An Efficient Scheduling For Diverse QoS Requirements in WiMAX

by

Xiaojing Meng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2007

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

WiMAX is one of the most important broadband wireless technologies and is anticipated to be a viable alternative to traditional wired broadband techniques due to its cost efficiency. Being an emerging technology, WiMAX supports multimedia applications such as voice over IP (VoIP), voice conference and online gaming. It is necessary to provide Quality of Service (QoS) guaranteed with different characteristics, quite challenging, however, for Broadband Wireless Access (BWA) networks. Therefore, an effective scheduling is critical for the WiMAX system. Many traffic scheduling algorithms are available for wireless networks, e.g. Round Robin, Proportional Fairness (PF) scheme and Integrated Cross-layer scheme (ICL). Among these conventional schemes, some cannot differentiate services, while some can fulfill the service differentiation with a high-complexity implementation.

This thesis proposes a novel scheduling algorithm for Orthogonal Frequency Division Multiplex/Time Division Multiple Access (OFDM/TDMA)-based systems, which extends the PF scheme to multiple service types with diverse QoS requirements. The design objective is to provide differentiated services according to their QoS requirements, while the objective can be achieved by adjusting only one unique parameter, the time window for evaluating the average throughput. By extensive simulation, it is shown that the proposed scheduling algorithm exploits the advantage of the PF scheme, enhancing the throughput, and distinguishes the services in terms of the average delay. Afterward, we prove the superiority of the new scheme over the conventional ones by showing simulation results.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors, Professor Pin-Han Ho and Kainam Thomas Wong, for their guidance and support during my gradate studies. Without their help, the achievements in my research would never have been possible.

I would also like to thank Professor Murat Uysal and Professor Liangliang Xie for being the readers of this thesis and for their insightful comments and suggestion.

Many thanks to my friends and all members in my research group for their friendship, support and helpful discussion. Thanks to the administrative support staff, namely Lisa Hendel, Wendy Boles and Karen Schooley.

I am forever indebted to my parents for their constant support and encouragement throughout the past years. This thesis is devoted to them.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation

WiMAX (Worldwide interoperability for Microwave Access) is one of the most emerging technologies for Broadband Wireless Access (BWA) in metropolitan areas by providing an exciting addition to the current broadband techniques for the last-mile access. It is demonstrated that WiMAX is a viable alternative to the cable modem and DSL technologies due to its high resource utilization, easy implementation and low cost. Furthermore, WiMAX not only enhances the existing features of the competitive cabled access networks, but provides high data rate applications with a variety of Quality of Service (QoS) requirements.

We are reaching the goal of realizing a unique wireless network to cover a big area. In a large scale wireless network, the radio resource must be shared among multiple users. The bandwidth allocation algorithms have been designed for the efficient utilization of the scarce radio resource. In addition, to support multimedia traffics, the Medium Access Control (MAC) protocols will co-ordinate the transmission of traffic flows. The channel characteristics of users and traffic flow requirements are largely diverse, motivating us to design an efficient MAC layer protocols that can improve the system performance due to the channel and traffic dynamics.

1.2 Research Challenges

Scheduling algorithms serve as an important component in any communication network to satisfy the QoS requirements. The design is especially challenged by the limited capacity and dynamic channel status that are inherent in wireless communication systems. To design an MAC layer protocol which can optimize the system performance, the following features and criteria should be concerned. [1-3]

- Bandwidth utilization

  Efficient bandwidth utilization is the most important in the algorithm design. The algorithm must utilize the channel efficiently. This implies that the scheduler should not assign a transmission slot to a connection with a currently bad link.

- QoS requirements

  The proposed algorithm should support different applications to exploit better QoS. To support delay-sensitive applications, the algorithm provides the delay bound provisioning. The long-term throughput should be guaranteed for all connections when the sufficient bandwidth is provided.

- Fairness

  The algorithm should assign available resource fairly across connections. The fairness should be provided for both short term and long term.

- Implementation complexity

  In a high-speed network, the scheduling decision making process must be completed very rapidly, and the reconfiguration process in response to any network state variation. Therefore, the amount of time available to the scheduler is limited. A low-complexity algorithm is necessary.

- Scalability

  The algorithm should operate efficiently as the number of connections or users sharing the channel increases.

Our protocol design is desirable to fulfill all of the above features.

1.3 Main Contributions

The major objective of this thesis is to develop an effective yet simple QoS scheduling scheme for multi-service networks that can be deployed and implemented with the less overhead. We explore the limitation on the proportional fairness (PF) scheduling scheme, and propose to extend the PF scheme for delay differentiation in IEEE 802.16 networks (or referred to as WiMAX). For meeting different delay constraints on each service type, T_i can be manipulated to serve this purpose. The analysis results are given by the simulations, which have shown that the proposed algorithm can differentiate services in terms of the delay constraints over a large scale network.

1.4 Thesis Outline

Chapter 2 introduces the background of IEEE802.16 Physical Layer (PHY) and Medium Access Control (MAC) layer protocols. An overview of previous related work is presented in Chapter 3. Chapter 4 describes the developed protocol. The analytical and simulation results by using MATLAB are validated in Chapter 5. Finally, the conclusions are drawn in Chapter 6.

Chapter 2
Wireless Metropolitan Area Networks Overview

In the early 2000's, BWA in Metropolitan Areas has been recognized as one of the most promising technologies that will be widely deployed in the world. In order to rapidly converge on a worldwide standard, several standards have been published. A number of options are provided in the IEEE 802.16 family. [15-16]

- **IEEE 802.16a:** The standard specifies the operation from 2GHz to 11GHz, both licensed and license exempts. Because the signals at lower frequency can penetrate barriers and thus a line-of-sight connection between the transceiver and receiver is not required, most commercial interests have focused mainly on the lower frequency ranges. Under this premise, IEEE 802.16a standard was thus completed in January 2001. It enables the WiMAX implementations with better flexibility while maintaining the data rate and transmission range. IEEE 802.16a also supports mesh deployment, which can extend the network coverage and increase the overall throughput.
- **IEEE 802.16b:** This extension increases the spectrum to the 5 and 6 GHz frequency bands, which provides QoS guarantee to ensure priority transmission for real-time applications and to differentiate service classes for different traffic types.

- **IEEE 802.16c:** As the Work Group's initial interest, IEEE 802.16c defines a 10 to 66 GHz system profile that standardizes more details of the technology. These high frequency bands have more available bandwidth, but the signals cannot diffract the obstacles and require line of sight deployment.
- **IEEE 802.16d:** Approved in June 2004, IEEE 802.16d upgrades the 802.16a standard. This extension aims to improve performance for 802.16 especially in the uplink traffic.
- **IEEE 802.16e:** This technology standardizes networking between fixed base stations (BSs) and mobile base stations (MSs), rather than just between base stations and fixed recipients. IEEE 802.16e enables the high-speed signal handoffs necessary for communications with users moving in vehicles. It promises to support mobility up to speeds of 70-80mi/h. The subscriber stations (SSs) could be personal communication devices such as mobile phones and laptops.

We only concentrate on some basic characteristics of IEEE 802.16d PHY and MAC protocols that are necessary for downlink scheduling algorithm design in the fixed network architecture. In the following sections, an overview on IEEE 802.16 PHY subsystems is provided.

2.1 PHY Technology in IEEE 802.16

IEEE 802.16 is a universal standard comprehending various types of network architecture. IEEE 802.16 defines two different network topologies each with a specific MAC protocol: the point to multipoint (PMP) mode and mesh mode. The mesh mode is optional in IEEE 802.16e, where data can be routed directly between two SSs. In the PMP mode, a central BS is capable of handling multiple independent SSs simultaneously. It does not need to coordinate with other stations. Nowadays, most WiMAX systems are equipped with the PMP mode where traffic only occurs between a BS and its SSs [17].

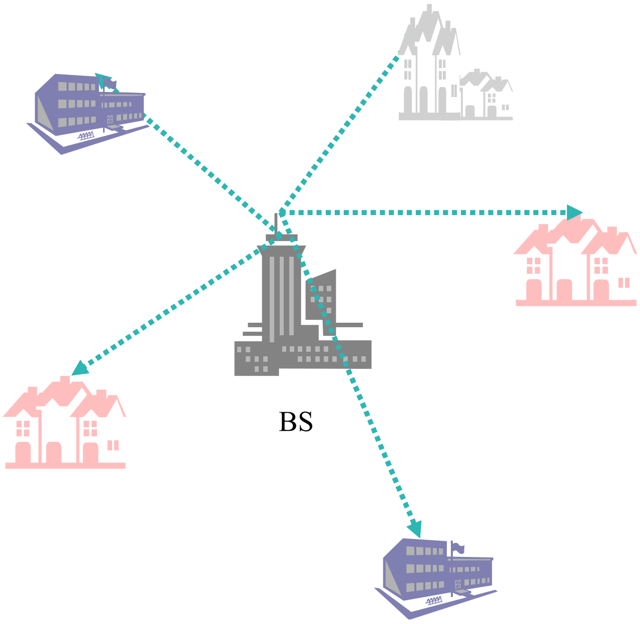


Fig. 2.1 Fixed PMP architecture network

2.1.1 Broadband Wireless Access Background

Several frequency bands for the initial 802.16 products have been identified. In IEEE 802.16a-2001, the frequency is addressed from 10 to 66 GHZ, which is available all over the world. Due to higher frequency, Line-of-Sight (LOS) propagation is a necessity. For a residential application, roof tops may be too low for a clear sight line to a BS. We must consider the multipath propagation affection. Recently, more interest is in the 2-11GHz projector. Design of the 2-11 GHz PHY is driven by the need for non-LOS (NLOS) operations. The standard defines three different air interfaces that can be used to provide a reliable end-to-end link:

- SCa: A single-carrier modulated air interface.
- OFDM: A 256-carrier orthogonal-frequency division multiplexing (OFDM). Multiple access of different SSs is time-division multiple access (TDMA)-based.
- OFDMA: A 2048-carrier OFDM scheme. But a subset of the carriers can be assigned to an individual user. It is referred to be OFD multiple accesses.

Table 1 summarizes the nomenclature for the various air interface specifications in standard.

Table 1 Air interface nomenclature

Designation	Applicability	Duplexing alternative
WirelessMAN-SC	10-66GHZ	TDD&FDD
WirelessMAN-SCa	Below 11GHZ,licensed band	TDD&FDD
WirelessMAN-OFDM	Below 11GHZ,licensed band	TDD&FDD
WirelessMAN-OFDMA	Below 11GHZ,licensed band	TDD&FDD

Among these three air interfaces, the two OFDM-based systems are more suitable for NLOS due to the simplicity of the equalization process for multicarrier signals [18]. In a multiple access communication system, transmission resources are shared among multiple users such that a resource management scheme is required. TDMA and Frequency Division Multiple Access (FDMA) are two well-known techniques for resource management based on the principle of time sharing and frequency sharing, respectively. When combined with OFDM (Orthogonal Frequency Division Multiplexing), they are called OFDM-TDMA and OFDMA (OFDM Access), respectively. The time and subcarrier assignment is illustrated in Fig. 2.2. All profiles currently defined by the WiMAX Forum specify the 256-carrier OFDM PHY. For this reason, the study focuses primarily on the 256-carrier OFDM/TDMA air interface, where each SS can take an OFDM symbol with all the subcarriers within a time slot exclusively. The access by multiple users is realized along the time domain. The advantage of OFDM/TDMA is that the number of physical time slots and the number of codes assigned are adjustable. It leads to different data rate armed with adaptive modulation and coding (AMC) available in PHY.

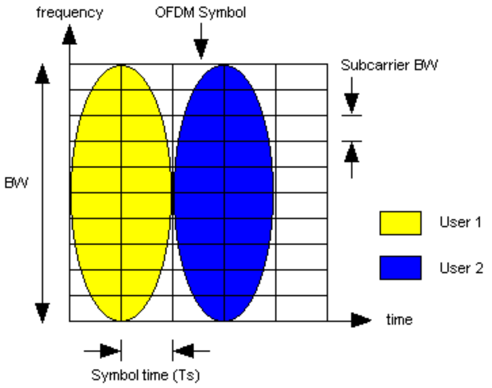


Fig. 2.2 (a) OFDM/TDMA

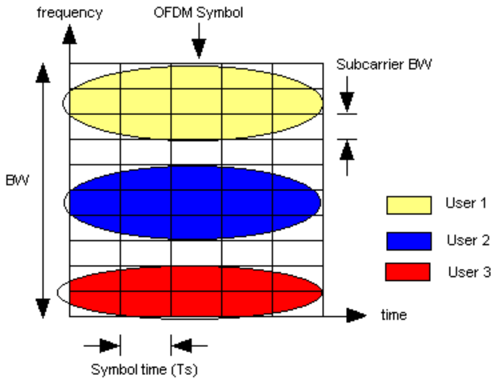


Fig. 2.2 (b) OFDMA

2.1.2 Adaptive Modulation and Coding (AMC)

The main objective of adaptive modulation and coding is to compensate for radio channel instability. It has been shown in [19-21] that adaptive modulation can effectively improve the bit error rate (BER) performance on radio channels which had suffered from shadowing and fading. The modulation schemes defined in the IEEE 802.16 standard in the downlink and uplink are binary phase shift keying (BPSK), quaternary PSK (QPSK), 16-Quarter Amplitude Modulation (QAM), and 64-QAM. The system could yield the best performance if the switching thresholds are selected carefully. Given the BER less than10^{-6}, the algorithm of choosing the optimal transmission mode is in accordance with Table 2 [22]. For WiMAX, the selection of a modulation scheme basically takes advantage of the radio channel measurements extracted by the SS which exercises the Channel State Information (CSI) and the retransmission procedure. The measurement on the Signal to Noise Ration (SNR) is performed on each frame preamble.

Table 2 Transmission Modes in IEEE 802.16

Modulation	Coding rate	W bits/symbol	Receiver SNR (dB)
BPSK	1/2	0.5	6.4
QPSK	1/2	1.0	9.4
	3/4	1.5	11.2
16 QAM	1/2	2.0	16.4
	3/4	3.0	18.2
64 QAM	2/3	4.0	22.7
	3/4	4.5	24.4

Based on a perfect channel measurement which can match the transmitter parameters to time varying channel conditions, a proper modulation and coding method can be chosen for the upcoming transmission so that the user bit rate can be maximized. AMC has been used to provide high-speed data transmission by many standard wireless networks, such as IEEE 802.11/16 and 3GPP/3GPP2 [23].

2.2 MAC Layer in IEEE 802.16

The MAC layer of IEEE 802.16 is designed to serve sparsely distributed stations with high data rates, where the SSs are not required to listen to the other stations like the MAC in IEEE 802.11. The BS schedules the transmissions of the corresponding SSs in advance. The MAC of WiMAX is reservation-based and contention-free. The SSs need to contend only when they access the channel for the first time at the connection admission control stage. The reservation-based resource allocation allows the WiMAX BS to serve a large number of SSs as well as the guarantee of QoS in the connection level for both uplink and downlink traffic. Compared with 802.16, Wireless Local Area Networks (WLAN) based on IEEE 802.11 terminals usually have intermittent traffic that contends every time before transmitting, where the efficiency is significantly impaired when more stations enter the network. In such a contention based resource reservation scheme, QoS could hardly be considered in the early standard until the advent of 802.11e. However, most WLAN networks deployed nowadays do not employ any QoS mechanism [24].

The main purpose of the MAC protocol is the sharing of radio channel resources among multiple accesses of different users. In IEEE 802.16, the MAC layer is divided into three sublayers: the service-specific convergence, common part sublayer, and security sublayer. The primary task of service-specific convergence sublayer is to classify external service data units (SDU) and associate each of them with a proper MAC service flow identifier and connection identifier. The MAC layer protocol is flexible and efficient over different traffic types. The common part sublayer is independent of the transport mechanism, which is the kernel bearing all the MAC characteristics. It is responsible for fragmentation and segmentation of each MAC SDU into MAC protocol data units (PDUs), system access, bandwidth allocation, connection maintenance, QoS control, and scheduling transmission, etc. The MAC also contains a separate security sublayer handling authentication, secure key exchange, and encryption.

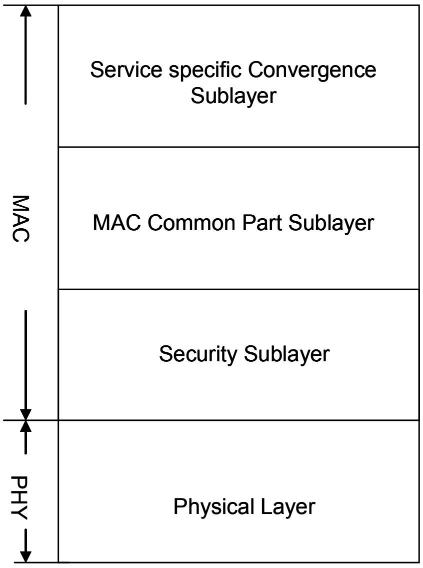


Fig. 2.3 IEEE 802.16 protocol layering

2.2.1 MAC Support of PHY

The basic distinction of MAC protocol is the duplexing techniques of uplink and downlink. The choice of duplexing techniques may affect PHY parameters as well as impact the features that can be supported. There are two approaches to implement it.

1) Frequency Division Duplex (FDD): In an FDD system, the uplink and downlink channels use separate subcarriers, which allows the terminals to transmit and receive simultaneously. The adoption of fixed duration frames in both uplink and downlink simplifies the design of bandwidth allocation algorithms.
2) Time Division Duplex (TDD): In this paper, TDD framing is illustrated in Fig. 2.4. The uplink and downlink transmissions share the same frequency while being allocated in each TDD frame according to an adaptive threshold. One TDD frame (Fig. 2.4) contains one downlink and one uplink subframe in a TDD frame separated by the threshold, which is divided into an integer number of physical slots (PSs). The downlink subframe comes first because it contains the bandwidth requests and transmission information directly sent from SSs to the BS, which forms a map for scheduling the uplink resources among all the SSs.

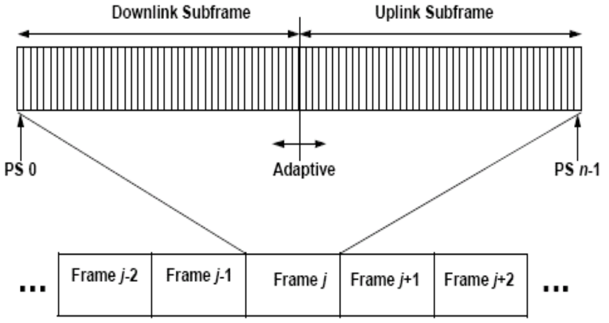


Fig. 2.4 The TDD frame structure in IEEE 802.16.

The structure of the downlink subframe using TDD mode is illustrated in Fig. 2.5 [22], [24]. The downlink subframe begins with a frame start preamble used by the

PHY layer for synchronization and equalization. The preamble is followed by the frame control section, containing DL-MAP and UL-MAP stating the resource allocation of the downlink and uplink. The DL-MAP specifies when the PHY layer transition occurs within the downlink subframe. The following portion carries the data, which are transmitted to each SS using a negotiated burst profile. Due to the dynamics of the bandwidth demand for the varieties of services, the burst profiles vary dynamically from frame to frame. Thus the system can support different levels of the data transmission. In the case of TDD, a time gap separates the downlink subframe and the uplink subframe.

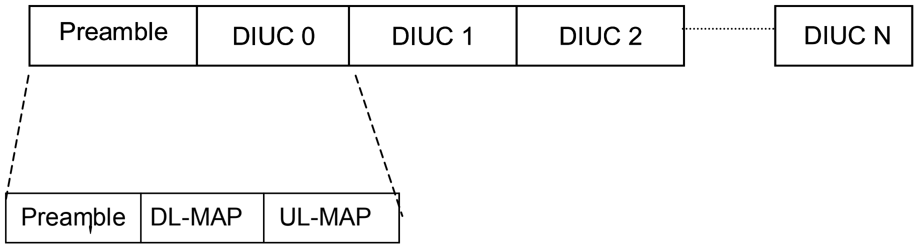


Fig. 2.5 The TDD downlink subframe structure in IEEE 802.16.

2.2.2 Scheduling Service

Scheduling services have represented the data handling mechanisms supported by the MAC scheduler for the data transport on a connection. To provide the service parameters respectively, the traffic management is necessary. The IEEE 802.16 standard divides all services in four different classes. Each group corresponds to a single service class, which is associated with a set of QoS parameters for quantifying the aspects of its behavior. Firstly, we outline these four service flows [22]:

1) Unsolicited grant service (UGS): It supports the constant bit rate (CBR) or fixed throughput connections at periodic intervals, such as T1/E1 and voice over IP

(VoIP) which needs to grant the constant bandwidth without any request. This service can guarantee the data throughput and the latency.

2) Real-time polling service (rtPS): It is a real time data stream comprising variable bit-rate (VBR) data packets which are issued at periodic intervals, such as the moving pictures experts group (MPEG) video. This application guarantees the minimum reserved rate and the latency, which are same as those of UGS. But the rtPS has to request transmission resources by polling, which means that the contention and the piggyback are not allowed.
3) Non-real-time polling service (nrtPS): This service is a delay-tolerant data stream consisting of variable-sized data packets, such as the file transfer protocol (FTP). The minimum data rate is required and the bandwidth request by polling is needed.
4) Best effort (BE): It does not provide any QoS guarantee, like the email or the short length FTP. There is no minimum resources allocation granted, where the occurrence of dedicated opportunities is subject to the network load. The channel access mechanism of this service is based on the contention.

2.2.3 Multiuser Diversity

Multiuser diversity was introduced to deal with time-varying fading channels in multiuser systems [9], [25]. In WiMax, because of the multiple users with different and time-varying channel conditions, there is a high probability that one or several users have very good links with the BS while the others have bad ones. By adapting to the channel conditions, the system throughput can be increased through the opportunistic scheduling. This is also referred to as the Adaptive Modulation and Coding (AMC), which can serve a large number of randomly distributed users with channel fading independently. In this case, the long term system throughput can be maximized by serving the users with the best channel when transmitting. The scheduler structure of downlink is depicted in Fig. 2.6. Although the proposed algorithm is presented in this paper is for downlink scheduling, this idea can be extended to uplink data transmissions.

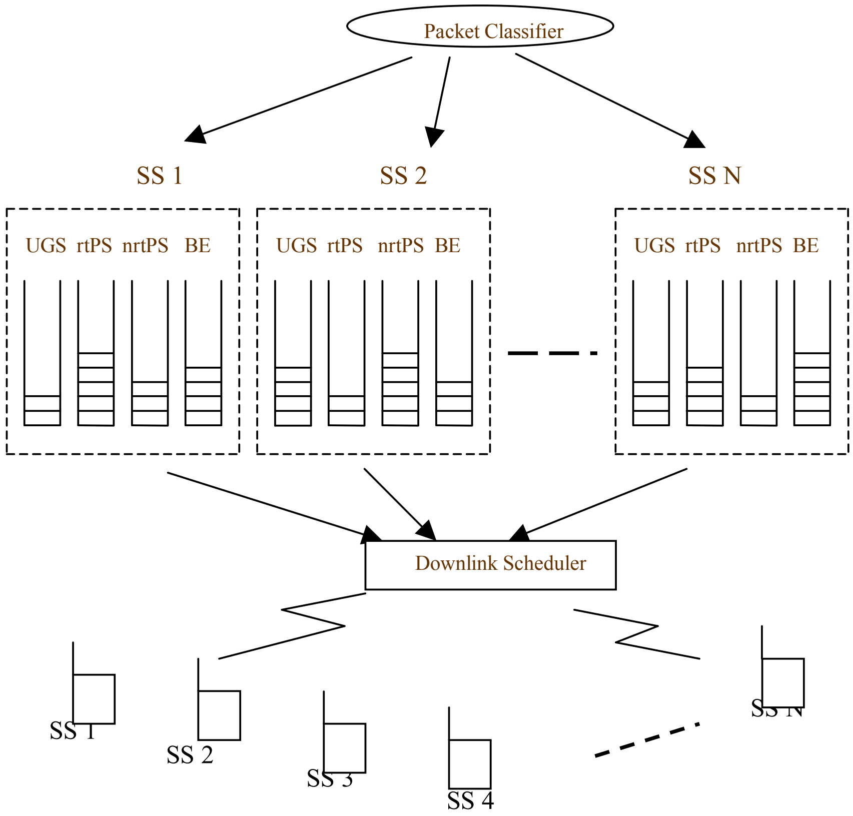


Fig. 2.6 Packet Scheduler structure

Chapter 3
An Overview of Scheduling Algorithms

Although there are a number of packet scheduling algorithms have been proposed for WiMAX network [4-6], the design of those algorithms are challenged by supporting different levels of services, fairness, and implementation complexity and so on. In this chapter, we investigate three widely adopted scheduling techniques, Round Robin (RR), Proportional Fairness (PF), and Integrated Cross-layer scheduling (ICL) in wireless networks.

3.1 Round Robin (RR)

RR is one of the simplest scheduling algorithms designed especially for a time sharing system, where the scheduler assigns time slots to each queue in equal portions without priority. Once a queue is served, it is not visited again until all the other SSs in the system have been served. RR can provide a fair resource access to each SS, and every queue is allocated with the same portion of system resources regardless of the channel condition. However, the RR scheduler has the same bandwidth efficiency as a random scheduler. Also, it cannot guarantee different QoS requirements for each queue.

3.2 Proportional Fairness (PF)

PF was proposed by Qualcomm Company, which was realized in the IS-856 standard for the downlink traffic scheduling (also known as High Data Rate (HDR)) [7-8]. The essential goals of this packet scheduling scheme are to enhance the system throughput as well as provide fairness among the queues under consideration. Proportional Fairness scheduling is based on one priority function:

$\mu_i(t) = \dfrac{r_i(t)}{R_i(t)}$, where $r_i(t)$ is the current data rate, $R_i(t)$ denotes an exponentially smoothing average of the service rate received by SS *i* up to slot *t*. Then the queue with the highest $\mu_i(t)$ is served at time slot *t*, where the average throughput of the queue is updated by $R_i(t+1) = (1 - \dfrac{1}{T_c})R_i(t) + (\dfrac{1}{T_c})r_i(t)$, where T_c is the time constant for the moving average. The average throughput of the queues that are not served at time slot *t* is updated by the relationship: $R_i(t+1) = (1 - \dfrac{1}{T_c})R_i(t)$. In general, T_c is assumed to be 1000 slots (1.66 seconds) in the CDMA-HDR system [7]. Taking a larger value of T_c makes the perceived throughput less sensitive to the short-term starvation on the queue, where the scheduler may wait for a longer period of time for a user turning back from a bad channel condition to a good one. When a huge number of users coexist in the system, we can obtain additional throughput gain by scheduling them to utilize the characteristics of fast fading channels, called multi-user diversity gain. This simple scheduler design enhances the overall throughput. In this way, the preference metric can be implemented for achieving PF among the SSs [9-10]. Although PF is simple and efficient, it cannot guarantee any QoS requirement such as delay and delay jitter due to its original design for saturated queues with non real-time data service.

3.3 Integrated Cross-layer Scheduling

Both RR and PF cannot manage the resource allocation and grants an appropriate QoS per connection. To resolve it, in the previous work, [11-13] proposed integrated QoS control architecture for IEEE 802.16 wireless systems. These scheduling schemes rely on different algorithms to handle different classes of services for matching their QoS requirements. To have a comprehensive introduction, a representative cross-layer scheduling algorithm with QoS support by [11] is briefed as follows.

The scheduler bases on a priority function for each queue, where the priority metric of each queue is updated according to its service status and channel condition in the PHY layer. Thus, the scheduler can provide the prescribed diverse QoS guarantees. First of all, the algorithm allocates a fixed number of time slots for the UGS queues, which is the requirement stipulated in the standard. The queues for real-time Polling Service (rtPS) are managed with an Earliest Deadline First (EDF) algorithm [14], which is sensitive to delay latency and reliable for real-time services. An opportunistic scheme which is similar to the PF algorithm is deployed for the queues supporting non-real-time Polling Service (nrtPS), while the queues for Best Effort (BE) traffic are managed based on a Best-Rate discipline. In order to differentiate the priority of the four types of services such that *rtPS* > *nrtPS* > *BE*, the class coefficients are assigned to the queues of each service type.

The algorithm is implemented according to the following formulas. The total time slots are allocated to Unsolicited Grant Service (UGS) data streams as N_{ugs} per frame. The residual time slots assigned to the rest QoS classes $N_r = N_d - N_{ugs}$, where N_d is total time slots in one frame.

The priority function for connection *i* at time slot *t* is defined as

$$\phi_i(t) = \begin{cases} \beta_{class} \dfrac{R_i(t)}{R_N} \dfrac{1}{F_i(t)} & \text{if} \quad F_i(t) \geq 1 \\ \beta_{class} & \text{if} \quad F_i(t) < 1 \end{cases}$$

Where $\beta_{class} \in [0,1]$ is the coefficient according to the service class, respectively. Based on the priority for different QoS classes, *rtPS*>*nrtPS*>*BE*, the coefficient can be set under the constraint $\beta_{rt} > \beta_{nrt} > \beta_{BE}$. $R_i(t)$ is the number of bits can be carried by one symbol at frame *t* via AMC, which is determined by the channel condition. R_N denotes the maximum number of bits in one symbol can be reached.

For each rtPS connection, *Fi(t)* is an indicator of the delay satisfaction. $F_i(t) = T_b - W_i(t)$, with R_N denoting the delay bound and $W_i(t) \in [0, T_i]$ denoting the longest packet waiting time. For simplifying the formula, we do not consider the guard time here. This priority function normalizes $\phi_i(t) \in [0, \beta_{rt}]$. When $F_i(t)$ <1, which means the packets in the queue *i* should be sent immediately to avoid the packet drop, the highest values β_{rt} has been set. For each non real time connection, $F_i(t)$ is the ration of the average transmission rate to the minimum reserved rate. $F_i(t) = \frac{\hat{\eta}_i(t)}{\eta_i}$, where $\hat{\eta}_i(t)$ is estimated by

$\hat{\eta}_i(t+1) = (1 - \frac{1}{Tc})\hat{\eta}_i(t) + (\frac{1}{T_c})r_i(t)$, $r_i(t)$ is the transmission rate at time *t*. At this time, $F_i(t)$ is an indicator of the data rate satisfaction. So if $F_i(t)$ <1, the packets of the *ith* stream should be sent to meet the rate requirement. The upper bound of priority value for nrtPS is β_{nrt}.

Because there is no QoS requirement for BE connections, the priority function for a BE connection is $\phi_i(t) = \beta_{BE} \frac{R_{i(t)}}{R_n}$. $\phi_i(t)$ only depends on the normalized channel quality regardless of the delay or rate performance.

This scheme provides a diverse QoS support for multiple connections. However, the author cannot offer how to set the upper bound of $\beta_{rt}, \beta_{nrt}, \beta_{BE}$, time slots number reserved for UGS *Nr* and delay bound T_i to get the optimal performance of the system. The equivalence of different priority functions for four types of services has not been proved. The scheduler is hard to be practically deployed due to its high implementation complexity.

Chapter 4
Scheduler Design

4.1 System Model

We propose a novel scheduler in this chapter, which is designed for a fixed PMP WIMAX system. Only one BS in the network serves all the SSs, therefore the inter-BS interference can be neglected. The downlink channel is shared by all SSs in a time division multiplexing manner, where a downlink scheduler is deployed at the BS to schedule the transmissions corresponding to the queues. The transmission occurs within a fixed-sized time frame. Only the selected queues can be severed within the frame. The air interface specification is OFDM which employs a fast Fourier transform (FFT) of size 256. All carriers are assigned to one queue for the data transmission in a time slot

Over a wireless fading channel, AMC described in 2.1.2 is employed at the PHY layer. The BS exactly knows the channel state information of all the SSs at each time frame. As specified in the standard, individual SS measures the SNR and feedbacks the information to the BS scheduler in each time frame. The BS receives the feedback signals from all the SSs to collect the current channel status. With the perfect channel state information, the BS scheduler makes the resource allocation decision for queues and selects the suitable adaptive coding and modulation on each traffic channel.

4.2 Channel Fading Model

An OFDM system transfers a broadband signal into parallel narrowband sub-channels, thus the frequency selective fading can be overcame. Therefore, we adopt the general Rayleigh channel model that is suitable for flat-fading channels as well as frequency-selective fading channels encountered with OFDM [26-27]. The average sensed SNR of SS k can be expressed as $SNR_k = \frac{P_k}{P_n}$, where P_n is the background noise variance [28]. The receiving power of SS k is given by $P_k = |h_k|^2 P_t$, where P_t is the total transmitter power of the BS and h_k is the channel gain, which reflects the effects of several physical phenomena including scattering, obstacles, and multipath propagation. In further detail, the channel gain from the BS can be written as $h_k = \sqrt{c d_k^{-\alpha} S_k} m_k$, where c is a constant incorporating the transmission and receiving antenna gains, d_k is the distance from the BS to user k, α is the path loss exponent, S_k is a random variable for the shadow fading effect, which is known to follow the log-normal distribution with zero-mean and variance σ_s^2 (dB) in the log-scale. The multipath fading effect m_k is modeled as an exponential random variable with a mean 1.0, which represents the Rayleigh fading channel. We also defined the median SNR at the cell edge, ρ, to represent the noise level of the wireless environment considered, which has been applied in previous research [29]

$\because \rho = cD^{-\alpha} P_t / P_n$, where D is the radius of the SS.

$$\therefore \overline{SNR_k} = P_k / P_n = P_t / |h_k|^2 P_n = \rho D^{\alpha} |h_k|^2 / c$$

$$\therefore \overline{SNR_k} = \rho (D/d_k)^{\alpha} S_k m_k^2 = \rho (D/d_k)^{\alpha} S_k$$

For a Rayleigh fading channel, the received SNR is modeled by an exponential random variant [30]. The probability density function is given by

$g(x) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda}) \qquad x \geq 0$, where $\frac{1}{\lambda}$ is the average received SNR.

4.3 Proposed Algorithm Scheduling

4.3.1 Priority Function

For the best effort traffic, the well known PF is an attractive discipline. In this section, the proposed scheduling algorithm, called Adaptive Proportional Fairness (APF) scheduling, is introduced, which aims to extend the PF scheduling to the real time service and provides diverse QoS requirements, The scheduling scheme is based on the Grant Per Type-of–Service (GPTS) principle, which aims to differentiate the delay performance of each queue. A novel priority function is devised for all the QoS guarantee queues, including UGS, rtPS, and nrtPS, for allocating time slots on the queues with the highest priority value. At the time interval *t*, the priority function for queue *i* is defined as:

$$C_i(t) = \frac{W_i(t)}{R_i(t)/(K_i(t)M_i)} \tag{4.1}$$

where M_i is the minimum rate requirement, $K_i(t)$ is the number of connections of the *i*th queue. $W_i(t)$ is the transmission capacity at time *t*, which is determined by the channel quality. As in Table 2, $W_i(t)$ is simplified to the number of bits that can be carried by one symbol via AMC. Each queue corresponds to one QoS requirement class, respectively. The estimated average throughput of queue *i* (denoted as $R_i(t)$) is updated by the simple exponential smoothing model as follows:

$$R_i(t) = (1 - \frac{1}{T_i})R_i(t-1) + (\frac{1}{T_i})r_i(t-1) \tag{4.2}$$

where $r_i(t)$ denotes the current throughput and T_i is a predefined system parameter associated with each class. Rather than making T_i constant in the PF scheme, we assign different values of T_i to each type to perform the different applications. In this thesis, we emphasize on explaining the proposed preference metric available to support diverse services with different QoS requirements. We assume the variation of channel conditions can allow us to select the same T_i for the same service type. The SSs, which are close to the BS, will not sacrifice the transmission of the others.

The key parameter of this algorithm is T_i in Eq. (4.2). The role of T_i is to distinguish the priority for different types of services. The available analysis of Eq. (4.2) has been shown in [31]. With a small T_i, the preference metric of queue i fluctuates significantly, making queue i being visited frequently. This feature is critical to distinct the queue with the delay constraint. For example, in the case T_i=2, according to Eq. (4.2), if the queue has not been served by the scheduler at previous n time frames, $R_i(t) = \frac{1}{2^n} R_i(t-n)$. $R_i(t)$ decreases dramatically, which lead queue i to win the transmission opportunity at the current time frame more likely. In contrary, with a large T_i, (for instance, when T_i=100) $R_i(t)$ decreases slightly although the queue has not been visited for a few time frames. Thus, the channel quality $W_i(t)$ is the main determining factor for the priority of each queue. In this case, the system throughput is enhanced and efficient bandwidth utilization is achieved. However, the scheme is not susceptible to the latency of the connections. With such a design, the task is finding a proper T_i to achieve the balance of delay and throughput for each class according to QoS requirements.

Quantity $R_i(t)/ K_i(t)$ specified in Eq. (4.1) normalizes the throughput of each connection in the ith queue. While $R_i(t)/(K_i(t)M_i)$ is an indicator of data rate satisfaction, which also reflects the delay satisfaction for real time connections. For a real-time service, $R_i(t)$ evaluated with a small T_i is sensitive to the waiting time of the data in queue i. Large value of $R_i(t)$ indicates high degree of delay satisfaction, which leads to low priority. Therefore, a variety of QoS demands for real-time and non-real-time applications are unified to $R_i(t)/(K_i(t)M_i)$, which plays an essential role in reflecting the instantaneous bandwidth requirements of queue i.

Jointly considering the effect of the current channel conditions and the transmission satisfaction of the previous time frames, the proposed preference metric not only keeps the good advantages of PF scheme, but also differentiates the services with diverse QoS requirements by selecting an appropriate set of T_i.

4.3.2 Time Slot Allocation

A cyclic MAC scheduler operates on one frame basis. There are *n* PSs in a given frame, where *n* is determined by the system parameter setting. The average throughput of each queue is tracked by its exponential moving average. At the beginning of each time frame, the BS calculates the preference metric, which is defined as the Eq. (4.1). The queue with the maximum preference metric will be selected for transmission at the next coming PS. Upon each visit, all the data sub-carriers are assigned to the corresponding queue and all data in the served queues are transmitted except that the remaining capacity of the current frame is not large enough to fully accommodate all data. In this way, the throughput of all connections is guaranteed. When the capacity of the current slot is larger than the data in the served queue, the data in the BE queue which belongs to the same SS can be transmitted using the remaining capacity. Once all the PSs in the current frame are exhausted, the rest queues that are not yet processed must be served later. Then, the scheduling process will repeat in the next frame.

Chapter 5
Analysis and Simulation Results

In Chapter 4, we have proposed the new scheduling scheme APF. The performance of the APF scheme is further investigated analytically and via computer simulations with MATLAB in this chapter. A WiMAX network model with a single BS is developed as the system model for the analysis and simulations. First, we examine the capability of the APF scheme to differentiate the QoS of different service classes such as the average delay. Second, the achievable system throughput of the APF scheme is evaluated. After that, the system performance in the presence of traffic load variation is considered. Also, we examine the scalability of the APF scheme to the system size in terms of the number of accommodated SSs. Last, the performance of the APF scheme is compared with three other abovementioned schemes.

5.1 Performance Analysis

We do not consider the channel conditions to simplify the analysis of the effect of T_i. This assumption is reasonable when T_i is rather small. Thus, the preference metric can be $C_i(t)$ rewritten as Eq. (5.1).

$$C_i(t) = \frac{K_i M_i}{R_i(t)} \tag{5.1}$$

where $R_i(t)$ denotes the average throughput at time frame t. When queue i has not been served in the continuous N_i time frames, $R_i(t)$ is updated by Eq. (5.2).

$$R_i(t) = (1 - \frac{1}{T_i})^{N_i} R_i(t - N_i) \tag{5.2}$$

Then,

$$C_i(t) = \frac{K_i M_i}{R_i(t)} = \frac{K_i M_i}{(1-\frac{1}{T_i})^{N_i} R_i(t-N_i)} = \frac{C_i(t-N_i)}{(1-\frac{1}{T_i})^{N_i}}$$

We assume that the values of the preference metric of the winning queue converge to a constant when the system is stable, proved by [33]. An approximate relationship between the inter-service time (denoted as N_i) and T_i is obtained as:

$$\frac{C_1(t-N_1)}{(1-\frac{1}{T_1})^{N_1}} \approx \frac{C_i(t-N_i)}{(1-\frac{1}{T_i})^{N_i}}, \; i \neq 1$$

The approximate relationship between N_l and N_i or D_l and D_i (D_i denotes the average delay of queue i) is given by Eq. (5.3).

$$\frac{N_1}{N_i} = \frac{D_1}{D_i} = \frac{\log(1-\frac{1}{T_i}) + \hat{\delta}_{1,i}(t)}{\log(1-\frac{1}{T_1})}, \; i \neq 1 \qquad (5.3)$$

where $\hat{\delta}_{1,i}(t) = \frac{\log C_1(t-N_1) - \log C_i(t-N_i)}{N_i}$.

The accurate computation for $C_i(t)$ is complicated due to the impact of many factors, such as the traffic load and arrival rate of connections. It must be noted that the value of $(1-1/T_i)$ is so small that $\hat{\delta}_{1,i}(t)$ in the Eq. (5.3) cannot be ignored. Thus, it is quite difficult to establish a mapping between T_i and D_i. Here, a simplified equation reformulated as Eq. (5.4) provides an approximate mapping relationship.

$$\frac{N_1}{N_i} = \frac{D_1}{D_i} = \frac{\log(1-\frac{1}{T_i})}{\log(1-\frac{1}{T_1})}, i \neq 1 \qquad (5.4)$$

5.2 Simulation Results

5.2.1 System Parameter Setting and Assumption

5.2.1.1 Channel Model

As mentioned in section 4.2, all the SSs are assumed to have independent Rayleigh fading channels. The average received SNR of each SS is derived by

$$\overline{SNR_k} = \rho(D/d_k)^{\alpha} S_k$$

Each SS radius D is set to 1(km). The path loss exponent α is 2, the standard deviation of large scale fading σ_s is set to 8 (dB) and the median SNR at the cell edge ρ is 0 (dB), referred to [28]. In simulation, 10 SSs are located. The distances away from the BS are listed in Table 3.

Table 3 Distances from SS to the BS

Index of SSs	1	2	3	4	5	6	7	8	9	10
Distance (km)	3.8	4.5	5	4.5	6	3.8	5	4	5	6

At each time frame, the received SNR is a random variable with an exponential distribution. The optimum modulation and coding rate for the corresponding channels are chosen from Table 2.

5.2.1.2 System Model

The whole bandwidth is assumed to be 20MHz and duration of one time frame is 2.5ms. In the OFDM 256 FFT mode, the duration of each OFDM symbol is 12.6µs and one PS is composed of four OFDM symbols. Thus, a single 2.5ms time frame consists of around 50 OFDM PSs, 20 PSs of them aligned to downlink data transmission. The duration of each frame is so short that it is reasonable to omit the channel fluctuation during one frame for a fixed wireless system. In each SS, there are four queues corresponding to four traffic classes. Each queue has an infinite backlog of data. In the simulation, all traffics arrive at the beginning of each frame. The generator sources K_i of queue i is constant. For each

UGS, rtPS and nrtPS connection, we assume that the arrival process to the queue follows a Poisson distribution with a given arrival rate, which is given in Table 4. For BE service, we assume the queues are always saturated.

Table 4 Input service flow of each SS

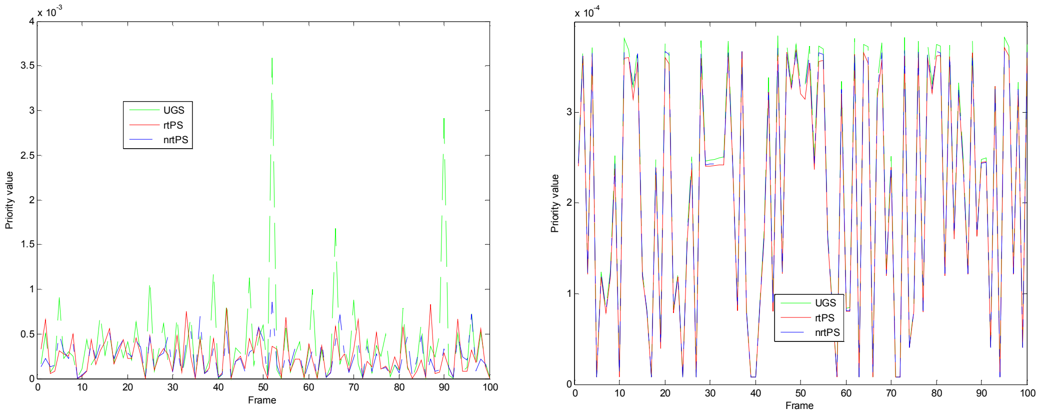
Service Type	**Average arrive rate of each connection (kbps)**	**Min.reserved rate (kbps)**
UGS	9.6	8
rtPS	80	64
nrtPS	5	4

5.2.2 Simulation Results

5.2.2.1 The Performance of APF

A. Capability of Differentiation and the delay performance of APF

We first investigate the effect of different T_i on the service priority to verify the capability of APF to differentiate the service classes. A large value of T_i makes APF scheme roughly behave as PF scheme. The system throughput is high, but the data in the queue may experience large delay. For a real-time application, conforming delay constraint is more critical than improving the long-term throughput. However, for a non-real-time application, the major concern is the long-term throughput. According to the delay constraint of each traffic type, the values of T_i should be set as $T_{UGS} < T_{rtPS} < T_{nrtPS}$. Two examples are illustrated in Fig. 5.1 to analyze the impact of T_i, which present the average priority value of each class in the full load system.



(a) (b)

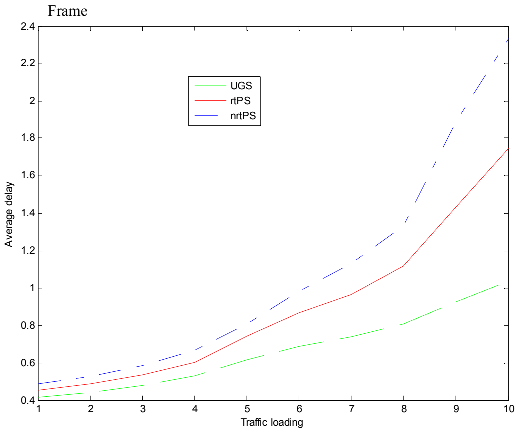
Fig. 5.1 Comparison for Priority value under different T_i sets

(a) T_i of UGS, rtPs, nrtPS are [2, 4, 8]

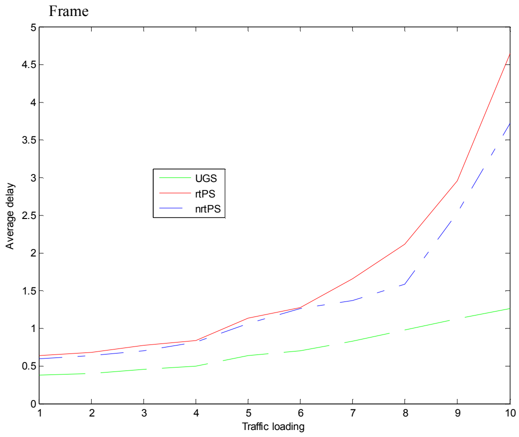
(b) T_i of UGS, rtPs, nrtPS are [200, 400, 800]

According to Eq. (4.1), the peak points in the above figures mean queues are served. As shown in Fig. 5.1, with T_i increasing, the priority values are not sensitive to the parameter T_i. Although the rule $T_{UGS} < T_{rtPS} < T_{nrtPS}$ is enforced and the ratios $T_{UGS} : T_{rtPS} : T_{nrtPS}$ are equivalent in two experiments, it is observed that, in Fig. 5.1 (b), the fluctuation characteristics of the priority values for all types of services are consistent, following the channel condition variation. In this case, the priority of the connection mainly depends on its channel quality. Corresponding to Fig. 5.1 (b), the figures of the related average delay are presented in Fig. 5.2 (b). Obviously, three types of queues cannot be differentiated in terms of the average delay when each T_i is large. In contrast, Fig. 5.1 (a) shows the UGS queues have the smallest average delay while the nrtPS queues have the largest average delay. Therefore, the selected T_i for UGS and rtPS connections should be so small that the priority values are susceptible to the waiting time of data. While the selected T_i for nrtPS should be large enough that the ranking is mainly determined by the channel quality. It is shown that APF can provide the service differentiation among multiple service classes by selecting a proper set of T_i.

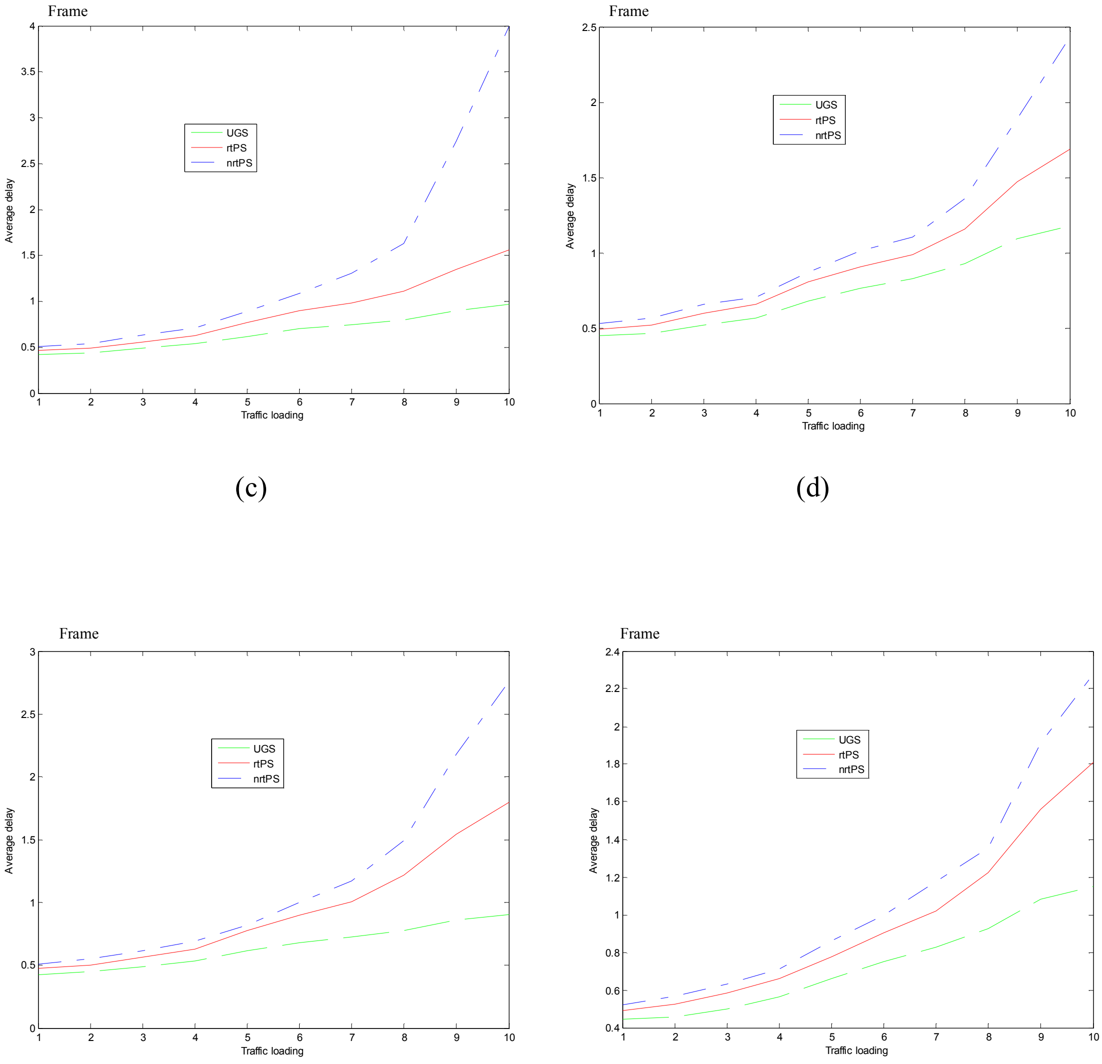
To further verify the effects of T_i, more simulations are conducted, shown in Fig. 5.2 (c)-(f). We change one of the T_i in a set and all other parameters are kept same. By the comparison of the three pairs, (a) and (c), (d) and (e), and (d) and (f), it is observed that the average day increases with the related T_i increasing and the average delay of the rest two service types decreases. It is clear that the average delay is coupled with T_i. This result makes T_i can be manipulated to provide service differentiation in terms of the average delay for each class.



(a)



(b)



(c)

(d)

(e)

(f)

Fig. 5.2 Comparison for Average delay under different T_i sets

(a) T_i of UGS, rtPs, nrtPS are [2, 4, 8],

(b) T_i of UGS, rtPs, nrtPS are [200, 400, 800]

(c) T_i of UGS, rtPs, nrtPS are [2, 4, 100]

(d) T_i of UGS, rtPs, nrtPS are [4, 10, 50]

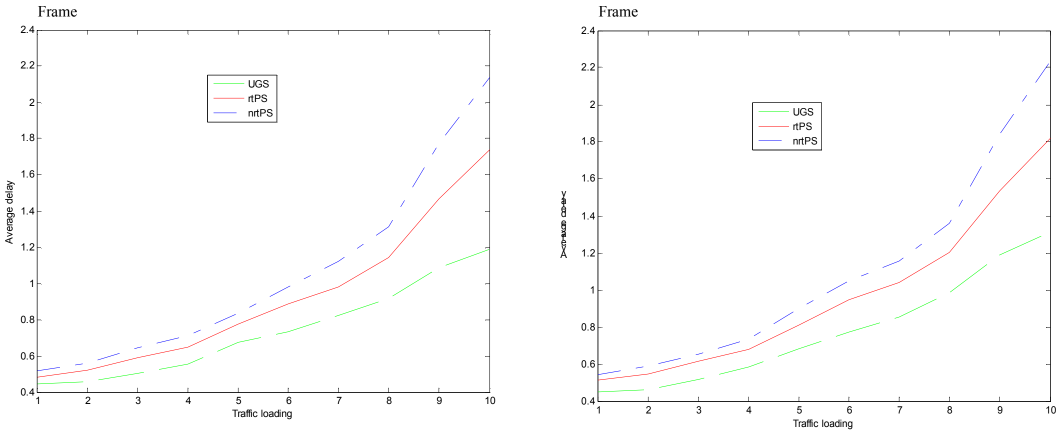
(e) T_i of UGS, rtPs, nrtPS are [2, 10, 50]

(f) T_i of UGS, rtPs, nrtPS are [4, 20, 50]

The following is a case study on examining Eq. (5.4). As seen from Fig. 5.2, Fig. (a), T_i = [2, 4, 8] and Fig. (d), T_i = [4, 10, 50], are very similar according to both the upward trend and the value of the average delay. Among Fig. 5.2 (a)-(f), the values of $\log\left(1-1/T_{ugs}\right):\log(1-1/T_{rtps}):\log(1-1/T_{nrtps})$ of two sets are closest. To further verify the Eq. (5.4), another two sets are selected, which are listed in Table 5 with the corresponding calculated values. With the quite equal ratio of $\log\left(1-1/T_{ugs}\right):\log(1-1/T_{rtps}):\log(1-1/T_{nrtps})$, Fig. 5.2 (a), (g) and (h) have shown three almost identical figures, which confirm us to approximate estimate T_i with the given delay constraint by Eq. (5.4).

Table 5 T_i and the Corresponding Analysis Values

T1	***T2***	***T3***	**log(1-1/*T1*):log(1-1/*T2*)**	**log(1-1/*T1*):log(1-1/*T3*)**	**Fig**
2	4	8	2.41	5.19	Fig. 5.2(a)
4	9	19	2.44	5.32	Fig. 5.2(g)
8	20	40	2.6	5.27	Fig. 5.2(h)



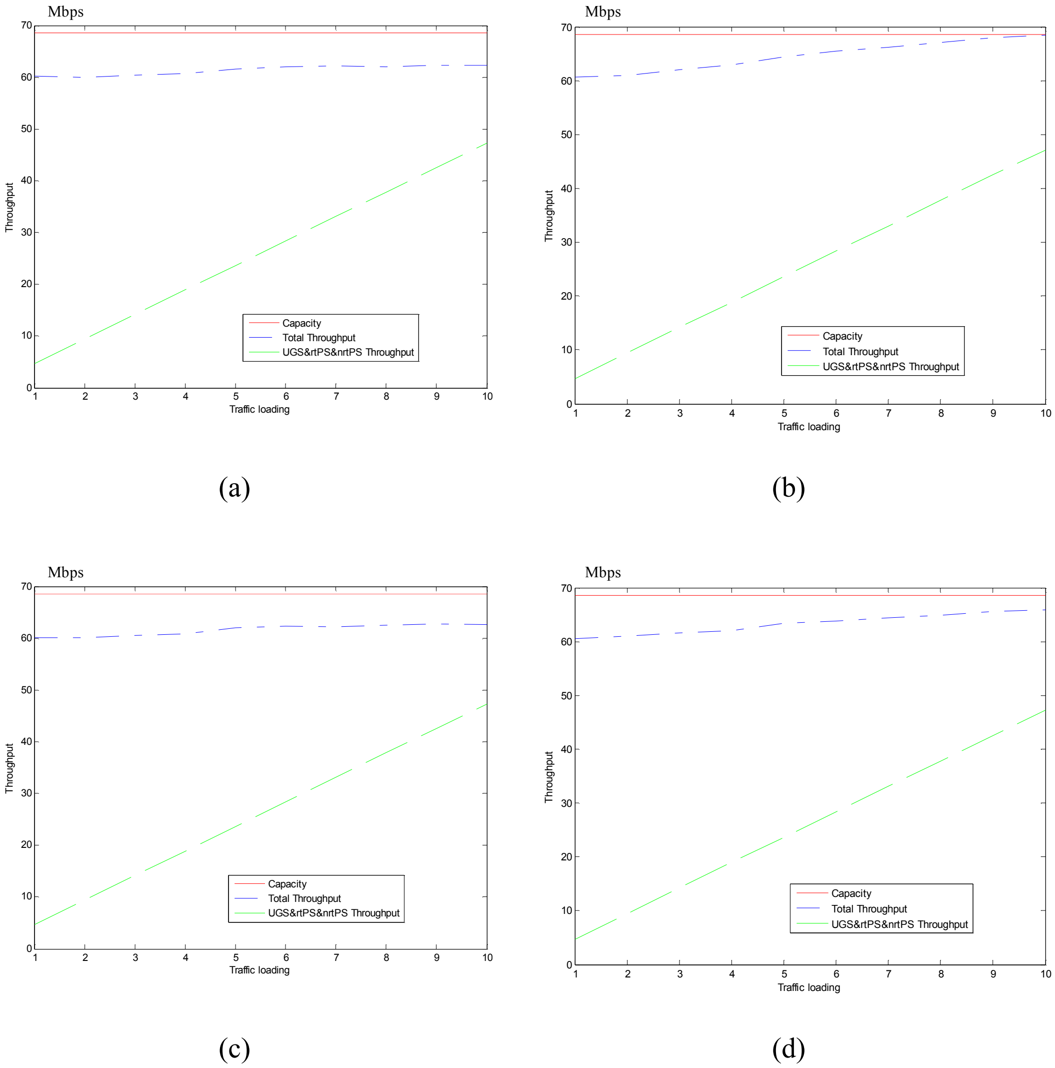
(g)

(h)

Fig. 5.2 Comparison for Average delay under different T_i set

(g) *Ti* of UGS, rtPs, nrtPS are [4, 9, 19]

(h) T_i of UGS, rtPs, nrtPS are [8, 20, 40]



(a)

(b)

(c)

(d)

Fig. 5.3 Comparison for Throughput under different T_i sets

(a) T_i of UGS, rtPs, nrtPS are [2, 4, 8],

(b) T_i of UGS, rtPs, nrtPS are [200, 400, 800]

(c) *Ti* of UGS, rtPs, nrtPS are [2, 4, 100]

(d) T_i of UGS, rtPs, nrtPS are [4, 10, 50]

B. The system throughput of APF

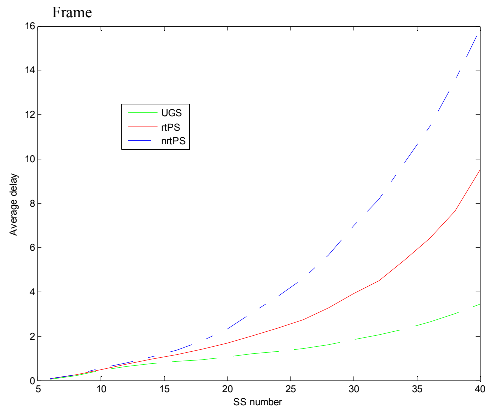
In this section, extensive simulation is conducted to investigate the relationship between T_i and the system throughput. Based on different T_i sets as shown in Fig 5.2, the corresponding capacity, the total throughput, and the cumulative throughput of UGS, rtPS and nrtPS are shown in Fig. 5.3. The system capacities are identical in all figures in Fig. 5.3 due to an ideal status where the selected queues are all in the best channel condition. The total throughput is defined as an effective data rata transmitted from the BS to all the SSs, which is depicted by the blue line. The data streams include UGS, rtPS, nrtPS and BE. Since the selected queue is served exhaustively, the cumulative throughput of UGS, rtPS and nrtPS shown by the green line should be equal to the sum of their arrival data rate. It is observed that the minimum throughput constraint of each connection can be exactly met, while the bandwidth efficiency is much improved compared with that using RR and ICL. However, the choice of T_i does not make much difference in the total throughput. This characteristic makes the manipulation of T_i only base on the delay constraint of each service type for QoS provisioning. Meanwhile, it shows that APF takes the advantage of opportunistic scheduling schemes, enhancing the system throughput regardless of the value of T_i.

C. The impact of the traffic load to APF

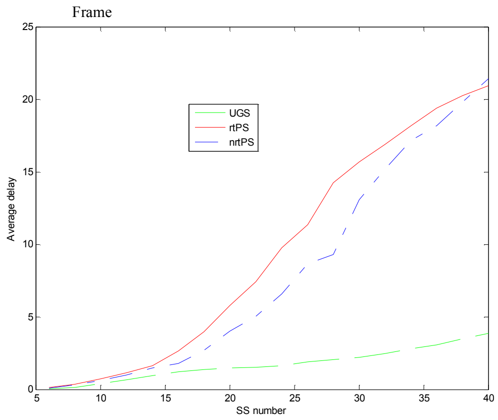
Next, we will investigate the impact due to the variation of the traffic load. Fig. 5.2 and Fig. 5.3 show the relation between the average delay and the throughput with the different traffic load. The Fig. 5.2 indicates that APF does not assign much more privileges to UGS and rtPS when the system is in the light traffic. The differences of the delay performance for the various classes are not apparent. Because when the capacity resource is big enough, all connections have the chance to be served to satisfy the transmission demands. Consequently, all queues achieve the small average delay. In Fig. 5.3, when the system is in the light traffic, the throughput of large T_i is slightly higher than that of the small $T_{i.}$ It is clear that the performance difference among the three service types is not significant in a lightly loaded system. In this case, the advantages of APF are not prominent. In addition, with enough resources, most conventional scheduling schemes can work well, such as RR, PF and etc. While in a heavy traffic system, the limited resource has to be carefully assigned to each connection. The scheduling scheme is especially critical to achieve the satisfied performance for each service. Under such circumstance, APF exhibits its advantage obviously that the scheduler can give the preference to the services in terms of their delay constraints.

D. The impact of the number of SSs to APF

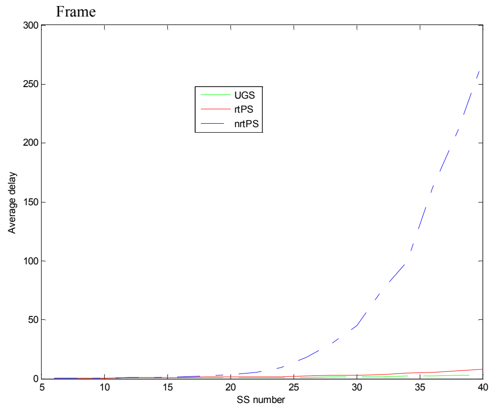
Both pictures in Fig. 5.2 and Fig. 5.3 present the algorithm performance when adding new connections to decrease the available capacity for each flow. The performance of the service classes with the low priority degrades prior to that of the service classes with the high priority. Scalability is achieved. We keep tracking the scalability of the proposed scheme by adding several new SSs, but decrease the number of connections to balance the system in a full traffic load. In this set of experiments, the numbers of SSs are varied from 6 to 40, but the number of PSs in one time frame keeps unchanged. Intuitively, the visited times for each queue decrease in the fixed time duration with the number of SSs increasing and each queue has to wait longer for being visited. Thus, the average delay for each queue increases.


Frame
Average delay
SS number
UGS
rtPS
nrtPS
0
2
4
6
8
10
12
14
16
5
10
15
20
25
30
35
40

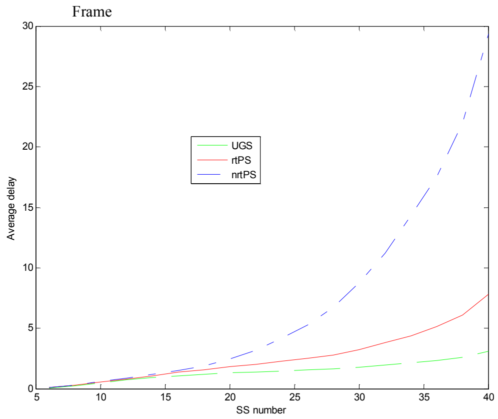
(a)


Frame
Average delay
SS number
UGS
rtPS
nrtPS
0
5
10
15
20
25
5
10
15
20
25
30
35
40

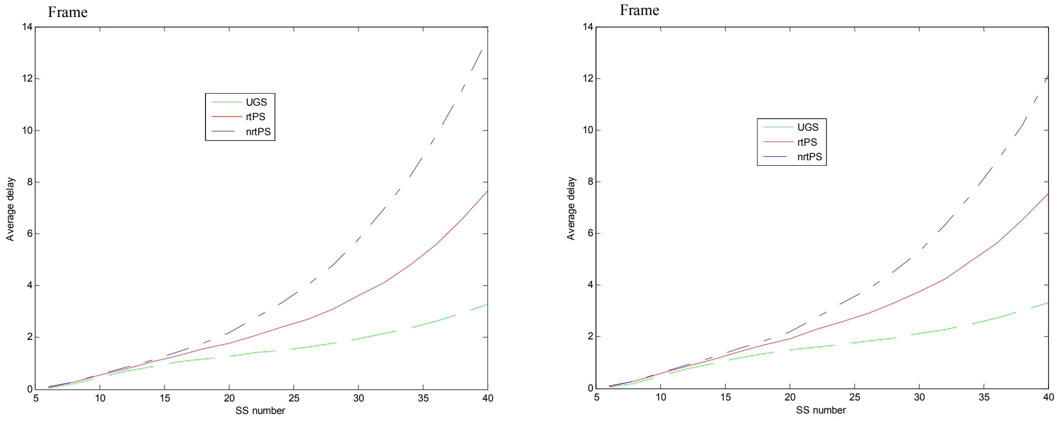
(b)


Frame
Average delay
SS number
UGS
rtPS
nrtPS
0
50
100
150
200
250
300
5
10
15
20
25
30
35
40

(c)


Frame
Average delay
SS number
UGS
rtPS
nrtPS
0
5
10
15
20
25
30
5
10
15
20
25
30
35
40

(d)



(e)

(f)

Fig. 5.4 Comparison for average delay under different T_i sets

(a) T_i of UGS, rtPs, nrtPS are [2, 4, 8]

(b) T_i of UGS, rtPs, nrtPS are [200, 400, 800]

(c) T_i of UGS, rtPs, nrtPS are [2, 4, 100]

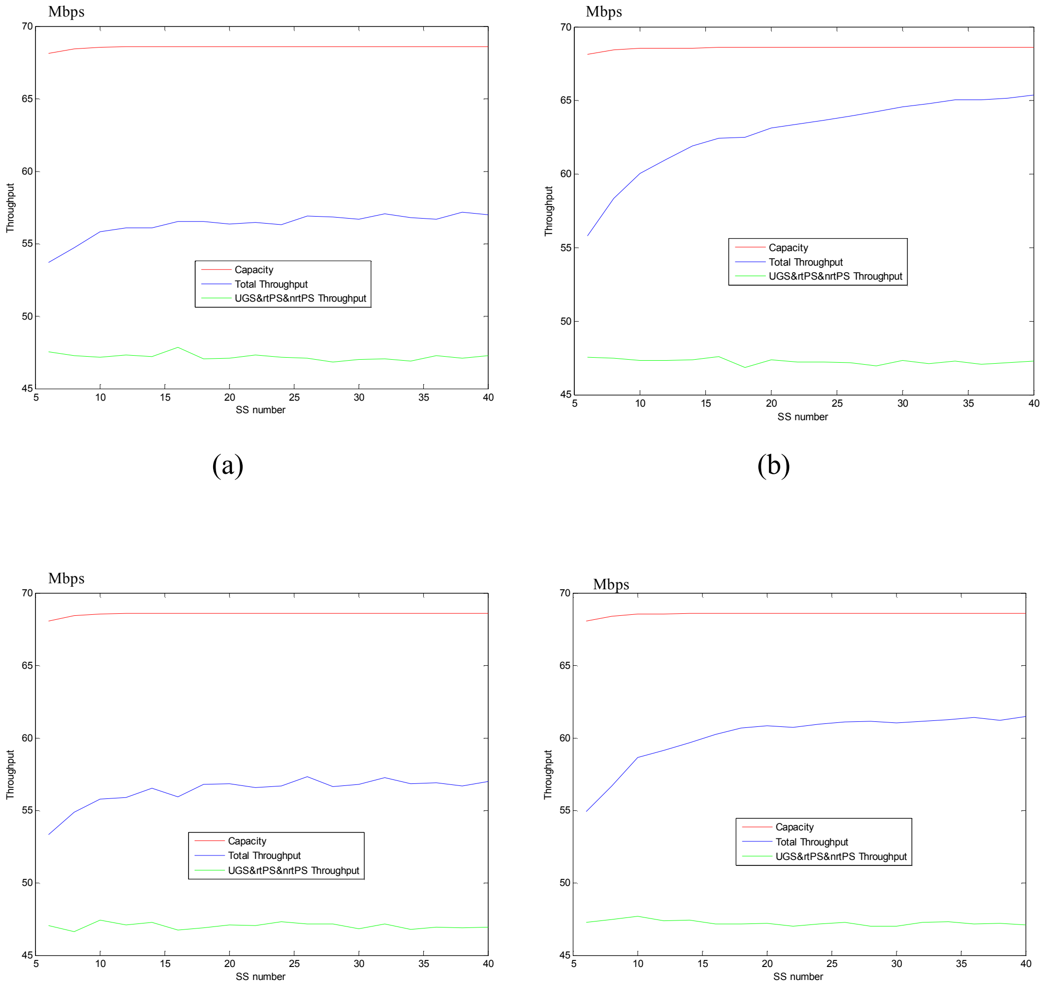
(d) T_i of UGS, rtPs, nrtPS are [4, 10, 50]

(e) T_i of UGS, rtPs, nrtPS are [4, 9, 19]

(f) T_i of UGS, rtPs, nrtPS are [8, 20, 40]

Although the performance of the average delay degrades with more SSs entering the system, the simulation results ensure that the system would not lose priority order for different QoS classes. The QoS requirements of queues in a high-priority QoS class can be satisfied prior to queues in a low-priority class. Simulation results show that all previous analysis and attributes of APF for the impact of T_i setting are held. Therefore, APF is flexible to the system size in terms of the number of accommodated SSs.

We continue with the simulation of the system throughput. By looking at Fig. 5.5, we see that in APF, as the number of users increase, the average throughput increases as well, however, the increase is more dramatic in the small scale system than in the large scale system [28]. When more users are in the wireless network, it is more likely to serve the queues whose channels are near the peaks. APF exploits the benefits of channel fluctuations of independent SSs.

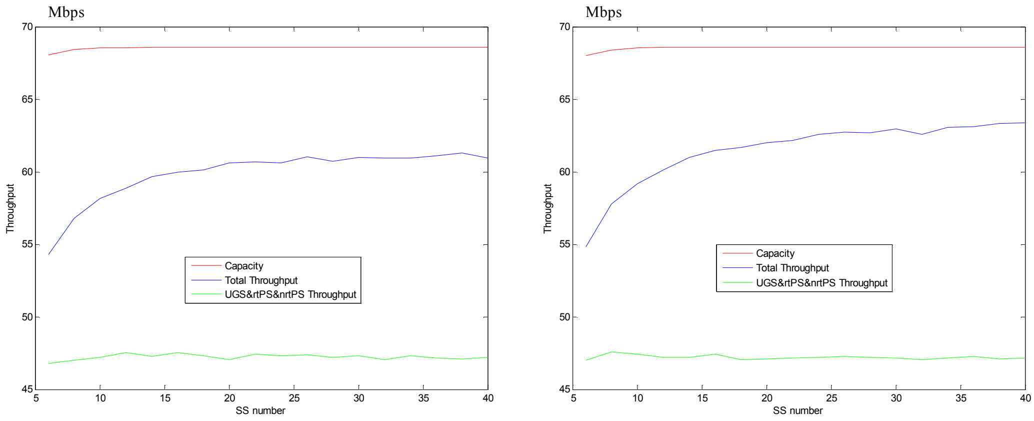


(a)

(b)

(c)

(d)



(e) (f)

Fig. 5.5 Comparison for Throughput under different T_i sets

(a) T_i of UGS, rtPs, nrtPS are [2, 4, 8]

(b) T_i of UGS, rtPs, nrtPS are [200, 400, 800]

(c) T_i of UGS, rtPs, nrtPS are [2, 4, 100]

(d) T_i of UGS, rtPs, nrtPS are [4, 10, 50]

(e) T_i of UGS, rtPs, nrtPS are [4, 9, 19]

(f) T_i of UGS, rtPs, nrtPS are [8, 20, 40]

Up till now, we have analyzed the performance of APF, addressing several important characteristics. Here, we summarize them as follows:

1. APF is to rank all service types by one uniform preference metric, where one parameter T_i to realize the service differentiation. Implementation complexity is low.
2. Controlling the average delay could be realized by dynamically adjusting T_i. In other words, a set of appropriate T_i for all service types can be selected to satisfy the delay constraints.
3. Taking the advantage of the conventional PF, APF can achieve a high system throughput regardless of the value of T_i. The minimum data rate of each type is definitely guaranteed.

4. The advantages of APF are quite obvious when the system is in heavy traffic load and large scale. Scalability is achieved.

In addition, the approach for determining the value of the operation parameter T_i is provided:

1. $\frac{N_1}{N_i} = \frac{D_1}{D_i} = \frac{\log(1 - \frac{1}{T_2})}{\log(1 - \frac{1}{T_1})}$ can be used to approximate estimate the value of T_i.
2. The small T_i can make significant influence on the preference metric. With T_i increasing, the channel condition $W_i(t)$ is a major factor in the preference metric gradually.
3. T_i is a compromise parameter between two competition requirements, the average delay and throughput. T_i of UGS and rtPS should be much smaller than that of nrtPS.
4. Enlarging any T_i of a set increases the average delay of the related queues monotonically and decreases those of the rest two service types.

5.2.3 Comparison with Other Scheduling Schemes

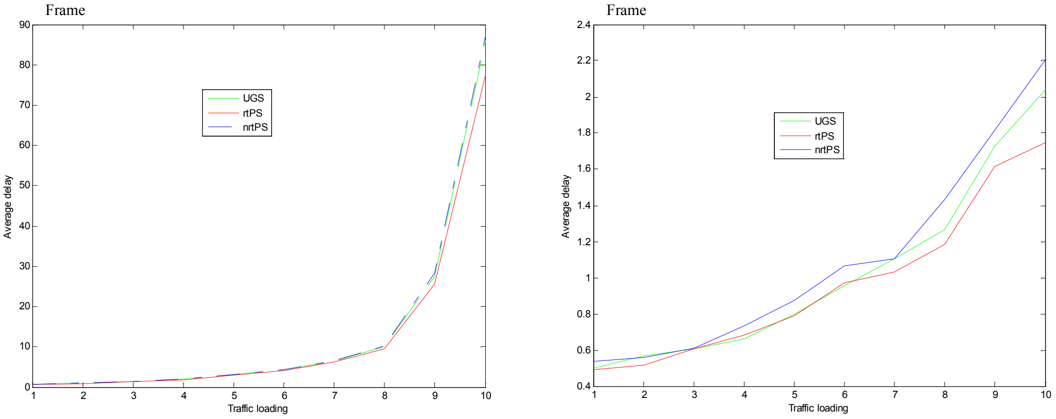
We compare APF scheduling with three schedulers aforementioned in Chapter 3, RR, PF and ICL. The basic algorithms were described. But for being the competitors, the actual scheme implementations are modified to be more robust.

5.2.3.1 Comparison with the RR and PF

RR visits each queue of UGS, rtPS or nrtPS equally. The amount of time slots assigned to each queue is determined according to the request. We do not schedule the resource frame by frame. Each queue is served in turn. The remaining capacity of the current slot will be allocated to the BE queue which belongs to the same SS.

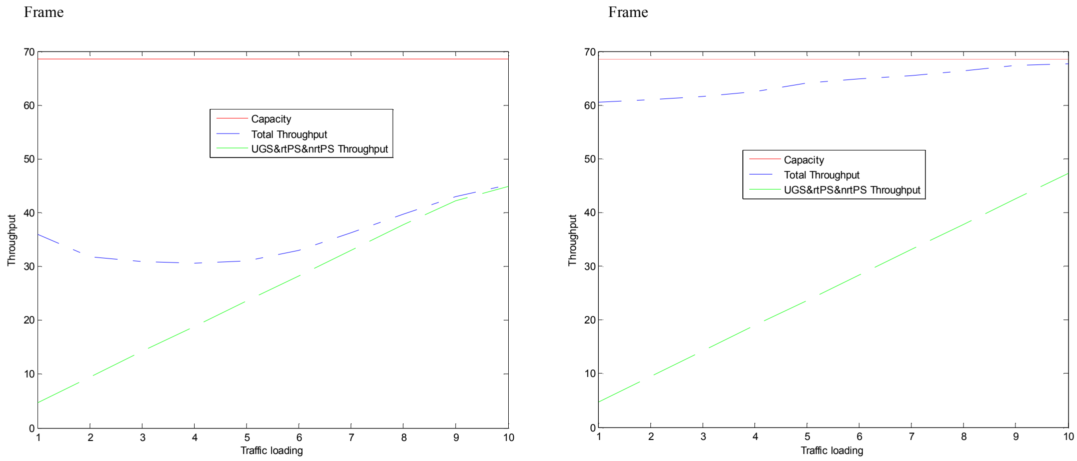
For PF scheduling, the priority metric is defined as an effective ratio of the current data rate to the average service rate: $\mu_i(t) = \frac{r_i(t)}{R_i(t)}$, where $R_i(t)$ is defined iteratively as

$R_i(t+1) = (1-\frac{1}{T_c})R_i(t) + (\frac{1}{T_c})r_i(t)$. Considering the minimum reserve data rates are not consistent for different traffic services, we normalize the average throughput $R_i(t)$ of each connection by dividing the minimum data rate. The new preference metric is rewritten as $\mu_i(t) = \frac{r_i(t)}{R_i(t)/K_i(t)M_i}$. It is quite same with APF except that T_c is a constant for all types of services. In the simulation, we pick up Tc =100.



(a) Average delay under RR

(b) Average delay under PF



(c) Throughput under RR

(d) Throughput under PF

Fig. 5.6 System performance under RR and PF

As we expect, RR and PF cannot prioritize the traffic classes. Fig. 5.6 (a) and (b) show both two scheduling schemes face the difficulties in providing three different delay performances. RR offers same visit times to all queues regardless of their channel conditions, resulting in the fairness and low channel capacity. Fig. 5.6(a) shows the system capacity is lower than that of the exception. Another disadvantage of RR scheme is no multiuser diversity gain. In contrary, Fig. 5.6 (d) shows that the total system throughput under PF scheme is nearly maximized. The inherent diversity gain of PF is analyzed in [32], which proves PF scheme is efficient by jointly considering the channel quality and fairness. Without the consideration of service constraints of different applications, PF is an attractive scheduling. This is the reason that PF has drawn lots of attention.

5.2.3.2 Comparison with the ICL

To resolve the difficulties to support multiple types of services, ICL is designed. The preference metrics are devised for each type of queues. In the simulation, there are fixed reserved time slots allocated for the UGS connections. ICL transmits UGS using RR principle among the SSs. For rtPS and nrtPS services, we follow the algorithm described in section 3.3. However, we do not assign a specific time slot to the BE which only use the remaining resource. The undetermined parameters are summarized as follows:

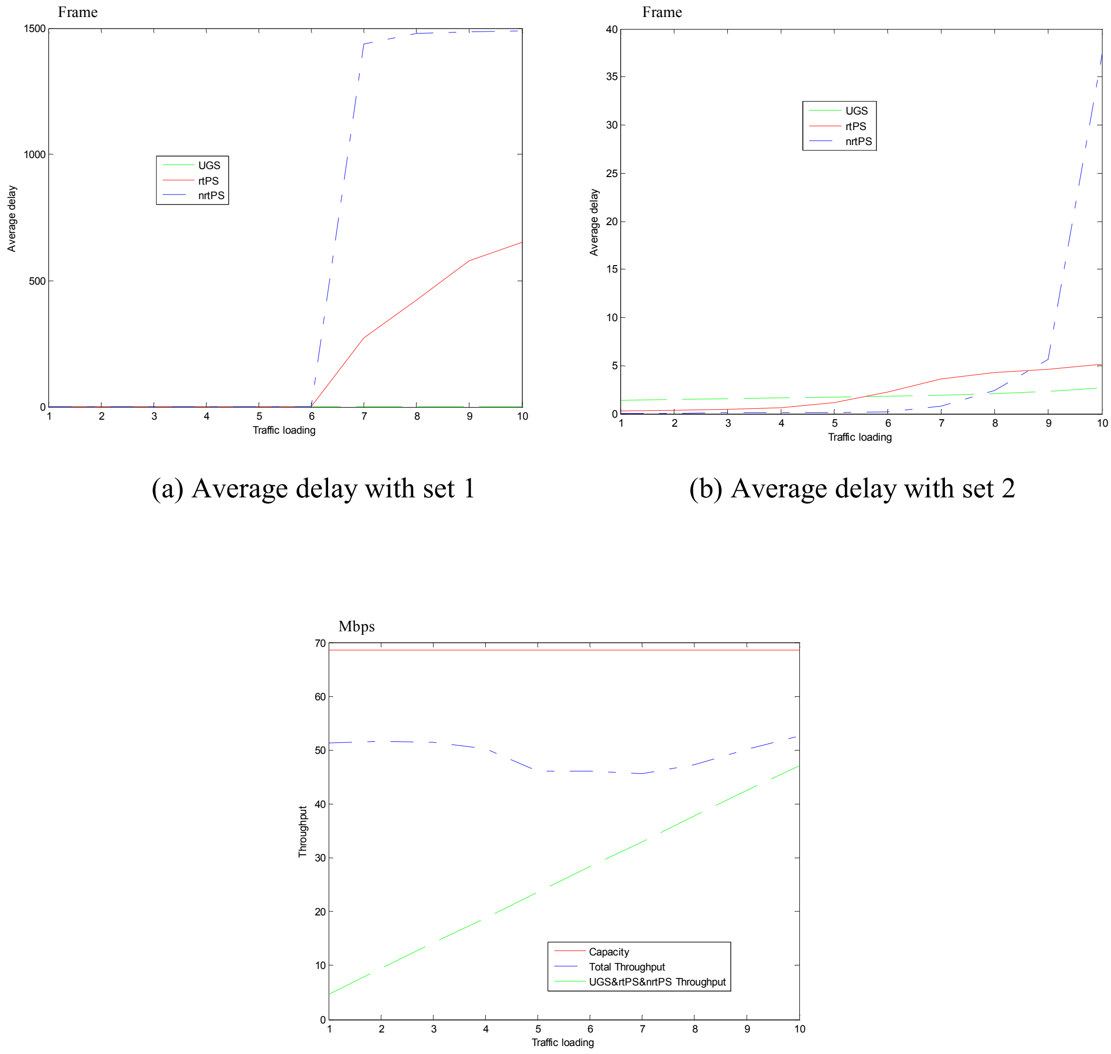
- N_r, the number of time slots reserved to the UGS.
- β_{rt} , β_{nrt} , the coefficient to the rtPS class and nrtPS class. Here, β_{BE} is not considered.
- T_b, the delay bound to the rtPS.
- T_c, the window size for estimating the average transmission rate.

Five parameters to realize the scheduling task are designed. However, [11] only set these parameters heuristically, which did not provide the effective analysis for the effects of the parameters. On the other hand, the preference metrics for rtPS and nrtPS have different physical meanings. It is not proven the rationale to put them together to determine the ranking of the queues.

Table 6 The Parameter Setting in ICL

Set	*Nr*	*βrt*	*βnrt*	*Tb*	*Tc*
1	3	0.8	0.6	10	100
2	3	0.7	0.6	10	100

To evaluate the effectiveness of ICL, the first set of the parameters listed in Table 6 is selected, which is suggested by [11]. The average delay of each service type is shown in Fig. 5.7(a). Although the ICL can distinguish the different types of services as system defined, the delay performance of each type is not good. We deploy extensive simulation with different sets of parameters and find any change among these five parameters yields a great impact to the system performance. The optimal result we can get is illustrated in Fig. 5.7 (b) and the corresponding parameters are listed in the set 2 of table 6. Although with the optimal or near-optimal parameter setting, the system performance in terms of the average delay and the system throughput is worse than that under APF. Without any approach to determine the parameters, it is possible that all the services are not qualified with their QoS requirements. The advantage of this scheme is only the capability to distinct the relative priority.



(a) Average delay with set 1

(b) Average delay with set 2

(c) Throughput with set 2

Fig. 5.7 System performance under ICL

Chapter 6
Conclusion and Future Work

6.1 Conclusion

In this thesis, a novel scheduling algorithm, called APF, for OFDM/TDMA-based systems, has been proposed. Instead of ranking queues with different scheduling schemes, one uniform preference metric is defined to assign a priority for each queue. This preference metric is conceptually based on the PF scheduling which makes it simple to implement. In addition, APF takes into account the QoS satisfaction in terms of average delay for each connection. We further investigate performance of the scheme through extensive simulation. It is observed that the proposed scheme gain merits in the aspects of service differentiation and throughput guarantees for all types of services. Furthermore, the guidelines on how to manipulate T_i with a given delay are provided, validated by simulation. Compared with RR, PF and ICL schemes, APF outperforms in service differentiation and QoS provisioning. Overall, APF is an effective, scalable, and simple scheduling algorithm.

6.2 Extensibility

In this thesis, APF is deployed to schedule the data transmission in downlink, however, the scheme can be readily extended to uplink. Moreover, although we evaluate the performance of APF in an IEEE802.16 network, the developed scheme is a general algorithm for an OFDM/TDMA-based network with various QoS features for different applications. It is suitable for future wireless networks, including cellular networks, IEEE802.11and IEEE802.15 wireless networks.

6.3 Future Work

Scheduling four classes of traffic services over WiMAX system is quite a challenging research topic. There are many issues that should be further investigated.

- For practical implementation, the accurate estimation of T_i deserves further research. We will investigate how the delay constraint can be absolutely guaranteed by manipulating the value of T_i.
- Fairness is another important metric to evaluate the performance of a scheduling scheme. The fairness measurement of APF will be implemented.
- Considering the average channel conditions among the SSs are various from each other, we may set different T_i for different connections to fulfill the QoS provisioning for each flow.
- A general traffic model is assumed in this thesis. However, the traffic model variation affects the performance of schemes. The impact should be investigated.
- The proposed APF performance depends on the perfect CSI. Unreliable channel measurement may result in wrong decisions being made by schemes. It is important to study the effects of imperfect CSI, e.g. estimation error and feedback delay.

It is important to study APF performance with respect to the critical parameters, e.g., T_i, practical traffic model, and the imperfect instantaneous CSI in the AMC module.

Appendix

List of Abbreviations

AMC: Adaptive modulation and Coding

BE: Best Effort

BER: Bit Error Rate

BPSK: Binary Phase Shift Keying

BS: Base Station

CBR: Constant Bit Rate

CSI: Channel State Information

FDD: Frequency Division Duplex

FDMA: Frequency Division Multiple Access

FFT: fast Fourier transforms

FTP: File Transfer Protocol

HDR: High Data Rate

ICL: Integrated Cross-layer scheduling

APF: Integrated Proportional Fairness

LOS: Line Of Sight

MAC: Medium Access Control

MPEG: moving pictures experts group

NLOS: Non-Line Of Sight

nrtPS: non-real-time Polling Service

OFDM: Orthogonal Frequency Division Multiplexing

OFDMA: Orthogonal Frequency Division Multiple Access

PDU: Protocol Data Unit

PF: Proportional Fairness

PHY: Physical Layer

PMP: Point to Multi-Point

PS: Physical Slot

QAM: Quarter Amplitude Modulation

QoS: Quality of Service

QPSK: quaternary phase shift keying

RR: Round Robin

rtPS: real time Polling Service

SCa: Single Carrier

SDU: Service Data Units

SNR: Signal to Noise Ration

SS: Subscriber Station

TDD: Time Division Duplex

TDMA: Time Division Multiple Access

UGS: Unsolicited Grant Service

VBR: Variable Bit-Rate

VoIP: Voice over IP

WLAN: Wireless Local Area Networks

WiMAX: Worldwide interoperability for Microwave Access

Bibliography

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

[2] H.Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-switching Networks," *Proceedings of the IEEE*, vol. 83, pp. 1374-96, Oct. 1995,

[3] R.Guerin and V.Peris, "Quality-of-service in Packet Networks: Basic Mechanisms and Directions," *Computer Networks*, vol.31, pp.169-89, Feb. 1999.

[4] J.Cheng, W.Jiao, Q.Guo, "A fair scheduling for IEEE 802.16 Broadband Wireless Access Systems," *ICC2005*, May 16-20, Seoul, Korea.

[5] Hawa, M, Petr, D.W., "Quality of service scheduling in cable and broadband wireless access system," *Tenth IEEE international Workshop on Quality of Service*, pp. 247-255, 2002.

[6] K.Wongthavarawat, and A.Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," *International Journal Communication System*, pp. 81–96, 2003.

[7] A.Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *Proc. IEEEE VTC* pp. 1854-1858, 2000, Tokyo, Japan.

[8] Bender,P., Black,P.,Grob,M., Padovani, R., Sindhushayana, Viterbi,A. (2000). “CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users,” *IEEE Communication Magazine* 38(7), pp. 70-77, 2000.

[9] P.Viswanath, D.N.C.Tse, and R. Laroia, “Opportunistic beamforming using dumb antennas,” *IEEE Transactions on Information Theory*, vol. 48, no.6, pp. 1277-1294, Jun. 2002.

[10] J.M. Holtzman, “Asymptotic analysis of proportional fair algorithm,” *in Proc. IEEE PIMRC 2001*, San Diego, CA, pp. 33-37.

[11] Q.Liu, S.Zhou, and G.B.Giannakis, “A Cross-layer scheduling Algorithm with QoS support in wireless networks,” *IEEE Transactions on vehicular Technology*, vol.55, No.3, May.2006.

[12] J.F Chen, W.H Jiao, Q.Guo, “An integrated QoS control architecture for IEEE 802.16 broadband wireless access systems”, *Global Telecommunications Conference, 2005. Globecom’05.IEEE*, vol. 6, pp 3330-3335, Dec.2005.

[13] D. Tarchi, R. Fantacci, M. Bardazzi, “Quality of Service management in IEEE 802.16 wireless metropolitan area networks”, in *Proc. of IEEE ICC’06*, Jun.2006, Istanbul, Turkey

[14] C.L.Liu, and J.W. Layland, “Scheduling algorithms for multiprogramming in a hard real time environment”, *Journal of the Association for Computing Machinery*, vol.20, no.1, pp. 44-61, Jan. 1973

[15]Steven J.Vaughan-Nichols, “Achieving Wireless Broadand with WiMAX” *IEEE Communication Magazine*, vol. 37, issue 6, Jun. 2004, pp.10-13

[16] Abichar, Z, Yanlin Peng, Chang, J.M, “The Emergence of Wireless Broadband”, *IT Professional*, vol. 8, issue 4, pp. 44-48, Jul.-Aug. 2006.

[17] C. Eklund, “IEEE standard 802.16: A Technical Overview of the Wireless MAN Air Interface for Broadband Wireless Access,” *IEEE Communication Magazine,* vol. 40, no. 6, pp. 98–107, Jun. 2002.

[18] A.Ghosh, David R.Wolter, Jeffrey G. Andrews and Runhua Chen, “Broadband Wireless Access with WiMAX/802.16: Current Performance Benchmarks and Future Potential,” *IEEE Communication Magazine*, pp129-136, Feb. 2005.

[19] A. J. Goldsmith and S.-G.Chua, “Variable-rate variable-power MQAM for fading channels,” *IEEE Transaction on Communications*, vol. 45, pp. 1218–1230, Oct. 1997.

[20] W. T. Webb and R. Steele, “Variable rate QAM for mobile radio,” *IEEE Transaction on Communications*, vol. 43, pp. 2223–2230, Jul. 1995.

[21] H. Matsuoka, S. Sampei, N. Morinaga, and Y. Kamio, “Adaptive modulation system with variable coding rate concatenated code for high quality multi-media communication systems,” in *Proc. IEEE VTC’96*, pp. 487–491.

[22] IEEE standard 802.16 working group, IEEE standard for local and Metropolitan Area Networks Part 16: Air Interface for fixed broadband wireless access systems, Oct. 2004

[23] J.Karaoguz, “High-rate wireless personal area networks,” *IEEE Communication Magazine*, vol. 39, no. 12, pp. 96-102, Dec.2001.

[24] Cicconetti, C., Lenzini, L., Mingozzi, E., Eklund, C., "Quality of service support in IEEE 802.16 networks," *Network IEEE*, vol. 20, Issue 2, pp.50-55, Mar.-Apr. 2006.

[25] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," in *Proc. IEEE International Conference,* pp. 331–335, Jun. 1995.

[26] Z.Kang, K. Yao, and F. Lorenzelli, "Nakagami-m fading modeling in the frequency domain for OFDM system analysis," *IEEE Communications Letters*, vol. 7, no. 10, pp. 484-486, Oct.2003.

[27] H.Liu, G.Q.Li, "OFDM-based broadband wireless networks: design and optimization", Hoboken, N.J.: *Wiley-Interscience*, c2005.

[28] Kin-Ghoo Choi, Saewoong Bahk, "Cell throughput analysis of the proportional fair scheduler in the single cell environment" *IEEE Transactions on Vehicular Technology*, Mar. 2006.

[29] S.Catreux, P.F.Driessen, and L.J.Greenstein, "Data throughputs using multiple-input multiple-output (MIMO) techniques in a noise-limited cellular environment," *IEEE Transactions on Wireless Communications,* vol.1, no.2, pp.226-234, Apr.2002.

[30] D.Bertsekas and R. Gallager, *Data Networks*, 2nd, Prentice Hall, 1991.

[31] H.J.Kushner and P.A.Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Transactions on Wireless Communications*, 2003.

[32] M.Castaneda, M.Joham, M.Ivrlaˇc, and Josef A. Nossek "Combining Multi-User Diversity with Eigen beamforming in Correlated Channels" In *Proc. European Wireless 2005*, volume 1, pages 138-144, April 2005.

[33] Y. Liu and E. Knightly "Opportunistic Fair Scheduling over Multiple Wireless Channels," In *Proc. IEEE INFOCOM'03*, San Francisco, CA, June 2003.