

d. Non-real-time Polling Service (nrtPS)

The nrtPS is designed for non-real-time service which can tolerate more delay, such as FTP, web-browsing, etc… There are several parameters used for this service: Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Maximum Latency, and the Request/Transmission Policy.

e. Best Effort Service (BE)

BE service is with the lowest QoS level. These kinds of service are designed to support data streams for which no minimum service level is required and therefore may be handled on a space-available basis. The mandatory QoS service flow parameters for this scheduling service are Maximum Sustained Traffic Rate, and Request/Transmission Policy.

2.2.5 Request-and-Grant Mechanism

The request-and –grant mechanism is different form QoS classes because of their characteristics and requirement.

In downlink, it is much easier to handle these procedures. BS has the precise information about traffic requirement. Therefore, BS can easily give bandwidth to downlink traffic. It's no need to do request and grant. BS will do everything.

In uplink, it is more complicated to do request-and-grant. Before listing those different QoS classes use which mechanism, we introduce the request-and-grant mechanism first.

Requests refer to the mechanism that MSs use to indicate to the BS that they need uplink bandwidth allocation. The Bandwidth Request message may be transmitted during any uplink allocation, except during any initial ranging interval. In the message will indicate the size of opportunities to be requested.

The Bandwidth Request message could be transmit through polling mechanism. Polling is the process by which the BS allocates to the MSs bandwidth specifically for the purpose of making bandwidth requests. MS can be polled individually, that is unicast polling. Or MSs can be formed into a group and polled together, that is multicast polling or broadcast. It is used when the bandwidth is insufficient for unicast polling. Besides polling, some MS get bandwidth through contention in the ranging subchannel. Only if MS contents successfully, they can get bandwidth to transmit.

After BS receives bandwidth request message, BS will allocate bandwidth to MSs according to their request if possible. However, some services may get uplink transmission bandwidth even if they don't send request message. This kind of service called Unsolicited Grant Service. System will allocate transmission opportunity in a constant period for users to transmit.

After introducing the Request-and-Grant mechanism, here listed which mechanism could be used in different QoS classes:

a. Unsolicited Grant Service (UGS)

MS is prohibited from using any contention request opportunities for this connection. BS assigns resource to MS to transmit in an unsolicited way in a prescribed interval. The BS shall provide unicast grants in an unsolicited manner.

b. Real-time Polling Service (rtPS)

The service offers real-time, periodic unicast request opportunities, which meet the flow's real-time needs and allow the MS to specify the size of the desired grant. If MS needs uplink transmission, it will send request message in the opportunities polled by BS. MS is prohibited from using any contention request opportunities for that connection.

c. Extended Real-time Polling Service (ertPS)

The BS shall provide unicast grants in an unsolicited manner like in UGS, thus saving the latency of a bandwidth request. However, whereas UGS allocations are fixed in size, ertPS allocations are dynamic. The BS may provide periodic uplink allocations that may be used for requesting the bandwidth as well as for data transfer. By default, size of allocations corresponds to current value of Maximum Sustained Traffic Rate at the connection. The MS may request changing the size of the uplink allocation.

d. Non-real-time Polling Service (nrtPS)

This kind of service offers unicast polls on a regular basis, which assures that the service flow receives request opportunities even during network congestion. The BS typically polls nrtPS on an interval on the order of one second or less. The BS shall provide timely unicast request opportunities but MS is allowed to use contention request opportunities.

e. Best Effort Service (BE)

BE service is with the lowest QoS level. It can't get resource through polling. It only get resource in the contention way.

2.2.6 Channel condition feedback

Due to the mobility, channel condition feedback is critical in 802.16e system. With the mobility, the channel condition may change rapidly. If there is no feedback information about channel quality, system can not decide which modulation coding scheme to be used. It might cause error or transmit inefficiently.

In 802.16e system, CINR could be an index to indicate channel quality. BS and

MS may measure the CINR to get the channel condition information and send it back. In distributed subcarrier permutation mode, to get downlink channel quality, MS is responsible to measure the preamble or a permutation zone and get the CINR value. Then MS will send REP-RSP message to BS to report the estimated CINR. Then BS can use the information to decide how to transmit data in the next frame. REP-RSP message could be sent in response to REP-REQ message from BS or in an unsolicited fashion. The REP-RSP message could be sent through CQICH. The location of CQICH is different from MSs and is assigned by BS. As to adjacent subcarrier permutation mode and applied for BandAMC mode, MS will measure CINR on different band and report the message through CQICH too.

If system wants to get uplink channel condition, MS can use UL-Sounding Zone and transmit data to let BS estimate the CINR.

2.2.7 Scheduling

Based on the information mentioned before, BS will do the scheduling decision and form DL-MAP and UL-MAP to indicate the scheduling result. System could consider the QoS requirement of different service and their channel condition to do the scheduling decision. In the spec, it doesn’t define the scheduling information and this part is left for design. And a proposed scheduling algorithm will be discussed in detail in this thesis.

2.3 Scheduling algorithm

The topic of this thesis is scheduling algorithm, so first of all we must have some basic knowledge about scheduling that how other people work.

A good scheduling algorithm shall utilize the system resource well and meet the different users' requirement. In the earlier system, there is only voice or data service. It is much easier to design scheduling algorithm. However, nowadays, it is common that there is mix-service in the system, not only a single service type. To handle the mix-traffic situation, the scheduling design will be more complicate to meet their requirement.

In this section will introduce some scheduling algorithm proposed by others [8] [9] [10] [11] [12] [13]. Different scheduling algorithm has different target to achieve. Some of them want to achieve fairness and some of them hope to maximize throughput or maintain the QoS. These algorithms have their own advantages, however, they have disadvantages in some aspects too.

- **Round-Robin (RR)**

  In round-robin scheduling algorithm, system schedules packets by users in a fixed sequence irrespective of the channel condition and services requirement. This algorithm provides fairness but ignore the channel condition and QoS requirement. It is hard to be used in mix-traffic system due to different QoS requirement.

- **MaxCINR**

  MaxCINR scheduling algorithm will schedule packets which belonged to the user with better CINR, meaning the channel condition is much better. This kind of scheduling algorithm takes channel condition into consideration and may provide good multiuser diversity to enhance system throughput and shorter transmission time due to efficient transmission. However, it is less fairness. If the power control doesn't implement well, some users might never get opportunity to transmit. Besides, it can't provide QoS guarantee especially to real-time service, which QoS definition is usually delay bound consideration. System will select user $m*$ that fulfill:

$$m_t^* = \arg\max_m CINR_m(t) \qquad (1)$$

- **Proportional Fair (PF)**

This algorithm considers channel condition and fairness across users. It provides a tradeoff between system throughput and user fairness. The scheduling decision will follow a ratio which is instantaneous data-rate over average data-rate and pick larger user $m*$:

$$m* = \arg\max_m \frac{R_m(t)}{S_m(t)} \qquad (2)$$

$$S_m(t) = (1-\frac{1}{W})S_m(t-1) + \frac{1}{W}R_m(t)\delta(t-m) \qquad (3)$$

, where $R_m(t)$ denotes the achievable instantaneous data-rate for user m, $S_m(t)$ denotes the moving average of data-rate at user m, and W donates the length of moving average. This algorithm provides both fairness and channel condition concerns, but it lacks QoS guarantee, especially for real-time service.

- **Modified Largest Weighted Delay First (m-LWDF)**

This algorithm is similar to the Proportional Fair scheduling algorithm, consider fairness and system throughput. More than that, it adds another factor, delay, into consideration. System will pick user $m*$ if

$$m_t^* = \arg\max_m a_m SVP_m(t) R_m(t) / S_m(t) \qquad (4)$$

, while a_m indicates the QoS level for user m, SVP_m indicates the starvation period, meaning the delay for user m. In this way, system can avoid users suffering from long delay.

- **Optimum Channel-Aware Scheduling with Service Differentiation (OCASD)**

This algorithm provides throughput gains while offering fairness and traffic delay constraints. However, this algorithm doesn't provide strict delay constraint which is important for delay-sensitive service. System picks connection x^* while

$$x_t^* = \arg\max_x \{ [\left| w_i - \frac{B_i(t)}{B(t)} \right| \cdot B(t)]^{\operatorname{sgn}(w_i - \frac{B_i(t)}{B(t)})} R_m(t) / d_i \} \qquad (5)$$

$$B_i(t+1) = (1 - \frac{1}{W}) B_i(t) + \frac{1}{W} \times L_i(t) \delta(t-i) \qquad (6)$$

, where w_i is the weight depending on the QoS requirements of traffic class i. d_i is the delay bound of the HOL packet which size if $L_i(t)$ in queue i. $B(t)$ is the aggregate of backlogged queues at time t. $R_m(t)$ denotes the achievable instantaneous data-rate for user m.

- **Traffic-Aided Opportunistic Scheduling (TAOS)**

This algorithm utilizes the file size information and user-diversity concept to try to reduce file completion time, which is waiting time plus transmission time. By reducing the transmission time, enhance the system throughput. System will pick user m^* if

$$m_t^* = \arg\min_m S_m(t) F_m / R_m(t) \qquad (7)$$

, where F_m donates the size of the HOL packet in user m's queue. This algorithm doesn't consider delay constraint and fail to ensure the QoS guarantee.

- **Early Deadline First (EDF)**

This algorithm considers completely the delay constraint. It ignores the channel quality so it might not use bandwidth efficiently. However, it can provide strict QoS guarantee for delay-sensitive service because it will transmit packet first which is much closer to their delay bound. System will transmit packet belonged to user m if

$$m^* = \arg\min_m \{ DB - Age - T_t \} \qquad (8)$$

, while DB means delay bound, Age is the time that the packet stayed in MAC, and T_t is the transmission time for this packet in transmitted in the current frame.

Chapter 3

Proposed scheduling algorithm

Just as mentioned before, the scheduling control mechanism will play an important role in the future communication system. Therefore, we propose an algorithm which will solve some of the challenges. In this chapter, we will introduce the proposed algorithm. And in chapter 4 and 5 will introduce the simulation setup and show the simulation results.

3.1 Goal

First of all, the algorithms mentioned in the former chapter designed to meet different goals, but cause some disadvantages. For example, the algorithm which considers the QoS requirement for real-time service strictly, but it ignores the channel condition and can't utilize bandwidth efficiently to enhance system performance. Besides, non-real-time service can't be taken into consideration because it doesn't have delay bound without modifying the algorithm and QoS definition. For those algorithms consider both the delay and channel condition, however, it didn't set the strict QoS requirement like EDF algorithm, not mention those algorithms which didn't consider delay at all. Therefore those algorithms without considering QoS strictly are hard to be used in the mix-traffic system due to lack of QoS guarantee for delay-sensitive service.

Hoping to be used in the wide band system, the ability to serve the mix traffic services is essential. System must concern different services' different QoS requirements and try to meet the requirements through scheduling design. Hoping to deal with mix traffic service, we give the QoS definition to different traffic services and unify the QoS requirement for different types of services. In this way, we can use the same indication to schedule users even if they are not belonged to the same type of services. Besides maintaining QoS requirement, utilize the bandwidth efficiently is important. If system utilizes the precious bandwidth in efficient way, the system will have larger throughput and might accommodate more users. Therefore, how to use the bandwidth is an important topic.

The topic mentioned before is what our proposed scheduling algorithm trying to solve:

1. Maintain the QoS requirement for different types of services.
2. After system make sure the QoS requirement be satisfied, try to enhance the system performance like throughput, capacity, etc… by scheduling users in a more efficient way.

For the straightforward thinking, the proposed algorithm will be a combination of EDF algorithm and opportunistic scheduling algorithm, which transmits packets with better channel condition. However, after we implement the straightforward thinking algorithm, which will be called **EDFOP algorithm**, we find out it will cause some problems. Therefore, we add some other mechanisms to get rid of the problems and form our proposed algorithm, **QoS-based opportunity scheduling**. Both of these two algorithms will be introduced in this chapter.

3.2 Proposed algorithm

The proposed algorithm will be a packet scheduling algorithm to try to meet packets' QoS requirement and try to enhance the system performance. It will be implemented in 802.16e system in downlink. BS will schedule the packets based on their traffic QoS information from upper layer, channel quality feedback, and ACK/NACK message from uplink subframe in the previous frame in the beginning of the current and form the DL-MAP message to indicate the scheduling result.

The proposed scheduling algorithm will be a two-stage scheduling algorithm. Based on the information forwarded to BS, BS will schedule packets first which need to be retransmitted because error occurs in the previous transmission. Then schedule the packets which are urgent and must be served first to maintain their QoS requirement or they might fail to meet the QoS. After that, if there still remains resource, BS will schedule packets based on the channel quality feedback to determine which packets to be transmitted will have better efficiency. In other words, the downlink subframe will be divided into three parts shown in Figure 3-1.

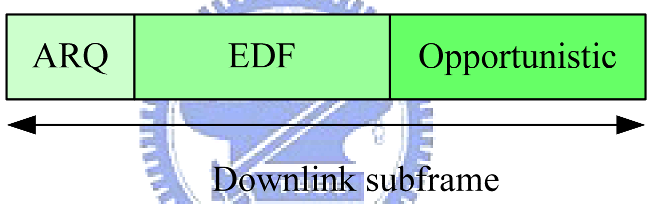

Figure 3-1 Portions divided in downlink subframe

Later we will introduce the proposed algorithm but before that, we must know the definition for QoS for different types of services.

3.2.1 QoS definition

The definition of QoS is always different between real-time service and non-real-time service. As to real-time service, the QoS definition is the delay concerns. For each packet, it has its own delay bound and must be transmitted before the bound or the packet will be dropped. As to non-real-time service, there is no delay bound requirement but they have the minimum reserved rate requirement. Although non-real-time service can tolerate delay, it still must be transmitted in an acceptable rate.

We propose a new definition to non-real-time service. The new definition is not only still related to the definition of minimum reserved rate, but it translates the minimum reserved rate concept into time concept. With the common definition for both real-time and non-real-time services, it will be much easier to do the scheduler design. Therefore, in our scheduling algorithm, we can use a simple but reliable way to maintain the mix-traffic services' QoS requirement.

We propose the "soft delay bound" concept for non-real-time service. We will translate the QoS definition from minimum reserved rate to delay bound concept. In this way, both real-time and non-real-time service can be controlled to meet their QoS requirement by their "delay bound".

The minimum reserved rate means how many data should be transmitted in a certain period to meet the QoS requirement. If we look the definition in other aspect, how many time before should system transmit the known size packet will meet the QoS requirement. And that time will be the soft delay bound for non-real-time service (9).

$$soft\ delay\ bound = \frac{packet\ size}{minimum\ reserved\ traffic\ rate} \qquad (9)$$

The soft delay bound will be an indication of the minimum performance that the non-real-time service should achieve. So, if the packet be transmitted successfully before the soft delay bound, it means the packet meets the QoS requirement. However, it doesn't mean that the non-real-time service can only achieve the defined performance. If there are still resource could be used, the non-real-time service can have better performance. Maybe the packet can be transmitted much earlier before the soft delay bound.

Why do we call it "soft delay bound"? That's because the different characteristic between real-time and non-real-time service. If the real-time service exceeds the delay bound, the packet will be dropped because it is useless even. However, non-real-time service must focus on the correctness. It can tolerate delay but doesn't like data loss. Therefore system will still transmit the non-real-time service packet even if it exceeds the soft delay bound.

Besides the drop and transmit issue, the soft delay bound of non-real-time service will be accumulated across the packets belonged to the same user. It can be expressed in the following formula, P denotes the packet size:

$$\frac{(P_1 + P_2 + \cdots + P_n)}{t} \geq \textit{minimum reserved rate} \qquad (10)$$

$$\begin{aligned} \textit{soft delay bound} &\leq \frac{(P_1 + P_2 + \cdots + P_n)}{\textit{minimum reserved rate}} \\ &= \frac{P_1}{\textit{minimum reserved rate}} + \frac{P_2}{\textit{minimum reserved rate}} + \cdots + \frac{P_n}{\textit{minimum reserved rate}} \end{aligned} \qquad (11)$$

It means, a single user might have several packets to be transmitted. The soft delay bound will accumulate across packets. If the first packet is transmitted much earlier before its' soft delay bound. The second packet of the user will have much longer soft delay bound because it take benefit from the first packet's fast transmission. However, if the first packet is transmitted really late even after the required soft delay bound, that means the packet doesn't meet the QoS requirement. Besides this, it will influence the next packet because the next packet will have shorter delay bound than normal because the former packet takes much longer time.

For example, there are two packets belonged to a user with the same packet size. Therefore both packets have the same soft delay bound. However, if packet one transmits quickly, the packet two will have longer delay bound shown in Figure 3-2. In other way, if packet one transmits late, packet two will have shorter delay bound or even the delay bound value is negative value shown in Figure 3-3.

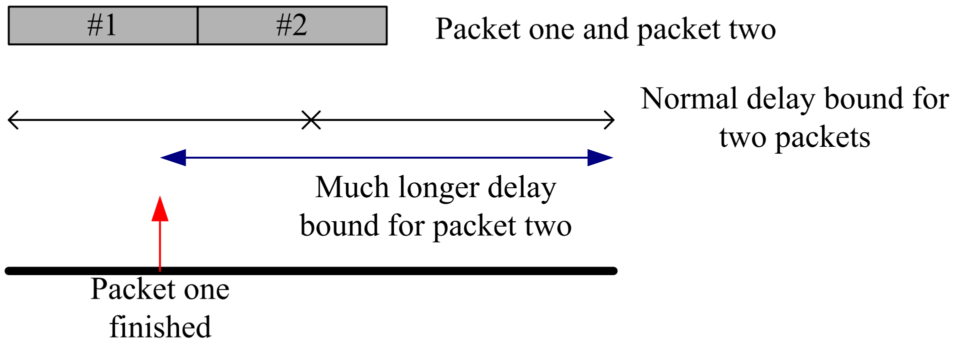

Figure 3-2 Soft delay bound effect-1

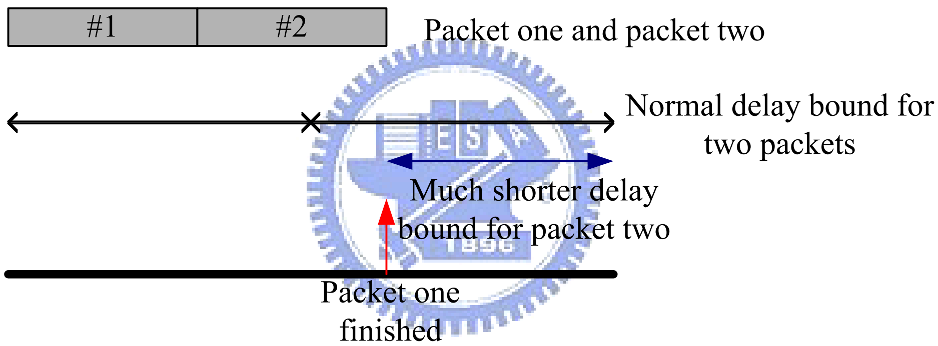

Figure 3-3 Soft delay bound effect-2

In 802.16e standard, it sets some parameters for different services to define the QoS. For real-time service, there is a parameter “maximum latency”, which can be used as the value of delay bound for VoIP and streaming service. For non-real-time service, there is a parameter ”minimum reserved traffic rate”, which can be used to conduct the soft delay bound value.

In this way, in our proposed algorithm, we use the delay bound as the QoS concern for both real-time and non-real-time services. If system transmit packets successfully before their delay bound, it means it maintain the QoS requirement of the packets.

3.2.2 EDF scheduling

In the beginning of the frame, the resource should be given to the packets which needed to be retransmitted. So, the PDU which needed to be retransmitted will be schedule first. After that, system will try to maintain the QoS requirement of packets in the system. In this thesis, the QoS requirement definition is delay bound for both real-time and non-real-time service. Hence, system will schedule packets which are urgent first. How to tell a packet is urgent or not? A simple definition for urgent packet is “if the packet doesn’t be transmitted in the current frame, it might fail to meet the QoS requirement” means that is might exceed the delay bound.

So system must find the urgent packets first.

How does the system find those packets? System will concern that when will the packet reaches delay bound for each packets, PDU retransmission time, and how many bits can be transmitted in the most robust burst profile. All the system wants to do is hoping transmit the packet successfully before the delay bound.

Therefore, system will use the information coming from upper layer indicated the traffic QoS information. Using the QoS definition mentioned before, each packet will have its own delay bound. System must transmit this packet successfully before the delay bound to maintain the QoS requirement. However, due to the unstable wireless environment, sometimes packet error will occur. To conquer the packet error, PDU retransmission is necessary. So, the retransmission time must be taken into consideration in the scheduling decision. System must conserve the retransmission time before delay bound because it might be useful when error occurs. So, the transmission for one packet must be much earlier than its real delay bound to make sure the correctness of transmission.

For example, in Figure 3-4, the packet delay bound could be extended to frame number four. However, to avoid transmission error, the PDU retransmission time is three so PDU could be retransmitted three times to combat the transmission error. Therefore the system should conserve three frames for retransmission. It will cause shorter delay bound for service. In the figure, the packet delay bound after conservation will make the packet transmission must occur before frame number one.

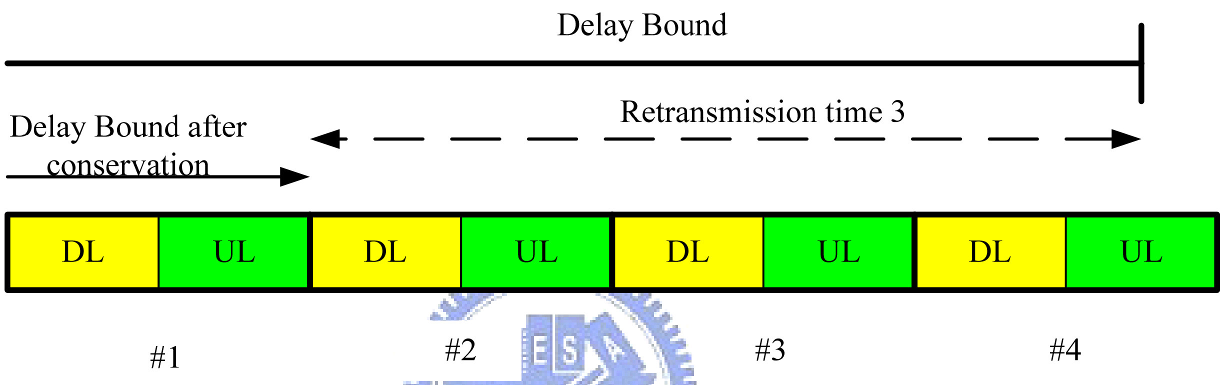

Figure 3-4 ARQ retransmission time causes shorter delay bound

The retransmission time varies from service. For service which data amount is less, the retransmission time will be less too because the error seldom occur and PDU retransmission will success easier. For example, VoIP services' PDU retransmission time can be set to one, while others will be set to three.

After system calculate the delay bound after conservation for PDU retransmission, however, due to TDD frame structure, the available transmission time is not much as we think because data only could be transmit in downlink subframe. So, system must try to find the really available time for transmission.

In Figure 3-5, after calculate the delay bound after conservation, there are three frame for packet transmission. However, packet only can be transmitted in the downlink subframe. The available transmission time is not 15ms (5ms per frame). The available transmission time is only 7.5ms (supposed that half of the resource assigned for downlink transmission).

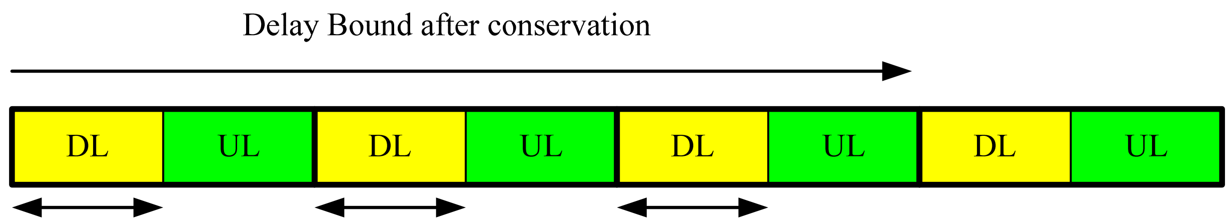

Figure 3-5 Transmission opportunity will be much shorter than imagination

System must use the available transmission time to transmit the packet. Then, system should consider the transmission time. It can easily calculate based on packet size based on traffic information and the modulation coding scheme decided based on the feedback information. To be conservative, system use the most robust modulation coding scheme to calculate the transmission time. System knows how many bits can be carried in single resource unit. And with the knowing of how many resource units in the frequency domain, system will know how many data can be transmitted in a slot time. Based on the information, system can calculate the transmission time for each packet.

In the last part, using the available transmission time minus the transmission time for packet, system can get a value. The value means before what time the system should transmit the packet in order to maintain the QoS requirement and the data correctness.

Let’s remind the goal of out algorithm. We hope to maximize the system efficiency, means throughput, but we must satisfy the QoS requirement. Therefore, system should find the urgent service which means it must be transmit right this frame or it will fail to maintain the QoS requirement. As for other packets, system may wait and observe the channel quality and transmit the packet when the channel quality is good enough.

So system will pick the "urgent packet" first to do the first time scheduling and let these packets transmit first to maintain their QoS requirement. But how?

$$num\_frame = \left\lceil \frac{life\_time}{frame\_length} \right\rceil \qquad (12)$$

$$transmit\_frame = num\_frame - \max(ARQretransmission) \qquad (13)$$

$$t_{availble} = DL\_subframe \times transmit\_frame \qquad (14)$$

$$t_d = \left\lceil \frac{packet\ size}{\min(bit\ \ per\ resource\ unit) \times total\,subchannel} \right\rceil \times symbol\_time \qquad (15)$$

$$t_{edf} = t_{available} - t_d \qquad (16)$$

$$t = t_{edf} - DL\_subframe \qquad (17)$$

$$if\ \ t \le 0, this\ service\ \ flow\ must\ be\ transmitted\ in\ this\ \ frame$$

In (12), firstly system calculates how many frames the packet can wait until its delay bound. Life time means the time to the packet's delay bound. Then, system conserves some frames for retransmission in (13). Due to the TDD frame structure, in (14) system calculates the available time for transmission. Then calculate the data transmission time in (15). In (16), use the value got before to calculate how many time does the system can wait for transmission but will not disobey the QoS requirement. Finally, using the value get from (16) minus downlink subframe time to determine if it can be transmit in the next frame. If the value t is larger the zero, it means it can wait for transmission in the next frame but still meet the QoS requirement. However, if the value t is less than zero, it means it must be transmitted right this frame or it will disobey the QoS requirement. So it is the urgent packet and must be schedule in the first round to meet the QoS requirement.

In this method, system picks the packets whose t is less than zero because they are the urgent packets and should be scheduled first. The scheduling decision of these packets will based on t_{edf}, the value indicate how many times the packet can wait for transmission. Here we use the earliest deadline first (EDF) concept. System schedules the packet with shorter t_{edf} because it will exceed the deadline earlier than others.

However, due to the non-real-time service will not be dropped even if it exceeds the delay bound, system will schedule real-time service first because it will be dropped and lose data if it exceeds the delay bound. After scheduling the real-time packets which are urgent, if there are still resources unused, system will schedule the non-real-time packets which are also urgent to try to meet their soft delay bound.

In this way, system does the first round scheduling to maintain the services' QoS requirement.

3.2.3 Opportunistic Scheduling

After the first stage of scheduling decision, if there are resources left unused, system will do the second stage scheduling to try to enhance the system performance and let services transmitted efficiently.

MSs in the system will measure the preamble information to detect the signal strength. Then they will report the downlink CINR information back to BS in the uplink subframe. Therefore, BS knows which burst profile should be used to transmit data to the specific MS. Based on the information, system knows how many bits could be transmitted in a resource unit. System also knows packet size which wants to be transmitted. Thus system can estimate how many resource units will be used if the system wants to transmit the specific packet.

$$N_o = \left\lceil \frac{Packet\ size}{estimated\ data\ amount\ per\ resource\ unit} \right\rceil \qquad (18)$$

In (18), N_o indicates how many resource units should be used to transmit the packet. If the value of N_o is less, it might means transmit the packet in this frame will be more efficient.

Why does here mention “might”? That is because the value N_o will be affected by two factors: packet size and estimated data amount per resource unit. It means the value of N_o is less may caused by smaller packet size, not for more estimated data amount per resource unit coming form better channel condition. If we just pick services with less N_o, sometimes it might be a wrong decision and system can not use resource efficiently.

To exclude the influence of packet size, system uses other parameters to help to figure out which packet to be transmitted is really efficient. First, set a parameter indicate how many resource units will be used if transmitting in the most robust burst profile. The parameter N_r denotes this situation (19).

$$N_r = \left\lceil \frac{data}{least\ data\ amount\ per\ resource\ unit} \right\rceil \qquad (19)$$

$$N_a = N_r - N_o \qquad (20)$$

Then use N_r minus N_o to get the parameter N_a (20). N_a means how many resource units can be saved additionally if system transmit the packet in this frame under the channel condition compared to the most robust burst profile. If N_a is larger, means it will save more resource units and transmission will be more efficient. In this way, we can exclude the influence of packet size and figure out which packet shall be transmitted if system likes to enhance the performance and efficiency.

If the algorithm stop right here, it only uses EDF concept combined with opportunistic scheduling concept to ensure QoS and enhance system performance. We named this algorithm as **EDFOP algorithm**. It means EDF algorithm combined with opportunistic scheduling concept directly.

However, sometimes if there are so many packets with similar delay bound, there are not enough resources to support them to transmit within their delay bound in the first stage scheduling. It means, maybe system can find out many packets which are all urgent and must be transmitted in the current frame. However, the resources in the current frame can't afford to transmit all these packets. Maybe only two packets can be transmitted in the frame then the frame is full, no resource left for other packets. We call this situation as "congestion".

The "congestion" is due to the similar delay bound among packets. In this situation, system will schedule other packets first maybe because of their transmission efficiency. But when the system figure out that those packets become urgent packets and try to schedule them, the resource is not enough to afford all of the packets transmission. If the situation occurs, the QoS guarantee will not easy be met. For real-time service, the packets will be dropped while they exceed their delay bound in the congestion situation. But for non-real-time service, the situation will be much more serious. That is because non-real-time packets will not be dropped if they exceed their soft delay bound. They will keep being transmitted. However, due to the soft delay bound will be accumulated. If the congestion occurs, system will try to transmit those urgent packets first even if those packets exceed their delay bound for long time. The following packets of the same user will have much shorter soft delay bound, even the negative value. Therefore it will affect the following packets' transmission seriously. Maybe the following packets can not meet their QoS requirement due to the accumulation from the previous packets.

Therefore, system must find some ways to avoid the congestion situation to maintain the QoS requirement. The straightforward EDFOP algorithm must be modified. What we modified is in the second stage of scheduling, hoping to separate packets delay bound to avoid the congestion.

In the second stage of scheduling, system will try to separate the delay bound between user packets but still transmits them in efficient way. Therefore, system divides packets into several groups according to their delay bound. Put packets with similar delay bound into the same group. System will still calculate the parameter N_a. With the information, pick packets with larger N_a in different groups in sequence. In this way, system separate packets to avoid the congestion due to similar delay bound and still can transmit packets in efficient way.

We describe the process in Figure 3-6. System separates packets into five groups according to their delay bound. Packets with similar delay bound will be assigned into the same group. Then system pick a packets with the largest N_a in group 1 which the delay bound are more urgent than other groups and schedule the packet first. Then if there are still available resources, pick the packet with the largest N_a in group 2. In this way, pick packet in different groups in sequence.

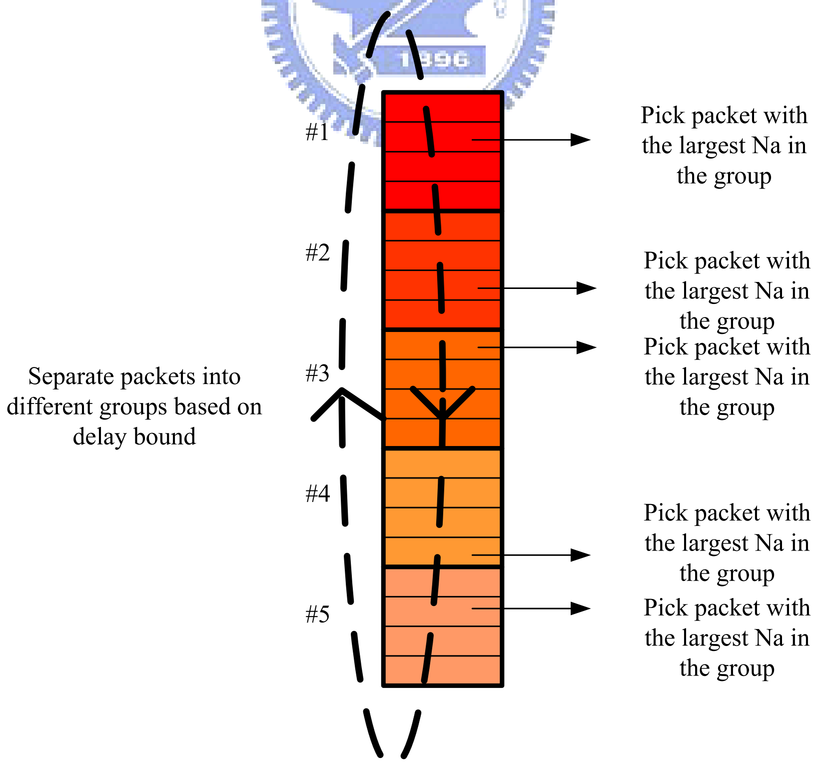

Figure 3-6 Packets separation to avoid congestion

After the modification, the algorithm is the proposed algorithm, named **"QoS-based opportunistic scheduling algorithm"**. It is because the algorithm uses the opportunistic scheduling concept but still focus on the QoS guarantee, give strict definition of QoS requirement.

3.2.4 QoS-based opportunistic scheduling algorithm

In this section, we introduce our proposed algorithm again and make a conclusion.

First of all, system must schedule PDUs which needed to be retransmitted first. Then system will do packet scheduling. This is a two-stage algorithm. In the first stage, system wants to make sure the QoS guarantee for packets. System will figure out if the packet is urgent or not? For urgent packet, system will schedule them according to their lifetime, meanings at what time the packet should be transmitted to make sure the QoS requirement, shorter lifetime will be schedule first. For those not urgent packets, system will observe their channel condition and hoping to transmit them in efficient way. After the first stage scheduling, if there are resources unused, system will do the second stage scheduling. In the second stage scheduling, system hopes to enhance system performance, like throughput. Therefore it will pick the packet which will be more efficient if it be transmitted. But to avoid the congestion that system can not afford enough bandwidth to users, system will divide packets into different groups according their delay bound. Then system picks packets which will be more efficient to transmit from each group and schedule them in sequence.

To achieve the functionality that the proposed algorithm expects, system must do cross-layer design to get enough information to support scheduling decision. It needs the traffic information from upper layer. It needs packet classifier in CS sublayer. It also needs channel feedback information and ACK/NACK message from physical layer. The whole structure is shown in Figure 3-7.

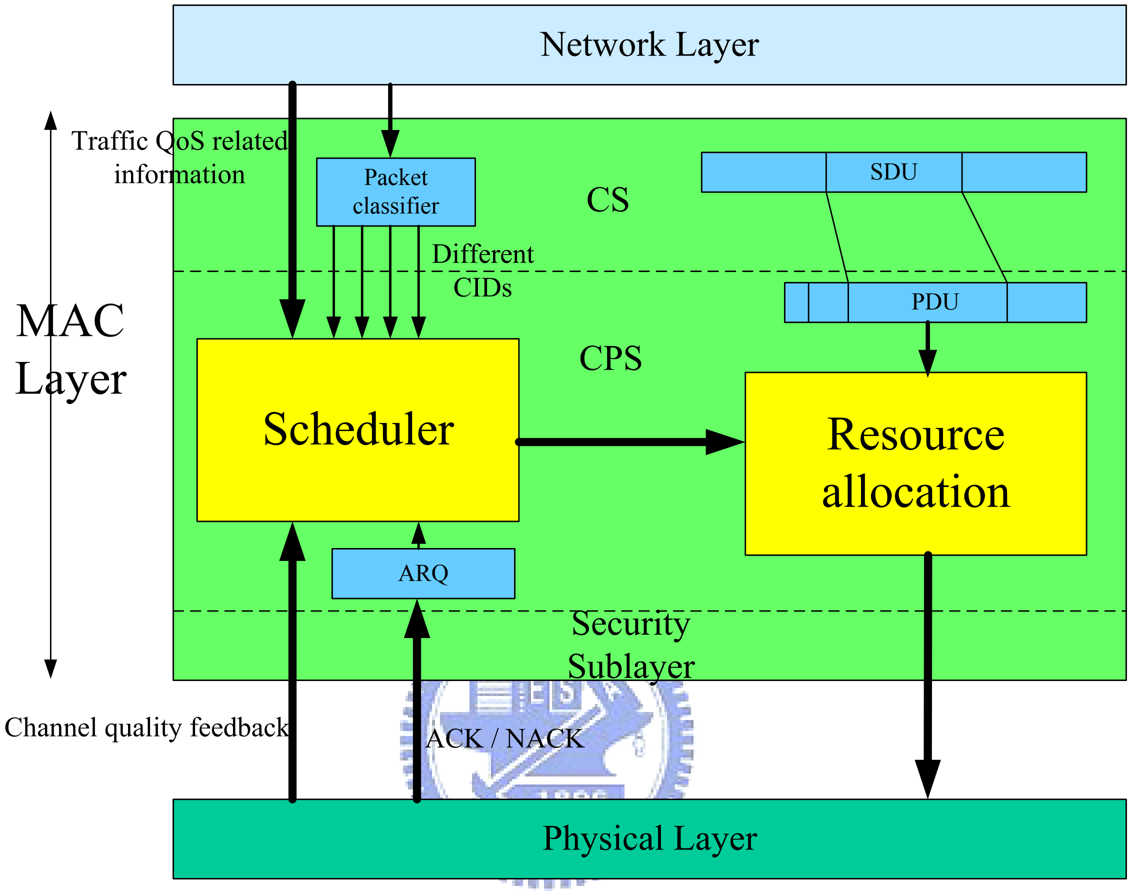

Figure 3-7 Cross-layer scheduling structure

In Figure 3-8 shows the proposed algorithm flow chart:

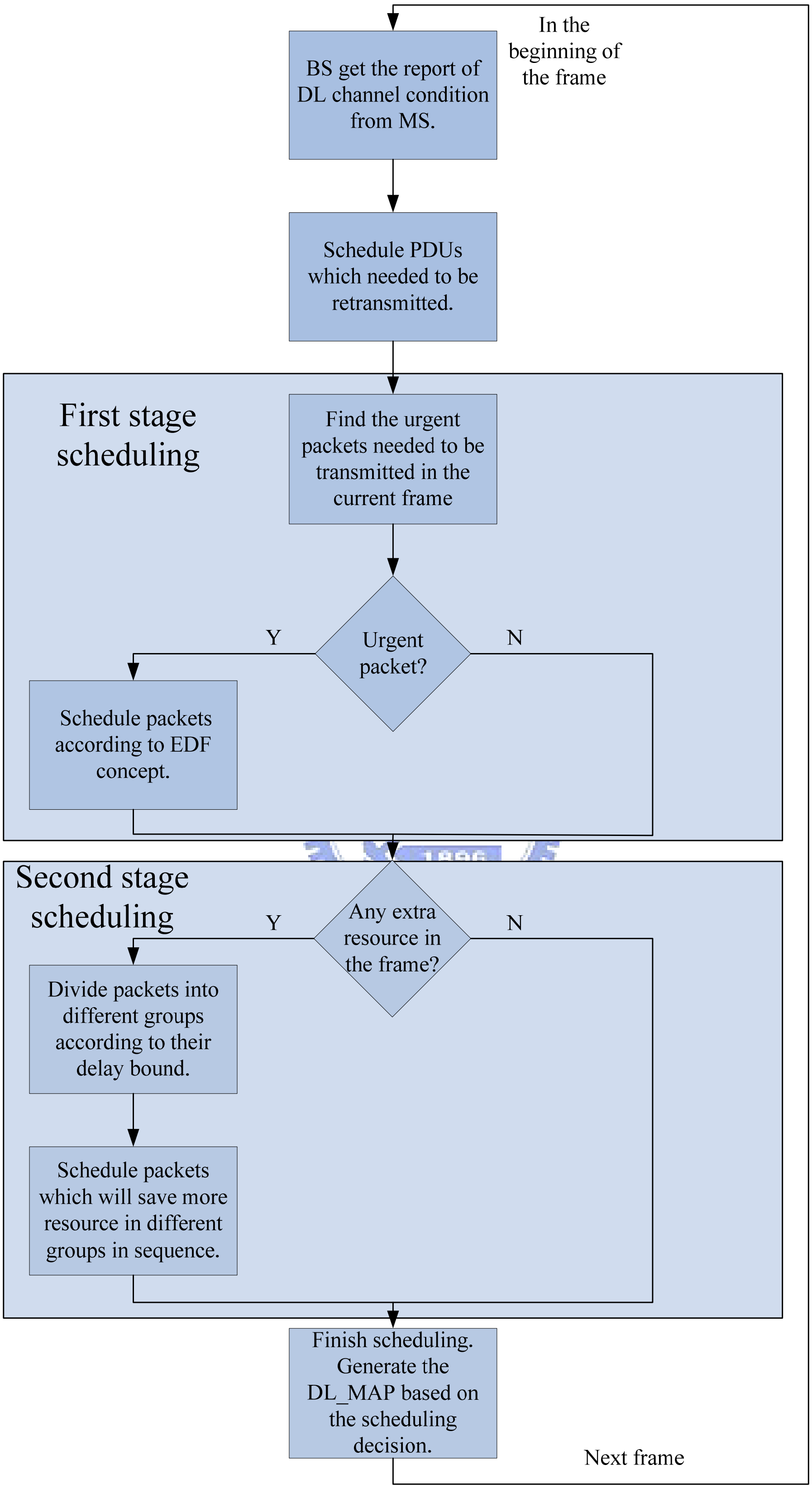

Figure 3-8 The QoS-based opportunistic algorithm flow chart

Chapter 4
Simulation Setup

In this chapter, the IEEE 802.16e system level simulation platform will be described. The details of system architecture and simulation parameters are going to be presented. Then, the Link Budget, such as path loss and shadow fading, is set to be suitable for IEEE 802.16e standard. The setting of basic radio resource managements, such as power control, rate control (AMC), scheduling controls, and handoff method, is showed in this chapter. Finally, the traffic models of simulation are introduced. The setting and the structure of the simulation platform is referred from [14].

4.1 The Architecture of Mobility Platform

In the simulation of a mobile system, the interference from other cells is an important element that would affect the overall performance of single cell. This effect need to be considered into the simulation. When the number of simulation cells increase, it will cause high load of the simulation time and computation. Between these two tradeoffs of accurate simulation and computing cost, we consider two-tier interference cells. According to approximate the cell coverage with a hexagon, we consider nineteen cells in our simulation as shown in Figure 4-1. From Figure 4-1, we can see that only the central cell completely has two-tier interference cells, the other cells can not find out the symmetric two-tier interference cells. Even if we use nineteen cells to simulate, only the statistic simulation value of the central cell can be referred, causing a lower simulated efficiency. Hence, we adopt a wrap around technique as shown in Figure 4-2. This technique can make any of nineteen cells owns complete two-tier interference cells. The main ideal is the lacks of any two-tier interference

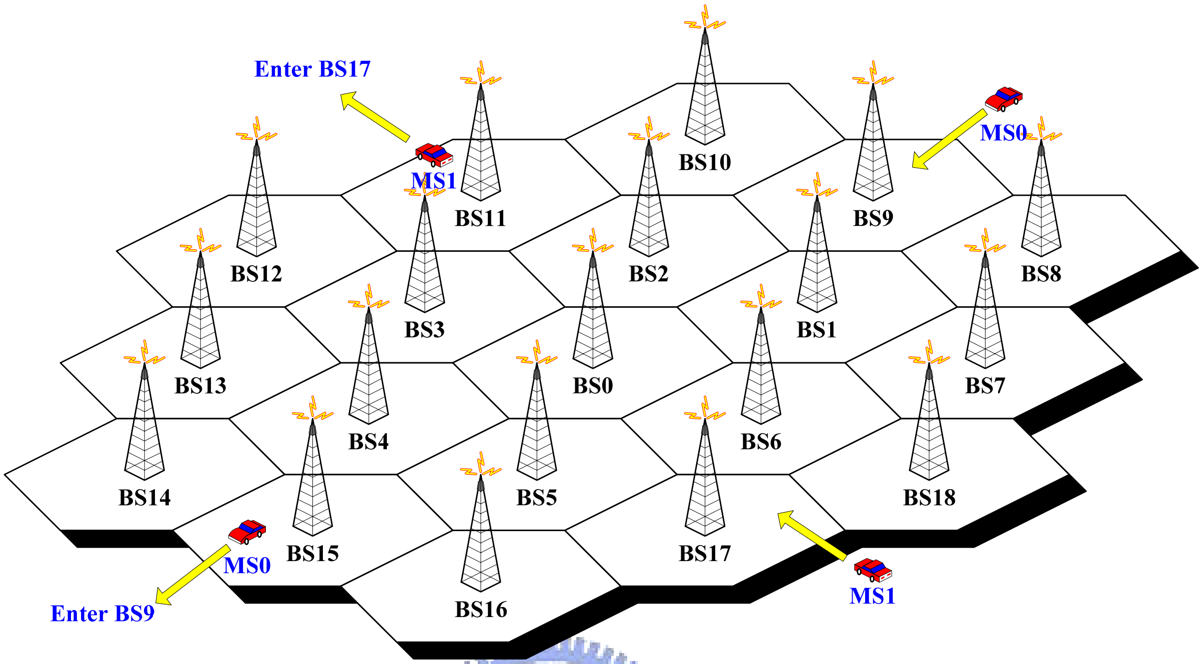

Figure 4-1 Cell Structure of System Simulation

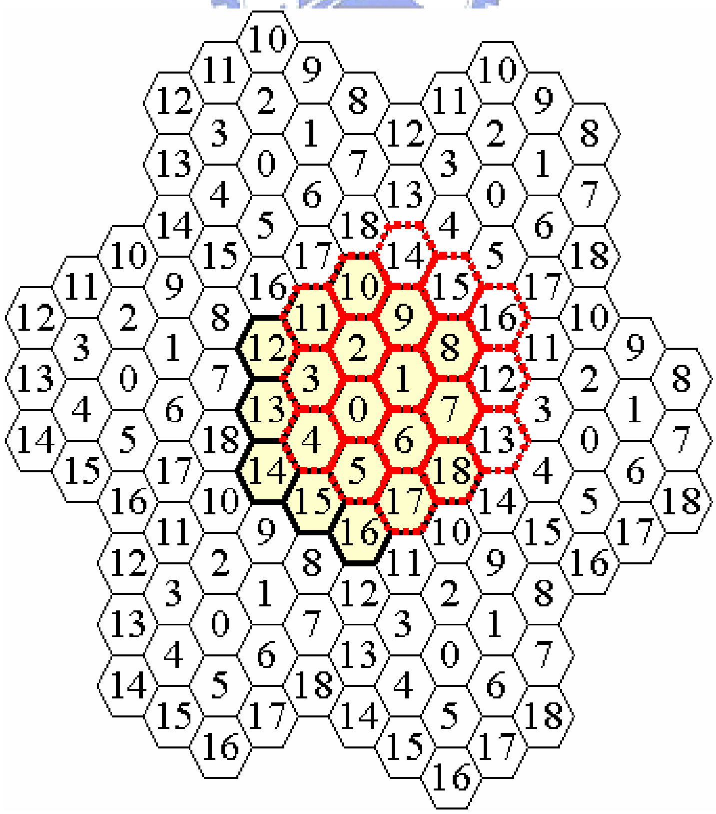

Figure 4-2 Example of Wrap Around

cells of a specific cell are copies from the simulated cells which are besides the two-tier interference cells of the specific cell. Through the clever arrangement, any cell has complete, symmetric, and different two-tier interference cells. Because the cell owns whole interference after wrap around, the statistic simulation value of nineteen cells would be meaningful.

The cell radius which we set is 1 km [15]. This approximate cell coverage is a result from a plan of Link Budget. The total cell bandwidth that we choose is 10 MHz. In our simulation platform, a cell is divided into three sectors as shown in Figure 4-3. Each sector has the different antenna direction and a regular pattern of deployment. The sector architecture in 802.16e system can reduce the transmission power of BS antenna and intercell interference. But it still has a small part of intercell interference in different sectors of distinct BSs due to subcarrier permutation. This characteristic is very difficult to simulate, so we assume a sector would produce interference to other sectors which have the same direction.

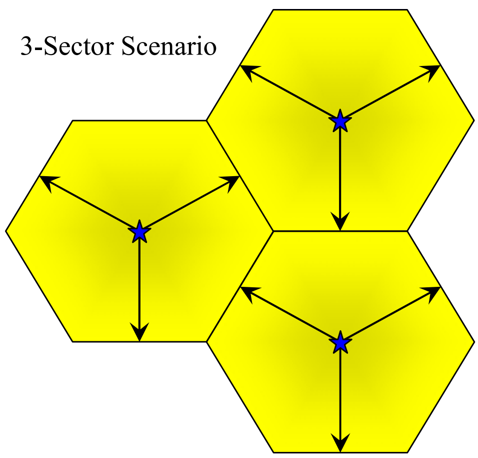

Figure 4-3 Example of sector deployment

In our simulation platform, the setting of antenna pattern uses the 3GPP's model [16] as shown in (21).

$$A(\theta) = -\min[12(\frac{\theta}{\theta_{3dB}})^2, A_m] where -180^\circ \le \theta \le 180^\circ \quad (21)$$

, where θ_{3dB} is $70°$ and A_m is $20dB$.

The cell's frequency reuse factor in our simulation platform will be set to 7 due to the link performance curve we adapted. In this setting, system can get rid of the BPSK modulation coding scheme which is not defined in 802.16e. Frequency reuse factor 7 need seven times bandwidth, 70 MHz. Figure 4-4 shows the deployment of frequency reuse factor.

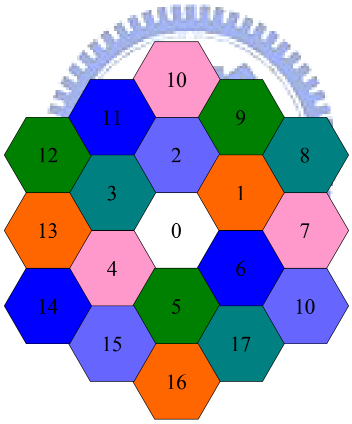

Figure 4-4 Example of the Deployment of Frequency Reuse Factor

4.2 The Architecture of Frame Transmission

In this thesis, we focus on the downlink transmission and use the OFDMA technique with TDD mode. The IEEE 802.16 standard can support an asymmetric downlink and uplink transmission of TDD mode, which adjusts the ratio according to the traffic loading of downlink and uplink transmission. In our simulation, we use a simple assumption. We assume the downlink and uplink transmission have a ratio of equality and they use a half frame, respectively. In 2.1.1, we introduce the time duration of a frame. In our simulation, the frame length we used is 5 ms. The frame structure we used is 1024-FFT OFDMA downlink carrier allocations –PUSC mode defined in the standard. The carrier distribution is shown in Table 4-1. In the 1024 subcarriers, only 720 subcarriers carry data information and other subcarriers are used for guard band, pilot and dc subcarrier.

Table 4-1 1024-FFT OFDMA downlink carrier allocations with PUSC

Subcarrier types	**number**
Total subcarriers	1024
DC subcarriers	1
Guard subcarriers	92 (Left), 91 (Right)
Sub-channels	30
Data sub-carriers within each sub-channel	24

The length and number of OFDMA symbols are defined according to WiMAX Forum [17] [18]. For 10MHz bandwidth and 1024 FFT sizes, the symbol period should be 102.9 μs and there will be 48 OFDMA symbols per frame. Assumed that we divide downlink and uplink equally, there will be 24 symbols for downlink transmission. While we use PUSC, two symbols form one slot, there will be 12 slots in time domain. And there will be 10 subchannels per sector (30/3 = 10). Hence, there will be 120 resource units per frame to transmit data and message.

4.3 Link Budget

The link budget settings of downlink transmission in our simulation are as far as possible to match the IEEE 802.16e real environment. In IEEE S802.16e-03/23 document [16], it makes deployment scenario assumptions for 802.16e, like Table 4-2. In our simulation, we adopt the outdoor vehicular scenario, which the BS transmitted power is 46 dBm, the BS antenna gain is 17 dBi, the SS (MS) antenna gain is 3dBi on the downlink transmission. The BS back off which is used to avoid the RF circuit working in the non-linear region due to the peak-to-average power ratio (PAPR) of OFDM system is 5 dB. The common usage value of thermal noise density is -173.93 dB / Hz. The receiver noise figure of MSs is 9dB [19].

Table 4-2 Link Budget Parameter of 802.16e system

Scenario / Parameter	Indoor	Outdoor to indoor	Outdoor vehicular
BS Tx power	27 dBm (0.5 W)	36 dBm (4 W)	46 dBm (40 W)
MS Tx power	17 dBm	17 dBm	27 dBm
BS ant gain	6 dBi	17 dBi	17 dBi
MS ant gain	0 dBi	0 dBi	3 dBi
BS ant height		15 m	30 m

In wireless channel, the transmitted signals will suffer the fading effect, which can change the original signals. The fading can be divided into three types: pathloss, shadow fading, and fast fading (multipath and doppler effect). In our simulation, we only consider the pathloss and shadow fading. The fast fading will be used in the future work. The pathloss model is used to present the signal strength decreases with increasing distance between transmitter and receiver. In Winner D5.4 document [20], it provides several pathloss models, such as Table 4-3. Because the cell radius is 1 km in our simulation and the signal transmission in 2~11 GHz is non-line of sight (NLOS), the C2 scenario is more suitable and we use it in our simulation.

Table 4-3 Pathloss Model Scenarios

Scenario		Path-loss [dB]	Shadow fading standard dev.	Applicability range
A1	LOS	$18.7\log_{10}(d[\text{m}])+46.8$	$\sigma = 3.1$ dB	$3\text{m} < d < 100\text{m}$
A1	NLOS	$PL(d) = 36.8\log_{10}(d[\text{m}])+38.8$	$\sigma = 3.5$ dB	$3\text{m} < d < 100\text{m}$
B1	LOS	$22.7\log_{10}(d[\text{m}])+41.0$	$\sigma = 2.3$ dB	$10\text{m} < d < 650\text{m}$
B1	NLOS	$0.096\, d_1[\text{m}]+65+$ $(2.8-0.024\, d_1[\text{m}])\log_{10}(d_2[\text{m}])$	$\sigma = 3.1$ dB	$10\text{m} < d_1 < 550\text{m}$ $w/2 < d_2 < 450\text{m}$
C2	NLOS	$35.0\log_{10}(d[\text{m}])+31.5$	$\sigma = 8$ dB	$50\text{m} < d < 5\text{km}$
D1	LOS	$PL(d) = 21.5\log_{10}(d[\text{m}])+44.6$ $= 40.0\log_{10}(d/d_{BP})+44.6+$ $21.5\log_{10}(d_{BP})$	$\sigma = 3.5$ dB $\sigma = 6.0$ dB	$30\text{m} < d < d_{BP}$ $d_{BP} < d < 10\text{km}$
D1	NLOS	$PL(d) = 25.1\log_{10}(d[\text{m}])+55.8$	$\sigma = 8.0$ dB	$30\text{m} < d < 10\text{km}$

The main reason forms shadow fading is from the shelters, like buildings, or mountain, on the signal transmitted path. According to the test result of the real wireless environment, we can know the variant of shadow fading is a log-normal distribution statistically. So, we can use the log-normal distribution to produce the shadow fading effect. The standard deviation of this distribution is based on the simulation environment. In our simulation, we use 8 dB for the standard deviation [20]. When the user is fixed, the shadow fading effect will not alter. On the other hand, the shadow fading changes with different locations at the mobile user. In the different simulation points, we can use the log-normal distribution to produce a value for the shadow fading, respectively. But this method has a problem, the time between two neighbor simulation points is very small so that the mobile user location will not change very obvious even at high mobile speed. It means the variance the shadow fading will not be large and have a correlated relationship between two neighbor time points. Hence, we use the concept of a correlation model, called Gudmundson's correlation model [21], in (22).

$$\rho(\Delta x)=e^{-\frac{|\Delta x|}{d_{cor}}\ln 2} \qquad (22)$$

,where ρ is the auto-correlation constant between two simulation sample points, Δx is the distance of two sample points and is a function of sampling times between them, sampling duration, and user speed. The d_{cor} is de-correlation distance and the values in the suburban macro, urban macro, and urban micro environments are 200m, 50m, and 5m, respectively. In our simulation, we use 5m as our parameter. Using log-normal distribution and correlation model, the simulation can get more actual shadow fading.

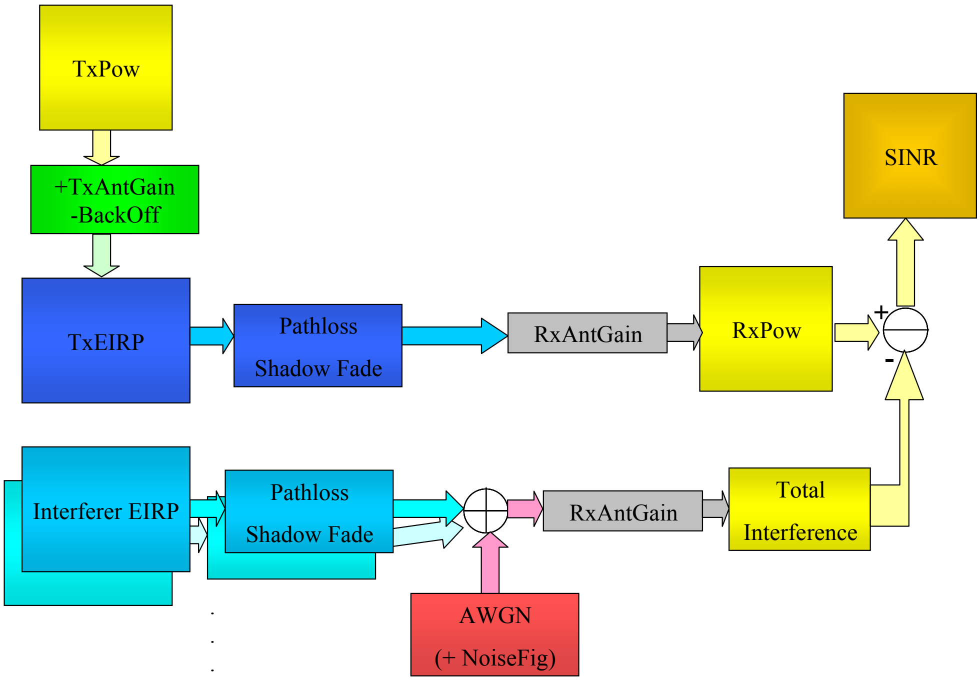

Figure 4-5 Example of SINR computation

In Figure 4-5, we present the flow of signal-to-interference and noise ratio (SINR) computation. The OFDMA technique uses the multiple carrier to transmit signal and we should compute carrier-to-interference and noise ratio (CINR), not SINR. But the MSs of 802.16 system with PUSC or FUSC mode only compute and report the sum of received CINR per carrier to BS, not individual CINR. Hence, the SINR and CINR are the same under these conditions. Finally, the mobility model we use is like below. The MS speed is 30 km/hr. Probability to change direction is 0.2 when position update. Maximum angle for direction update is 45^{o}.

4.4 Basic Radio Resource Management

The purpose of radio resource managements is to raise the efficiency and reliability of wireless transmission. In our performance analysis, we will use the basic radio resource management method as follow.

- **Power Control:** We assume the BS use maximum and fixed power to transmit signals. The power of per subcarrier is the same.
- **Rate Control (AMC):** The adaptive modulation and coding scheme is a major method to keep the quality of wireless transmission. The IEEE 802.16e standard supports a variety of modulation and coding scheme, which we introduce in 2.1.4. In our simulation, we only get the link performances of QPSK, 16-QAM, and 64-QAM modulation schemes, so we use these modulation schemes. The coding scheme is used to correct errors in the receiver and we use convolution code (CC) with 1/2 code rate. From 2.1.4 and 4.2, we can get one slot using QPSK, 16-QAM, and 64-QAM with CC 1/2 can carry 48, 96, 144 bits, respectively.
- **Channel assignment:** The OFDMA frame structure has two dimensions, the slot with two OFDMA symbols and one subchannel, for channel assignment. In our simulation, we obey the definition of 802.16e standard that we introduce in 2.1.2. The basic principles are to segment the data after the modulation block into blocks sized to fit into one OFDMA slot, and map the slots in the subchannels with higher priority than that in the OFDMA symbols. In other words, the data mapping method is frequency first.
- **Subcarrier permutation:** In our platform, we use the distributed subcarrier permutation. If we use the permutation formula and series to implement the permutation, it is too complex and low efficient in simulation computation. So, we use the statistic method to simulate the permutation effect. The interference of one slot produced by the other cells will be dispersed to all subframe.
- **Scheduling method:** The scheduling algorithm is critical in the thesis. Besides our proposed scheduling algorithm, we build some baseline algorithm to do the comparison. In our simulation, we choose round robin (RR), proportional fair

(PF), MaxCINR, early deadline first (EDF), which have been discussed in 2.3. Besides those algorithms, we also build the proposed algorithm in 3.2.4 and the EDFOP algorithm which is similar to the proposed one but without the delay bound separation mechanism in 3.2.3. There will be divided into six groups according to the delay bound in the proposed QoS-based opportunistic scheduling algorithm. And each group will cover 30 ms.

- **Handoff method:** In this thesis, handoff is not a weight-bearing point. So, we use the simplest method: hard handoff. This method is "Break-Before-Make".
- **PDU segmentation:** In our simulation, we will generate SDU according to the traffic model. Then system will segment the SDU into PDU to achieve the target error rate in (23).

$$PER_{target} = 1-(1-BER)^{bits} \qquad (23)$$

BER is the bit error rate which is available according to channel condition feedback. Target packet error rate is different from application. Based on those information, system can calculate the number of bits per PDU and segment the SDU according to the result. However, the simulation set an upper bound to PDU size, 180 byte. That's because sometimes retransmit a large PDU might block the opportunity for real-time service transmission. That will cause packets drop hence try to limit the upper bound to avoid the situation occurring.

- **ARQ retransmission:** In our simulation, we only implement the ARQ retransmission and don't use the HARQ. When the PDU is error, the ARQ retransmission will work. However, if system finds out that the retransmission will exceed real-time service packet's delay bound, it will drop the packet because it is meaningless to retransmit the PDU. We will set the retransmission times based on the application. For streaming service and non-real-time service, the retransmission time will be set to three. For voice service, the retransmission time will be set to one. That's because the different packet size will cause different error rate. For voice service, which with smaller packet sizes, will have lower error rate than others. Hence it is not necessary to set retransmission time as others.

4.5 Traffic Models

In IEEE 802.16e standard, the downlink data traffics are divided into four QoS classes, such as real-time CBR, real-time VBR, delay-tolerant VBR, and BE. The details are described in 2.2.4. In simulation, we build FTP service to stand for delay-tolerant VBR service, voice and streaming service for real-time VBR. The FTP traffic model adopts 3GPP2 model [22] as shown in Table 4-4. Both of the non-real-time services' minimum reserved rate will be set to 60kbps according to [23]. The VoIP traffic model uses G729-1 codec [24] as shown in Table 4-5. The FTP services use TCP/IP protocol to transmit, so the FTP packet needs to add 20 bytes TCP header and 20 bytes IP header. The VoIP services user RTP/UDP/IP protocol to transmit. The VoIP packet must add 12 bytes RTP header, 8 bytes UDP header, and 20 bytes IP header.

Table 4-4 FTP Traffic Model

File size (S)	Truncated Lognormal	A=0.35, u =14.45, M=5m
Reading time (D_{pc})	Exponential	L=1/180

Table 4-5 VoIP Traffic Model

codec	Framesize(byte)	Interval(ms)	Rate(bps)	Delay bound(ms)
G729-1	20.0	20.0	8k	20.0

Finally, we use Table 4-6 to summarize this chapter and present the arrangement of the parameter setting in our simulation platform.

Table 4-6 The Parameter Setting in Simulation Platform

Parameters	**Value/Comment**
Cell layout	Hexagonal grid, 19 cells (wrap around)
Sectors per cell	3
Frequency reuse factor	1x7
Available bandwidth	70 MHz in 1x7 reuse
Antenna pattern	70° with 20 dB front-to-back ratio, according to [16]
Cell radius	1 km
Transmitter/Receiver	Downlink (from BS to MSs)
Duplex	TDD mode
DL/UL subframe ratio	1:1
Frame length	5ms, according to [1]
Frame structure	1024-FFT OFDMA downlink carrier allocations with PUSC, according to [1]
OFDMA symbol length	102.9 μs, according to [17] [18]
OFDMA symbols per slot	2 symbols
BS Tx power	46dBm (40 Watt), according to [15]

BS Antenna gain	17 dBi, according to [15]
BS back off	5 dB, according to [20]
Thermal Noise Density	-173.93 dB/Hz, according to [20]
MS Noise Figure	9dB, according to [18]
MS Antenna gain	3 dBi, according to [15]
Pathloss model	35.0log(d[m])+31.5, 50m<d<5km, according to [20]
Shadow fading model	Log-normal distribution with STD=8dB and Gudmundson's correlation model, according to [21]
Mobility model	MS speed : 30 km/hr Probability to change direction : 0.2 Max. angle for direction update : 45^{o}
BS Power contol	Max power
AMC	QPSK+CC 1/2, 16-QAM+CC 1/2, 64-QAM+CC 1/2, according to [2]
Channel assignment	Frequency first, according to [2]
Scheduling control	Round Robin (RR), Proportional Fair (PF) Max CINR (MC), Early Deadline First (EDF) EDFOP, QoS-based Opportunistic
Handoff	Hard handoff
Traffic model	FTP [22][23] VoIP [24]

Chapter 5

Simulation Results

In this chapter, we will show the simulation results for with different scheduling algorithms to see the advantage of the proposed algorithm – QoS-based opportunistic scheduling algorithm. We will focus on some factors such as throughput, packet delay rate for non-real-time services, packet drop rate for real-time services and AMC usage percentage, etc… We use those results to prove that the proposed algorithm could ensure the QoS requirement and enhance the system performance as well.

The goals of designing the algorithm are to maintain the QoS requirement, especially in the mix-traffic, and to enhance the system performance. In the following simulation results, it will show the proposed algorithm will really outperform others excluding MaxCINR in throughput. Then we will observe if the QoS be ensured in the proposed algorithm. Does it perform better than the one without strict QoS guarantee, the PF algorithm?

5.1 Throughput Performance

In this thesis, observing the non-real-time service throughput as the indication to see if the algorithms enlarge the system performance. Besides the plots of throughput, it also shows the percentage of the AMC usage to see how different algorithms choose users to transmit. In this section will show the MAC throughput of non-real-time service under single traffic and mix-traffic situation.

5.1.1 MAC Throughput and AMC Usage Comparison for FTP Users Only under Single-traffic Environment

MAC throughput indicates the real data that have been transmitted. It ignores the header and the retransmitted data and reflects the real data which is really useful.

In figure 5-1 shows the MAC throughput under different scheduling algorithms. We can find out that the MaxCINR algorithm could provide the best throughput performance and great user diversity effect, MAC throughput increase as users increase. For EDFOP algorithm and QoS-based opportunistic algorithm, both of them consider the channel condition so the user diversity effect also reflects while FTP users are less than 25. However, both of them try to maintain the QoS requirement for services. Hence they will allocate packets to transmit to meet the QoS requirement even if the user without good channel condition. Therefore, if the traffic loading is getting heavier, they might support more users with poor channel quality to meet QoS requirement. That's why they can't achieve the throughput performance as MaxCINR. However, they still perform better than other scheduling algorithm. PF algorithm considers both channel condition and fairness. However the fairness factor will limit the user diversity effect and make it performs more poor in throughput than MaxCINR, EDFOP and QoS-based opportunistic algorithm, but still perform better than RR and EDF. For RR and EDF, both of them don't take channel condition into consideration. Therefore the throughput performance is not as good as others.

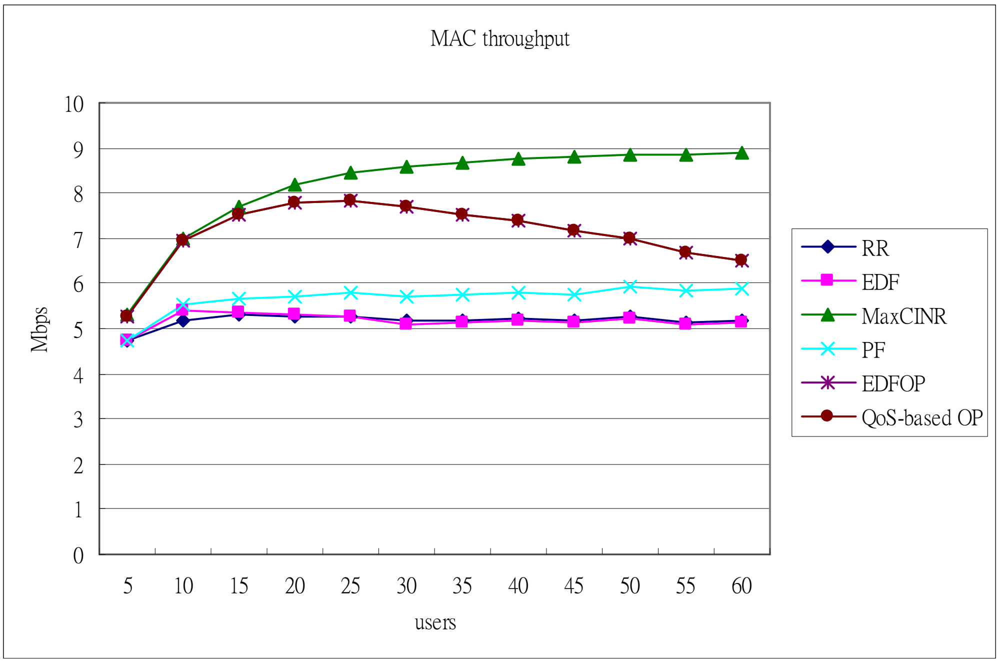

Figure 5-1 MAC throughput for FTP service

From Figure 5-2 to Figure 5-7 show the AMC usage percentage for different modulation coding scheme under different scheduling algorithms. The usage of modulation coding scheme will directly reflect why the MAC throughput varies from different algorithms.

For RR and EDF algorithm, the AMC percentage remains similar while the users increase from 5 to 60. So do the throughput performance they showed before. That is because they schedule packets regardless channel condition. For PF algorithm, usage of 64QAM will have more percentage and usage of QPSK will have much less percentage than it did in RR and EDF because the scheduling algorithm will take channel condition into consideration. Even if the PF algorithm take fairness into consideration, the packets belonged to good users still have higher priority than bad users. However, due to the fairness consideration, it can't take advantage of user diversity concept completely.

The one which fully uses user diversity concept is MaxCINR algorithm. We can figure out that while there are more users in the system, the more efficient modulation coding scheme is used more. That is because the modulation coding scheme is directly related to channel condition. Therefore it can achieve the best throughput performance.

For EDFOP and QoS-based opportunistic algorithm, they use users diversity concept but try to maintain services' QoS requirement at the same time. So, while the traffic load is not so heavy, less than 25, they will use more efficient modulation coding scheme to transmit packets. However, while the traffic load is getting heavier, to maintain the QoS requirement, they have to transmit packets even if they belonged to users have poor channel condition. Therefore the usage of modulation coding scheme which is more robust, such as QPSK, will be increasing. That causes the throughput will decrease while too many users in the system. However, in this way it can ensure the QoS requirement of users, which will be showed in the next section, but still have better efficiency excluding MaxCINR, which can not ensure QoS at all.

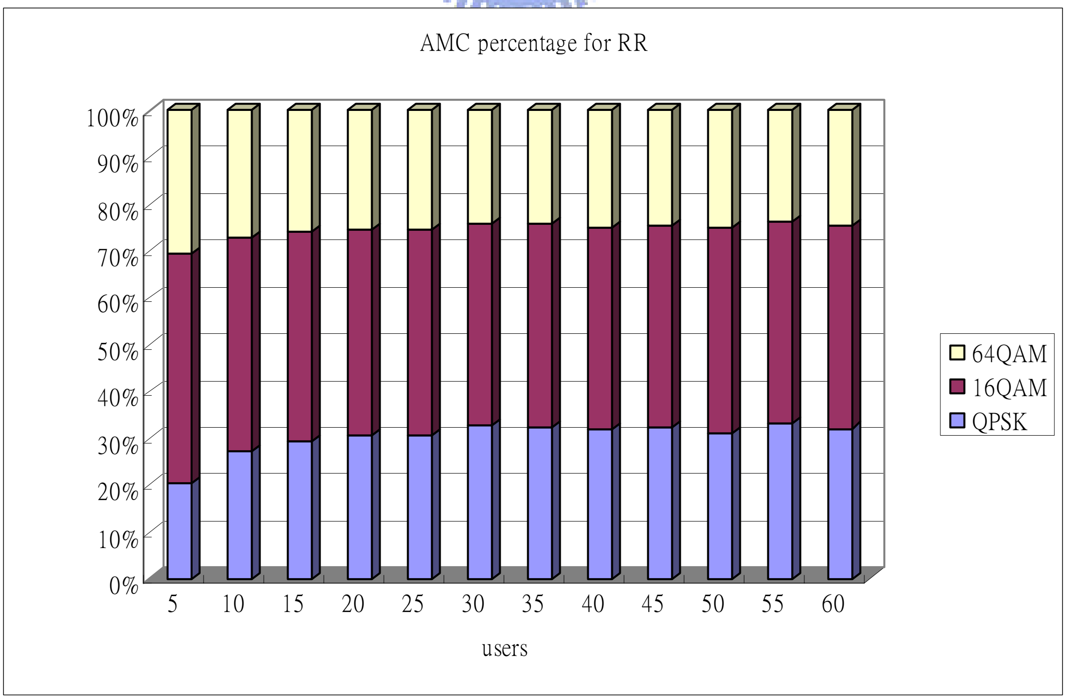

Figure 5-2 AMC usage percentage for RR

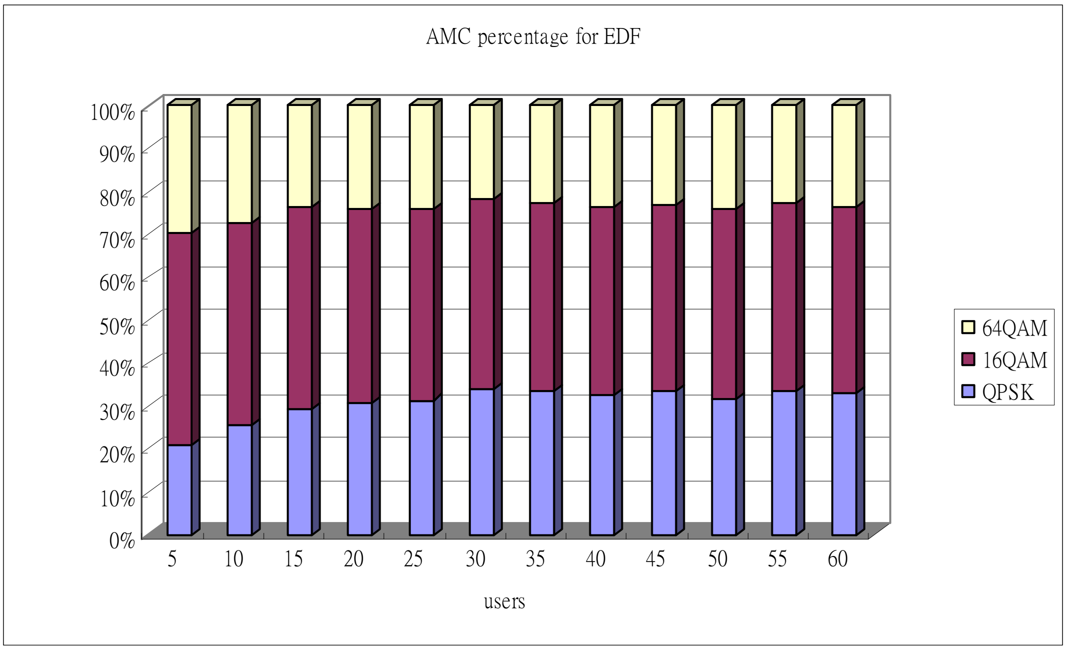

Figure 5-3 AMC usage percentage for EDF

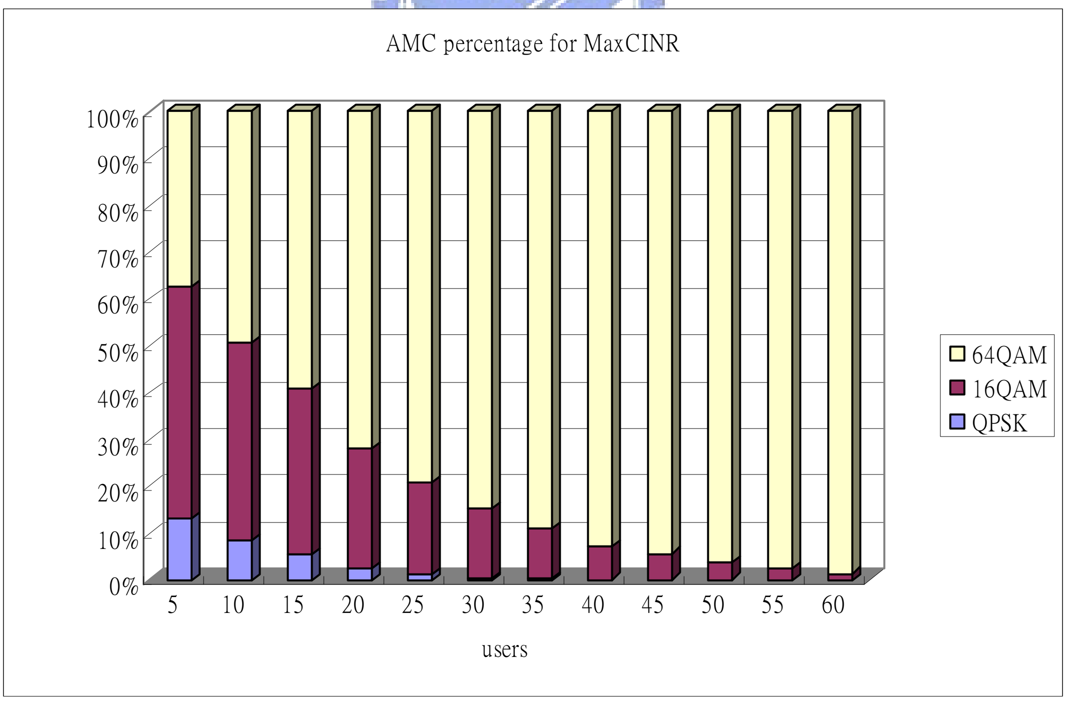

Figure 5-4 AMC usage percentage for MaxCINR

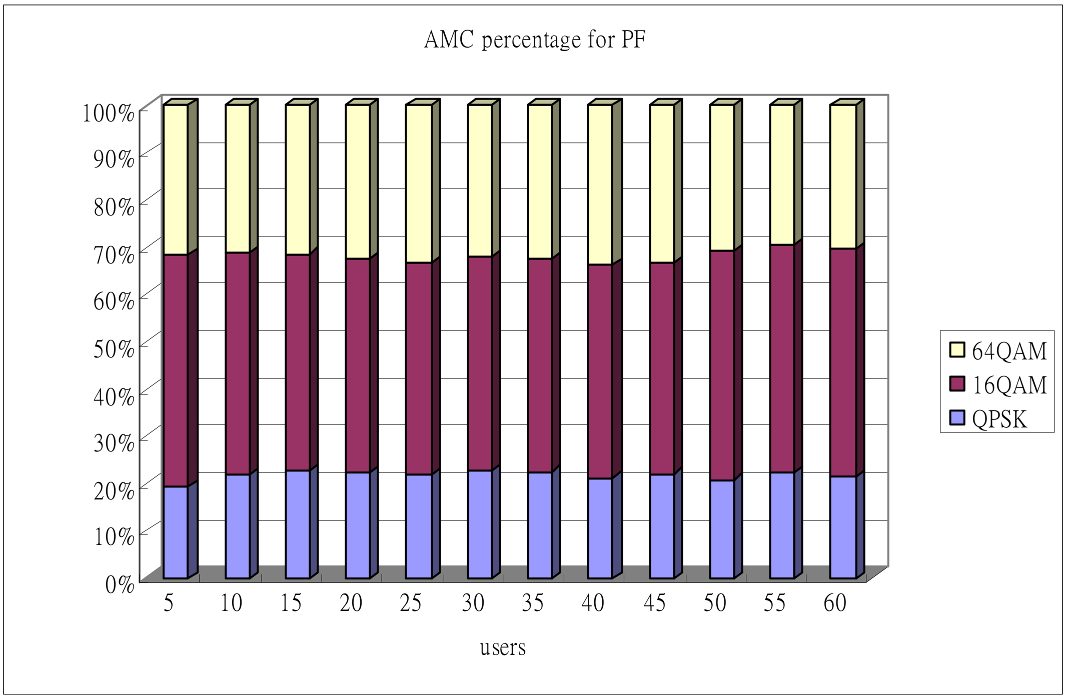

Figure 5-5 AMC usage percentage for PF

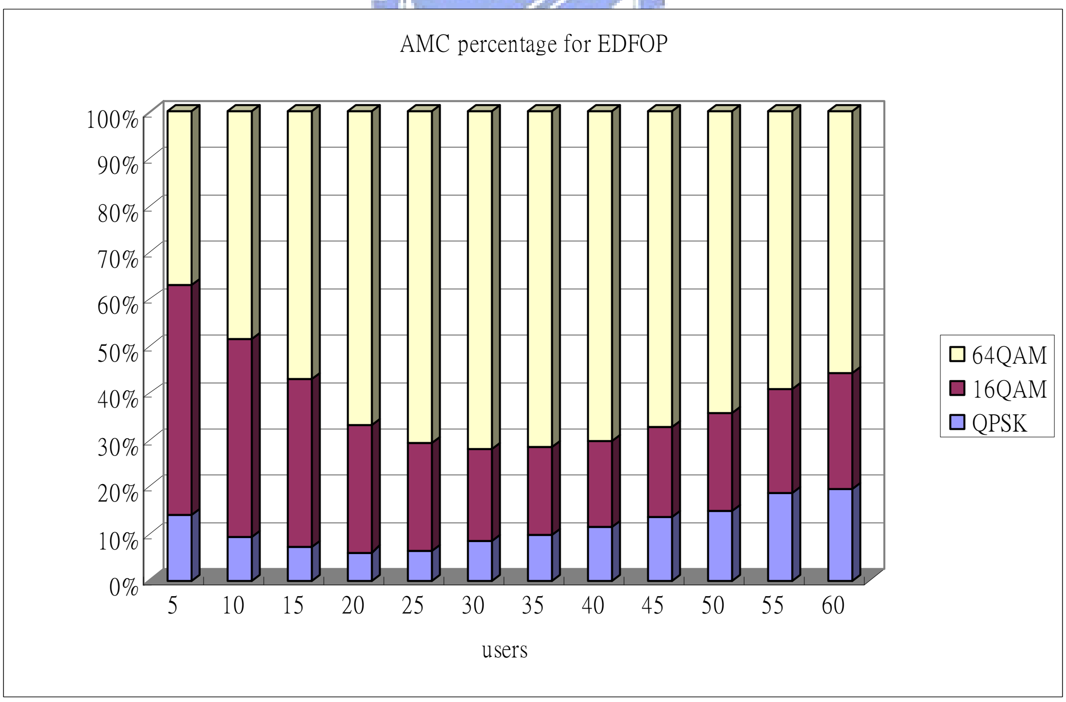

Figure 5-6 AMC usage percentage for EDFOP

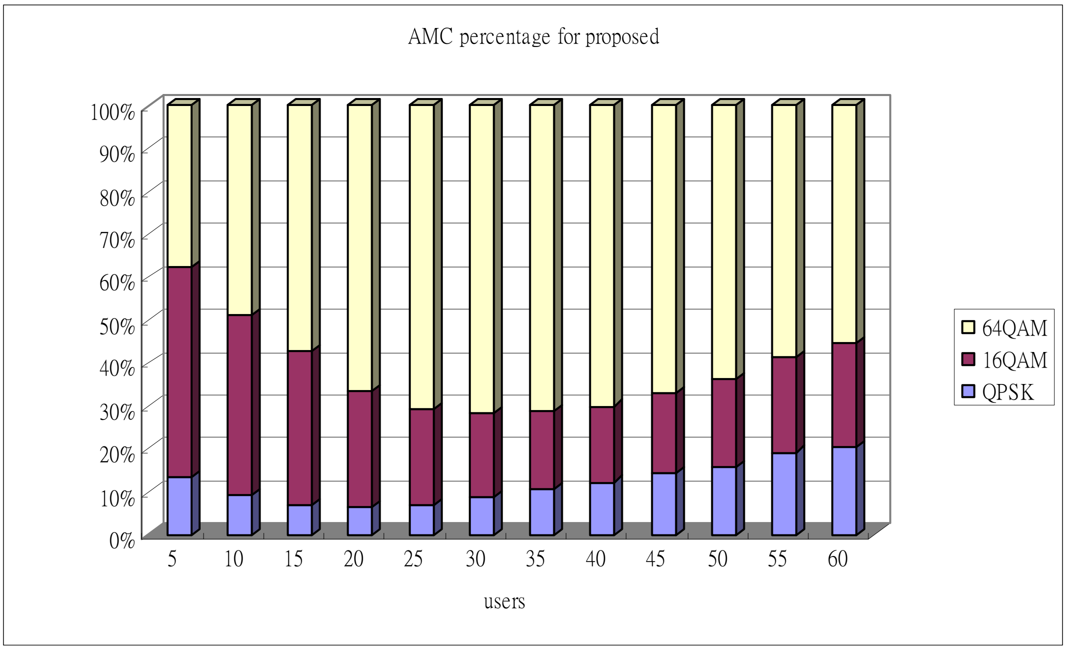

Figure 5-7 AMC usage percentage for QoS-based Opportunistic

5.1.2 MAC Throughput for FTP Users under Mixed-traffic Environment

Hoping to achieve higher throughput is the target of system design. In this way, the system may use the bandwidth more efficient even in the mix-traffic situation. From figure 5-9 to figure 5-11 show the MAC throughput for non-real-time service when there are 10, 20, and 30 FTP users with the increasing number of VoIP users in the system. The simulation results show that the non-real-time services MAC throughput will decrease as the number of real-time services increase. It is reasonable because transmission of real-time service will take the resource of transmission opportunity from non-real-time service. Although the throughput performance decrease, the influence of different scheduler design is still obvious.

For MaxCINR algorithm, which fully adopts user diversity concept, have the best non-real-time service throughput in all because it never care the QoS of users.

The real-time service will not have higher priority than non-real-time. There is only slightly decreasing in MaxCINR in throughput performance. The EDFOP and QoS-based opportunistic algorithm consider the channel condition and transmission efficiency, transmit in the best opportunity, outperform than PF, RR, and EDF in a big margin while there are not so many real-time users in the system. However, to maintain the QoS requirement, both of real-time and non-real-time users, it forces system unable to adopt user diversity concept but to transmit specific user in urgent to meet the QoS requirement. Therefore the non-real-time service throughput might keep decreasing and less than some algorithm while traffic load becomes heavier, because it allocates more resource to urgent packets. For example, when there are 10 FTP users, the throughput will be less than RR after the VoIP users exceed 80. When there are 20 FTP users, the throughput will be less than RR after the VoIP users exceed 70. And in 30 FTP users, the situation happens while there are 60 VoIP users. While the VoIP users keep increasing, the throughput will even close to what EDF algorithm performs. That is because the EDFOP and QoS-based opportunistic algorithm will use the strict definition to make sure the QoS guarantee. Therefore, they will easily give the transmission opportunity to those which are urgent to transmission. Hence it will sacrifice the opportunity to enlarge system throughput. However, it seems the two algorithms are adaptive to environment. While the traffic load is light, it will try to maximize the system throughput. However, while the traffic load is getting heavier, the goal of algorithms will turn to maintain the QoS requirement. We will see the evidences that they will maintain QoS requirement in the following section. As to PF algorithm, is perform better than RR and EDF scheduling algorithm while there are not so many real-time service users. When there are more and more real-time service users in the system, the throughput of non-real-time service will decrease, and might become even less than RR algorithm. That is because PF will give more resource to real-time service users than RR will. Because real-time service, in this case, the VoIP service, the data size will be much less than non-real-time service, FTP. Therefore, to calculate the average data, real-time service will have much less data which have been transmitted than non-real-time service did. Hence, real-time service easily gets the transmission opportunity over non-real-time service. The resource will be assigned to real-time service in the PF algorithm more easily. It causes throughput of non-real-time service degrades much faster than RR algorithm. As to EDF scheduling algorithm, it perform worst in throughput due to it will assign resource to real-time-service first and easily due to the much shorter delay bound, and it doesn't take channel condition into consideration at all.

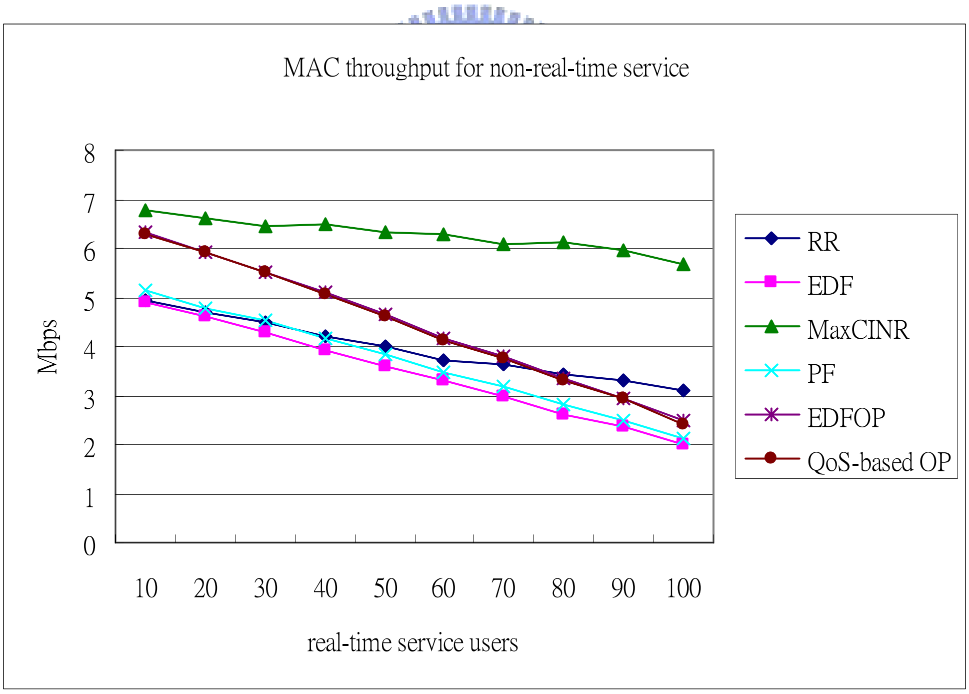

Figure 5-8 FTP throughput under fixed FTP users 10

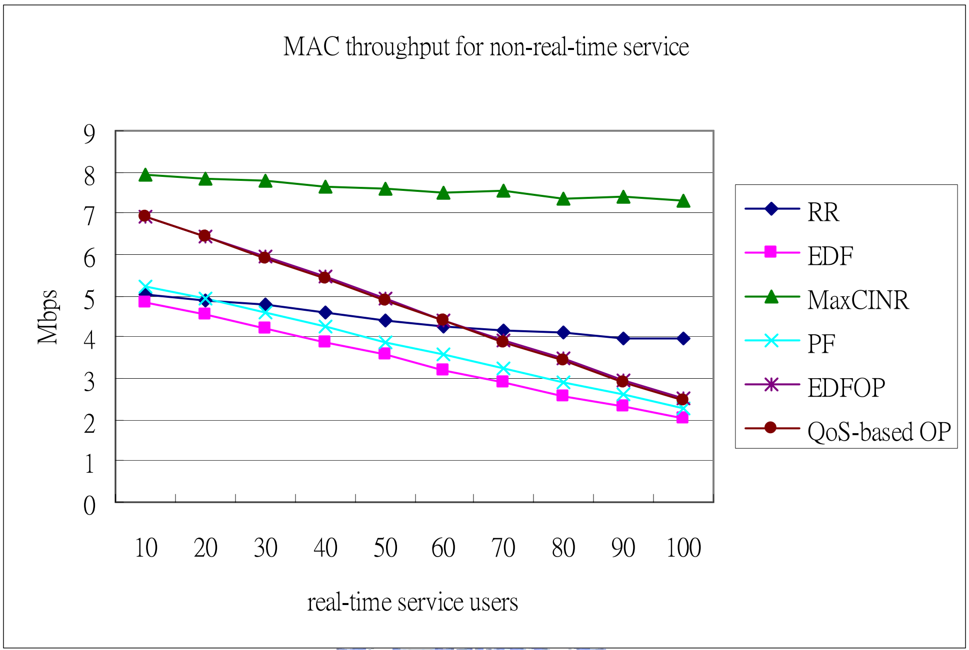

Figure 5-9 FTP throughput under fixed FTP users 20

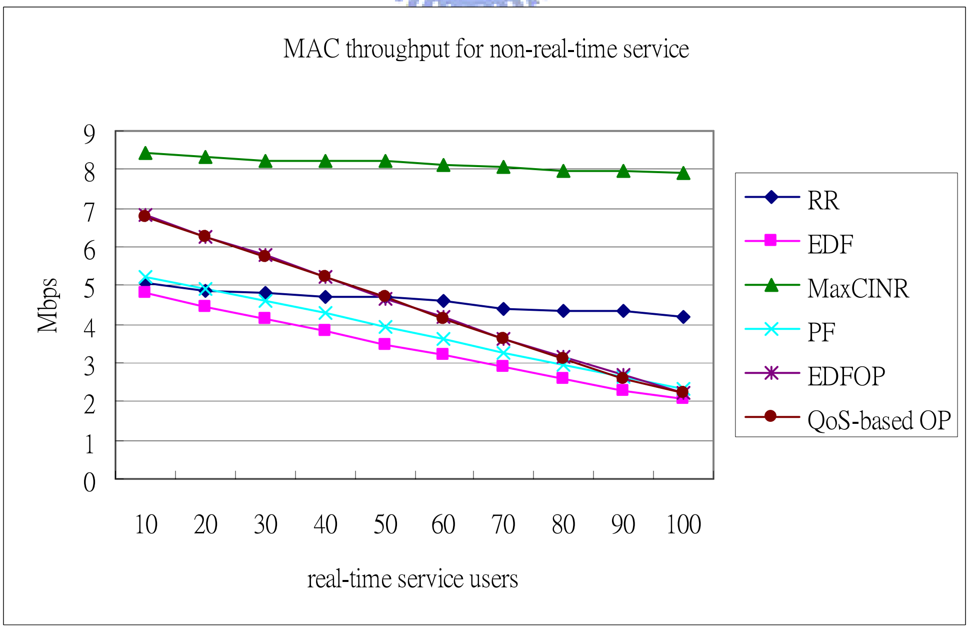

Figure 5-10 FTP throughput under fixed FTP users 30

From the figures above, we can find out that while in the mixed-traffic environment, PF algorithm can not provide good throughput performance. In most of time, it performs worst the RR. In the proposed QoS-based opportunistic algorithm, it can provide better throughput performance because it transmit packets only when they are really urgent or with good channel condition. It will release resources to the users with good channel quality to enhance throughput. However, the throughput of the proposed algorithm still degrades with VoIP users increase and the throughput might less than RR or PF algorithm. That is because it will support strict QoS requirement for users, both VoIP and FTP.

5.2 QoS Guarantee

After seeing the performance of throughput, we can figure out the proposed QoS-based opportunistic algorithm can achieve better performance in throughput than RR, EDF, and PF algorithm. That is one of the goals of proposed algorithm. The other goal of the proposed algorithm is try to maintain services' QoS requirement, no matter in single-traffic or mix-traffic. Hence, in this section, we will see if the QoS guarantee be made.

In this thesis, the indication of QoS will be defined as packet delay rate (PDR) for non-real-time service and packet loss rate (PLR) for real-time service. PDR means how many percentage of packets exceed the soft delay bound and PLR means how many percentage of packets exceed the delay bound.

In the section will show the plot of PDR for non-real-time only traffic and PDR, PLR for mix-traffic environment.

5.2.1 QoS Guarantee for FTP Users under Single-traffic Environment

In Figure 5-12 shows the simulation result of PDR for FTP users only. For MaxCINR algorithm, although it can achieve best throughput performance, it has large PDR because it doesn't take QoS requirement into consideration. Some of the users might never get the transmission opportunity at all. The QoS guarantee can not be achieved in MaxCINR algorithm although it has good performance in system throughput.

In this plot, we can figure out the difference between EDFOP and QoS-based opportunistic algorithm, which the latter one adds the delay bound separation mechanism to avoid congestion. While the users increasing, the greater possibility that packets will have similar soft delay bound and cause congestion. Hence EDFOP algorithm will cause congestion and congestion will cause many packets delay occurring much earlier than QoS-based opportunistic algorithm. While there are only 15 users in the system, the PDR starts to increase. In the proposed algorithm, due to the separation of the packets whose delay bound is closed. The congestion situation will not happen as early as EDFOP. The PDR of the proposed algorithm increases while the cell loading is too heavy, in figure 5-12, it rises after 50 users in the cell, about the same time that PF algorithm's PDR rising.

The EDF algorithm is the one which can meet the QoS requirement best. In figure 5-12 it has the best PDR. But it sacrifices the throughput performance mentioned before.

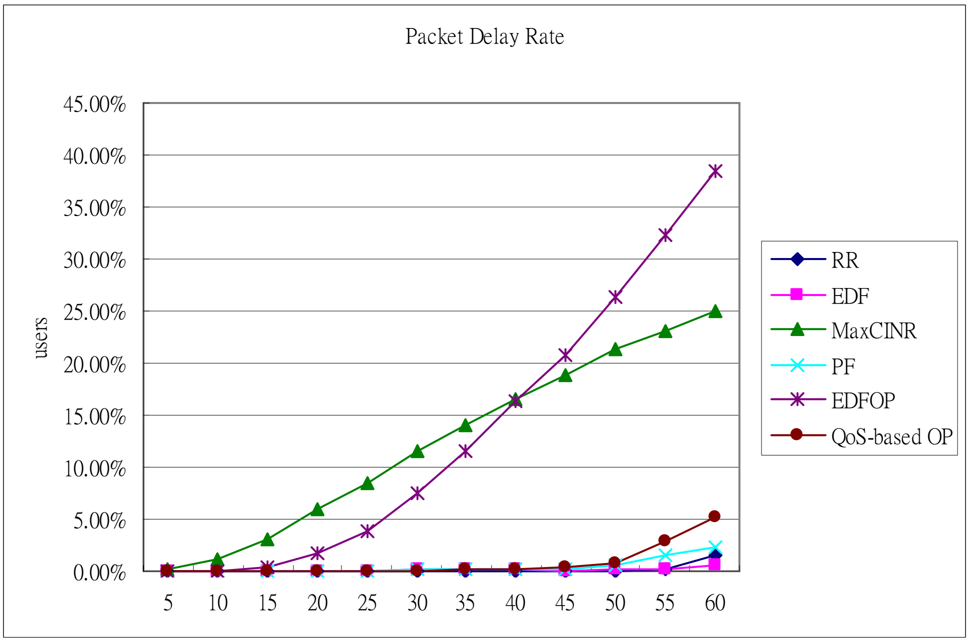

Figure 5-11 Packet delay rate for FTP service

5.2.2 QoS Guarantee under Mix-traffic Environment

For mixed-traffic service, hoping to achieve different services' QoS requirement and enhancing the system performance. We had seen the throughput performance before and figured out that the proposed QoS-based opportunistic algorithm can achieve much better throughput performance than EDF and PF algorithm. Moreover, it still has better throughput performance that RR while there are not so many VoIP users in the system. The reason we gave before is that the scheduler tries to maintain the QoS so it will give the resource to urgent packets first, most of time, the VoIP packets. Here we will see if our argument is correct. All of the simulations here fix the number of FTP services, 10, 20, and 30 while increasing the VoIP service step by step to see the influence.

5.2.2.1 PDR for FTP users under Mix-traffic Environment

In the thesis, use soft delay bound concept as the QoS definition of non-real-time service. From figure 5-16 to figure 5-18, show the PDR of FTP users, which denote the QoS of non-real-time service, while there are 10, 20, and 30 FTP users in the system.

The PDR of MaxCINR algorithm doesn't vary a lot from VoIP users' number. That is because it always picks users with good channel quality regardless of traffic information. It's obvious that this algorithm can not provide QoS guarantee. As for other algorithms, the PDR will be rising as the VoIP users increasing. However, the PDR of EDFOP algorithm is the one which rises first of all. That is because packets with similar delay bound congest in one frame. And the effect will accumulate to affect the following packets. Due to the characteristic of non-real-time traffic, system keeps transmitting packets which already exceed the delay bound and causes delay again and again.

While in the proposed QoS-based opportunistic algorithm, with delay bound separation, it can avoid the congestion that caused no sufficient resources to transmit urgent packets. Hence it performs much better than EDFOP algorithm in QoS guarantee for non-real-time service. However, the PDR will rise while there are more and more VoIP users in the system and the performance is worse than PF, RR, and EDF. Because the proposed one will support real-time services' QoS first, which will be dropped and must be supported to avoid data loss. Therefore, it will give the transmission opportunity to VoIP users more easily than PF and RR algorithm. As to EDF and RR algorithm, they consider the QoS factor or just transmit in fixed sequence to schedule packets. They perform better in QoS guarantee of non-real-time service than proposed one but it lack of the system performance consideration. The proposed one tries to transmit users in efficient way to enhance throughput, it must sacrifice the QoS guarantee in non-real-time services. That will be the trade-off in the scheduler design.

While the more FTP users in the system, the harder that to achieve the QoS guarantee. The PDR will rise earlier.

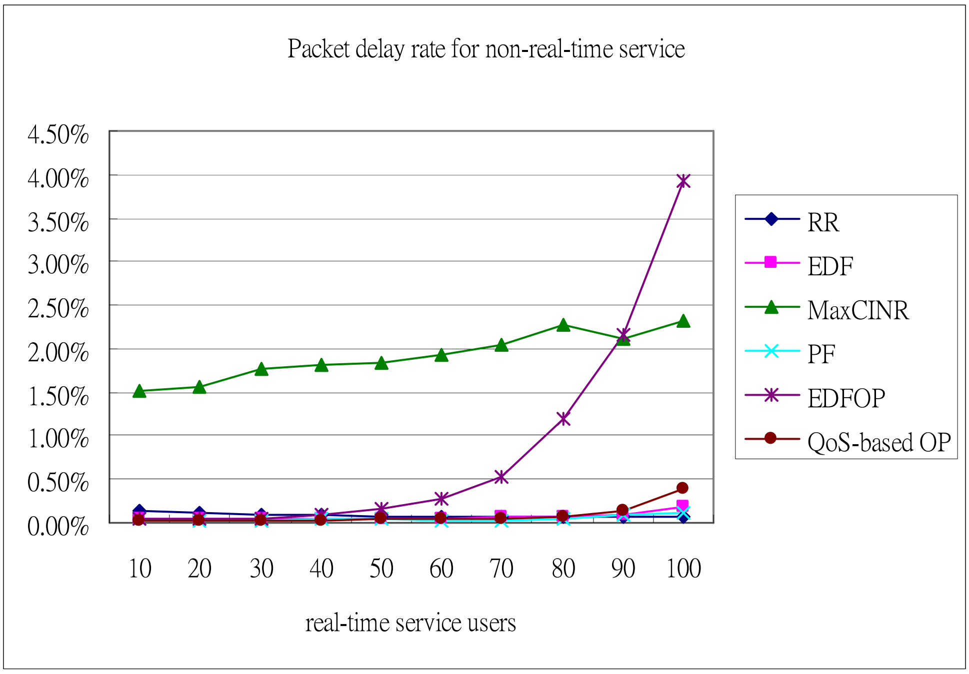

Figure 5-12 FTP packet delay rate under fixed FTP users 10

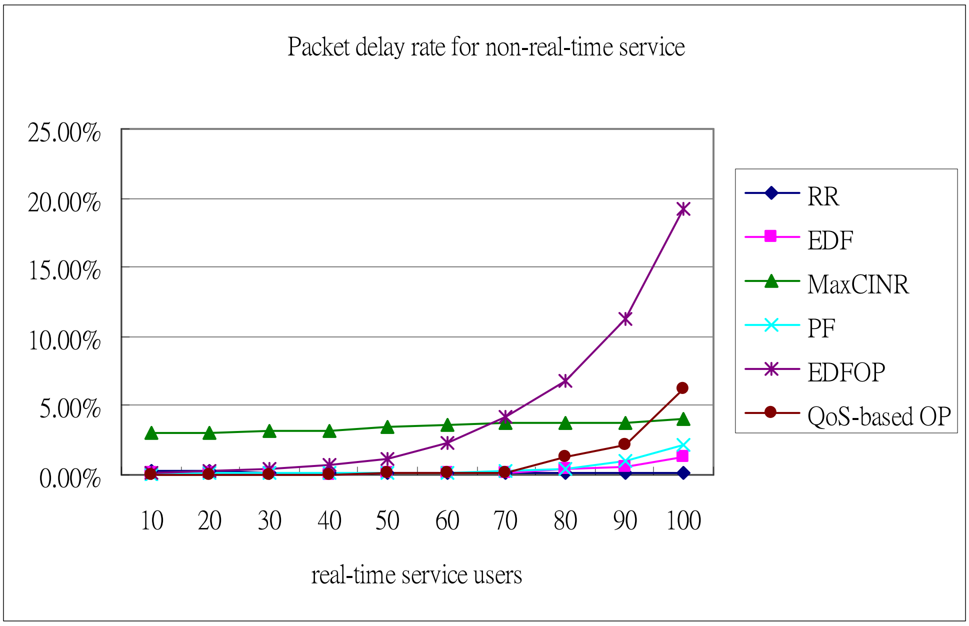

Figure 5-13 FTP packet delay rate under fixed FTP users 20

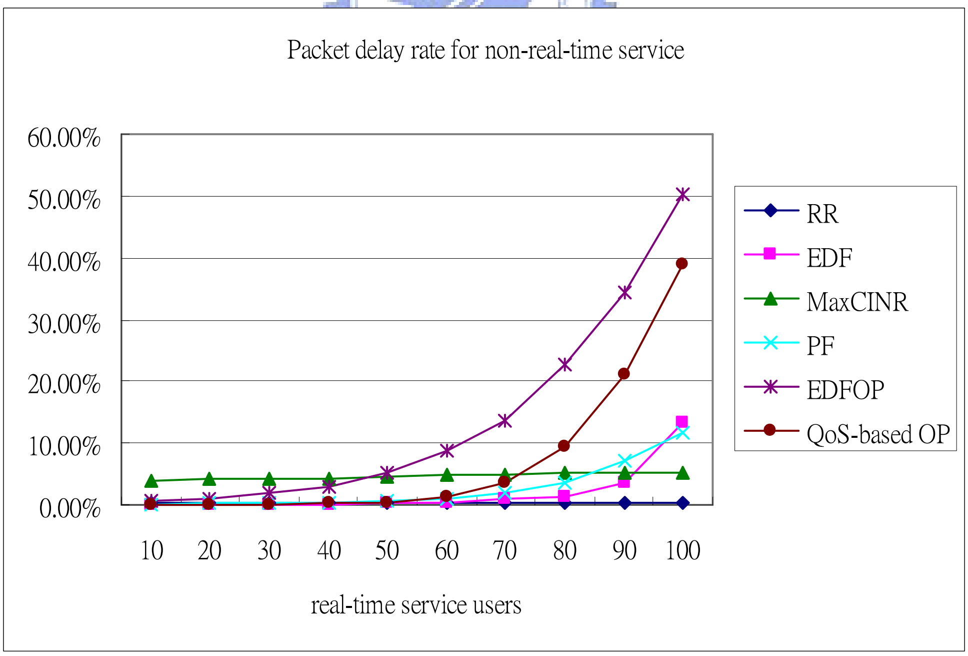

Figure 5-14 FTP packet delay rate under fixed FTP users 30

5.2.2.2 PLR for VoIP users under Mix-traffic Environment

In this thesis, use PLR as the indication of the QoS requirement for real-time service. From figure 5-19 to figure 5-24, show the PLR of VoIP users, which denote the QoS of real-time service, while there are 10, 20, and 30 FTP users in the system.

It is obviously that RR and MaxCINR algorithm can't meet the QoS requirement of VoIP service anymore in the mix-traffic condition. The PLR is much higher than other algorithm. They can not meet the QoS requirement for real-time service under mix-traffic environment. Therefore, let us focus on the other algorithms.

The disadvantage of PF is that it doesn't set the strict QoS guarantee mechanism as EDF concept. It only uses the fairness concept to try to maintain the QoS requirement but it doesn't work in the mix-traffic environment because there are different QoS requirement for different traffics. Hence it performs worse than other algorithm which use EDF concept, the stricter mechanism, to remain QoS guarantee, such as EDF, EDFOP, and the QoS-based opportunistic algorithm.

For EDF, EDFOP, and QoS-based opportunistic algorithm, which use the time concept as the QoS definition, they still have difference in PLR performance. We can figure out that no matter how many VoIP users in the system, EDFOP algorithm always has lower PLR than EDF algorithm. Compared with the mechanism of those two algorithms, EDFOP algorithm adds the mechanism to transmit packets which are more efficient. In this way, the reserved resource can be used to transmit more data. Therefore with the ability to save more resource to transmit more data, the less PLR the system has due to more resource to be used. The QoS-based opportunistic algorithm has similar performance with EDFOP while there are not so many VoIP users. For example, if the VoIP users exceed 80, the PLR of the proposed algorihm will rise quickly but EDFOP won't when there are 10 FTP users in the system.

What cause this situation that the two algorithms designed with similar concept but perform differently. That is the delay bound separation mechanism being implemented in the QoS-based opportunistic algorithm but EDFOP isn't. With the mechanism to separate packets according to their delay bound, the VoIP packets will have shorter and really closed delay bound. Therefore they will be assigned to the same group. However, FTP packets will have much longer delay bound and as the time goes by, they will be assigned into different groups. Back to QoS-based opportunistic algorithm, it will pick the one in the group which with shorter delay bound, then pick another packet in the next group. In this mechanism, most situation that only one VoIP packet is picked to transmit, then the opportunity of transmission will be given to FTP packets. This situation will be more obvious if there are more and more VoIP users in the system because although there are many VoIP packets wait to be transmitted, system picks FTP packets to transmit because all VoIP packets are in the same group. However, the EDFOP algorithm just picks packets with good efficiency. The VoIP packets will have larger probability to be picked than QoS-based opportunistic algorithm. That's why it will have lower PLR than the proposed algorithm while there are so many VoIP users in the system. However, without the mechanism to divide packets into different groups, EDFOP algorithm performs really badly in PDR of FTP packets.

While there are more FTP users, the higher possibility that the FTP packets be picket to transmit in the second stage of scheduling in the proposed QoS-based opportunistic algorithm. Therefore, the difference between the proposed one and EDFOP is getting more obvious in PLR.

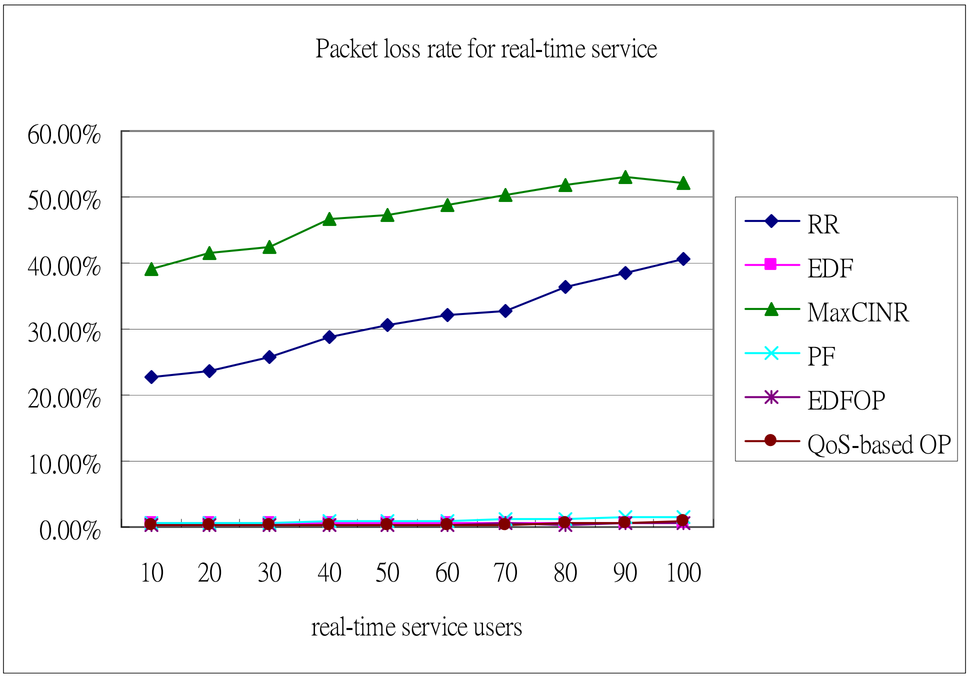

Figure 5-15 VoIP packet loss rate under fixed FTP users 10(1)

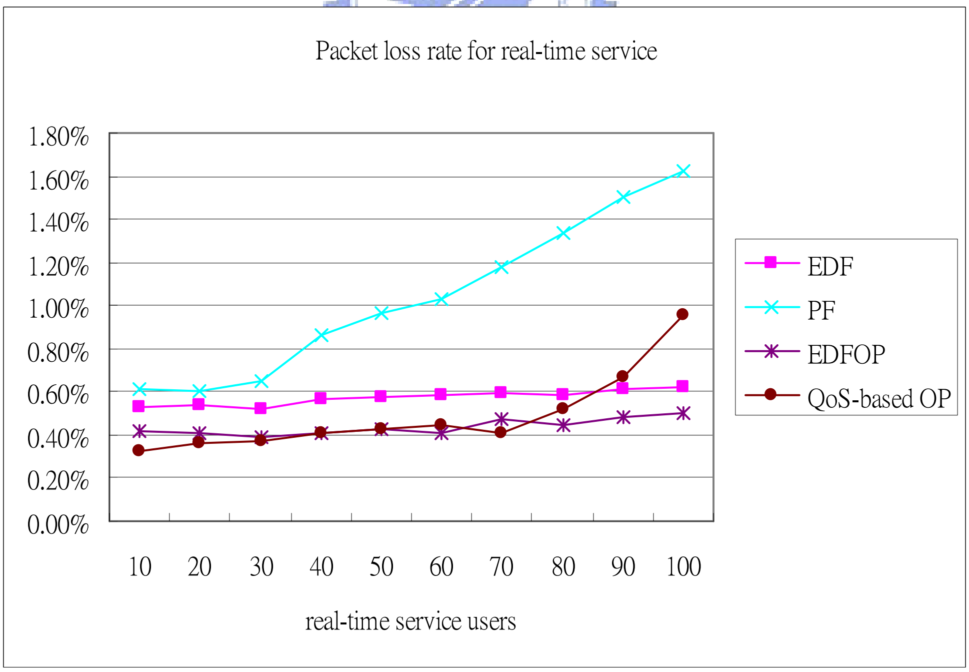

Figure 5-16 VoIP packet loss rate under fixed FTP users 10(2)

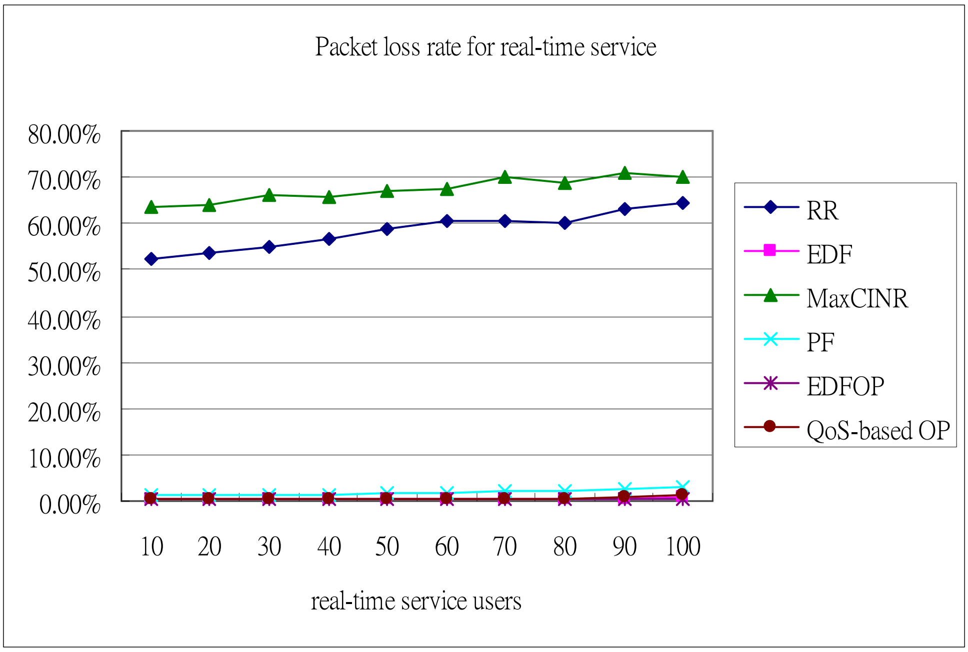

Figure 5-17 VoIP packet loss rate under fixed FTP users 20(1)

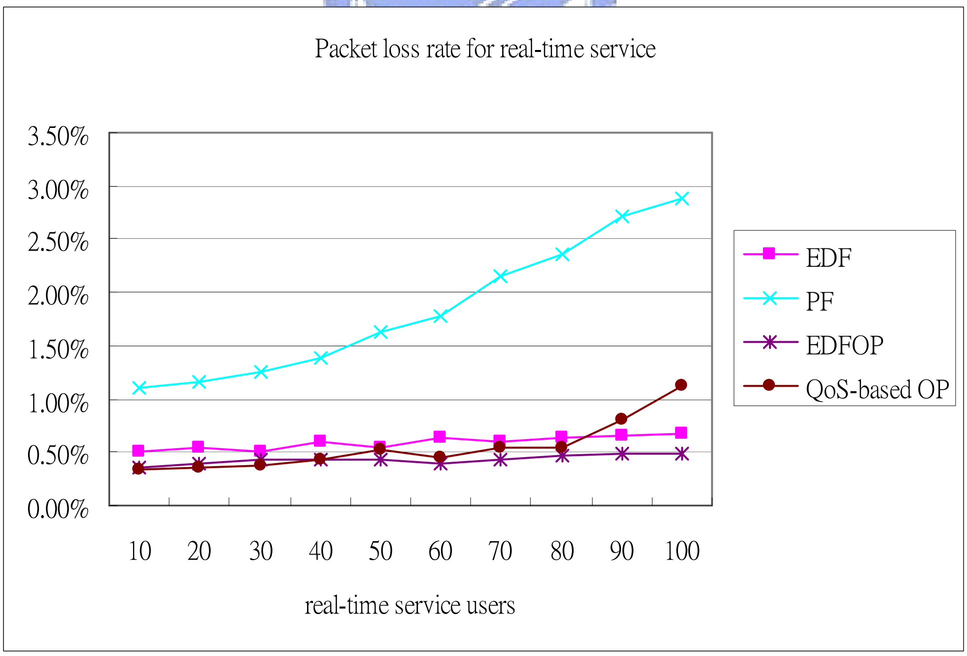

Figure 5-18 VoIP packet loss rate under fixed FTP users 20(2)

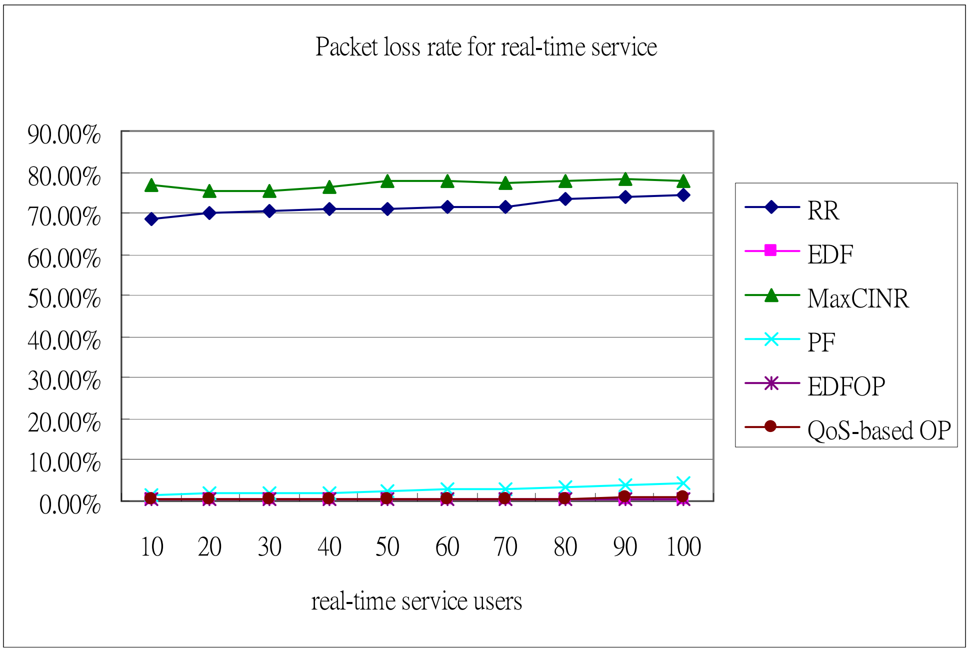

Figure 5-19 VoIP packet loss rate under fixed FTP users 30(1)

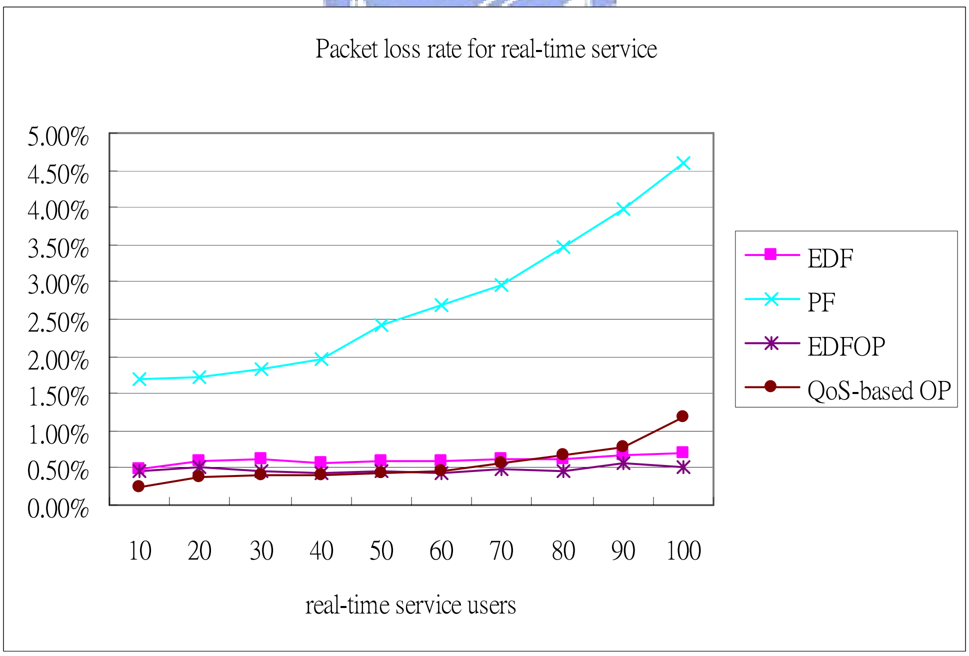

Figure 5-20 VoIP packet loss rate under fixed FTP users 30(2)

5.3 Conclusion of Simulation Results

From the simulation results, we can find out that the proposed QoS-based opportunistic algorithm have the ability to adapt to the environment. While the traffic load is not heavy, it can perform well in system throughput. While the traffic load is heavy, it will try to meet the QoS requirement for each packet, no matter real-time or non-real-time service. In throughput performance, although it can't perform as well as MaxCINR, it still performs much better than others if the environment is available. However, the MaxCINR can not maintain the QoS but the proposed one can. As to QoS concerns, it set the strict QoS requirement thus it performs much better than PF algorithm in the time-sensitive service. Moreover, it has better ability to handle the mix-traffic environment to maintain their QoS than EDFOP, which is created straightforwardly, because of the mechanism to divide packets into different groups. Although EDF algorithm emphasizes the QoS guarantee, however, it lacks the ability to enhance system throughput and due to efficient transmission, the proposed algorithm also performs better than EDF in maintaining QoS for real-time service in some situation.

Therefore, the proposed QoS-based opportunistic scheduling algorithm is really suitable to be used in the mix-traffic environment to ensure users' QoS and have the good ability to enhance system performance.

Chapter 6

Conclusion and Future Work

In the thesis propose a new QoS definition for non-real-time service. Using the soft delay bound concept, system can translate the QoS requirement from minimum reserved rate to delay bound which is widely used for real-time service. With the same QoS definition, scheduling mix-traffic will be much easier.

Then we propose a scheduling algorithm, QoS-based opportunistic scheduling algorithm using the concept of EDF and user diversity. The goal of this proposed algorithm is trying to maintain users' QoS requirement but maximize the system performance, throughput at the same time. The proposed algorithm is implemented in the WiMAX 802.16e system, with OFDMA physical layer.

The simulation result shows that the proposed algorithm has good performance in the mix-traffic environment to maintain the QoS requirement of different kinds of users. Besides, it has better performance in throughput for non-real-time service to utilize bandwidth efficiently. As for mix-traffic, it maintains the QoS for voice users with low PLR and FTP users with low PDR. While the traffic load is not so heavy, it can provide better throughput for non-real-time service. While the traffic load becoming heavier, the throughput performance of non-real-time service will start degrading because the resource will be used to maintain the users' QoS requirement.

Therefore, the proposed algorithm is really adapted for the traffic loading. While the traffic loading is not heavy, it will use user diversity concept to increase system performance but still maintain QoS requirement. If the traffic loading is heavy, it will try to maintain users' QoS requirement first. The system performance will no be the major consideration in this case.

While other algorithms have their own obvious disadvantages, the QoS-based opportunistic algorithm gives a method to handle mix-traffic environment with good performance but without the obvious disadvantages. It is a really competitive technique in the future wireless communication environment.

The proposed algorithm will be more efficient and robust than others. But it's still not good enough. In the future, the algorithm should also consider the fast fading and frequency select fading, which could be used as user diversity in the time and frequency index. The 802.16e system has the BandAMC structure to take advantage of frequency select fading. If the system takes fast fading and frequency select fading into consideration, it can upgrade the system performance more. Besides this, the call admission and power control of users are both important issues. The call admission control will block users when the system is overloaded. In this way, it can maintain the QoS of users further. The power control will enhance transmission power to the specific users, which is urgent to transmit but bothers by the poor channel condition. Both mechanism can enhance the performance in maintaining the QoS and will be important issues in the future work.

Reference

[1] IEEE, "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems", IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001), 2004 Page(s):0_1 - 857

[2] IEEE, "IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1",IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor 1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004), 2006 Page(s):0_1 - 822

[3] David Paranchych, "WiMAX / 802.16 Overview", NORTEL, Wireless Technology & Standards, Feb. 10, 2006.

[4] Vijay Dube, "CAN OFDM ENHANCEMENT DRIVE WIMAX MOBILITY FORWARD?", Wavesat, Oct 1, 2005.

[5] GuoSong Chu; Deng Wang; Shunliang Mei, David, "A QoS Architecture for the MAC Protocol of IEEE 802.16 BWA System", IEEE International Conference on Communications Circuits & System and West Sino Expositions, vol.1, Page(s): 435--439,

[6] Dong-Hoon Cho, Jung-Hoon Song, Min-Su Kim, Ki-Jun Han "Performance Analysis of the IEEE 802.16Wireless Metropolitan Area Network", Proc. First Int'l Conf. Distributed Frameworks for Multimedia Applications (DFMA '05), Page(s): 130-137, Feb. 2005.

[7] Jianfeng Chen, Wenhua Jiao, Hongxi Wang "A service flow management strategy for IEEE 802.16 Broadband wireless access systems in TDD mode", IEEE International Conference on Communications (ICC2005)

[8] Wengerter, C.; Ohlhorst, J.; von Elbwart, A.G.E., "Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA", Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st Volume 3, 30 May-1 June 2005 Page(s):1903 - 1907 Vol. 3

[9] Seokhyun Yoon; Youngkwon Cho; Chan-Byoung Chae; Hyunwoo Lee, "System level performance of OFDMA forward link with proportional fair scheduling", Personal, Indoor and Mobile Radio Communications, 2004. PIMRC 2004. 15th IEEE International Symposium on Volume 2, 5-8 Sept. 2004 Page(s):1384 - 1388 Vol.2

[10] Seungwan Ryu; Byunghan Ryu; Hyunhwa Seo; Mooyong Shin, "Urgency and Efficiency based Packet Scheduling Algorithm for OFDMA wireless system", Communications, 2005. ICC 2005. 2005 IEEE International Conference on Volume 4, 16-20 May 2005 Page(s):2779 - 2785 Vol. 4

[11] Chiussi, F.M.; Sivaraman, V., "Achieving high utilization in guaranteed services networks using early-deadline-first scheduling", Quality of Service, 1998. (IWQoS 98) 1998 Sixth International Workshop on 18-20 May 1998 Page(s):209 - 217

[12] Amoakoh Gyasi-Agyei, "Multiuser Diversity Based Opportunistic Scheduling for Wireless Data Networks", IEEE Communications Letters, Vol. 9, No. 7, July 2005 Page(s):670 - 672

[13] Amoakoh Gyasi-Agyei, "Comparison of Opportunistic Scheduling Policies in Time-Slotted AMC Wireless Networks". IEEE International Symposium on Wireless Pervasive Computing (ISWPC), Phuket, Thailand, 16-18 Jan 2006 Page(s):6 pp.

[14] Meng-Shiang Lin, "Radio Resource Management in Mobile WiMAX System"

[15] Tony Klinkert, "802.16e System Requirements Feedback from an operator

perspective", IEEE 802.16 Presentation SubmissionTemplate (Rev. 8.3), IEEE S802.16e-03/23, 2003-03-10.

[16] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Spatial channel model for Multiple Input Multiple Output (MIMO) simulations (Release 6)", TR 25.996 V6.1.0, 2003-09.

[17] “Mobile WiMAX – Part I: A Technical Overview and Performance Evaluation”, June 2006

[18] “Mobile WiMAX – Part II: A Comparative Analysis”, May 2006

[19] Harri Holma; Antti Toskala,"WCDMA for UMTS – Radio Access For Third Generation Mobile Communication", ISBN 0-471-72051-8 , John Wiley & Sons , 2000, pp.157~159.

[20] Winner, "Final Report on Link Level and System Level Channel Models", IST-2003-507581 WINNER D5.4 v. 1.4, Nov. 18th, 2005.

[21] Gudmundson, M., "Correlation model for shadow fading in mobile radio systems", Electronic Letters, pp.2145-2146, Vol. 27, Issue 23, November 1991.

[22] “cdma2000 Evaluation Methodology Revision 0”, 3GPP2 C.R1002-0 Version 1.0, December 10, 2004

[23]”3G & The Future of Wireless”, PDAs & Cell Phones, Vol.8 Issue 5,Page(s) 48-51, May 2002

[24] Winner, “Intermediate requirements per scenario”, IST-2003-507581 WINNER D1.2, 02/03/2005.