OPTIMAL PACKET SCHEDULING IN HIGH SPEED DOWNLINK PACKET ACCESS SYSTEM

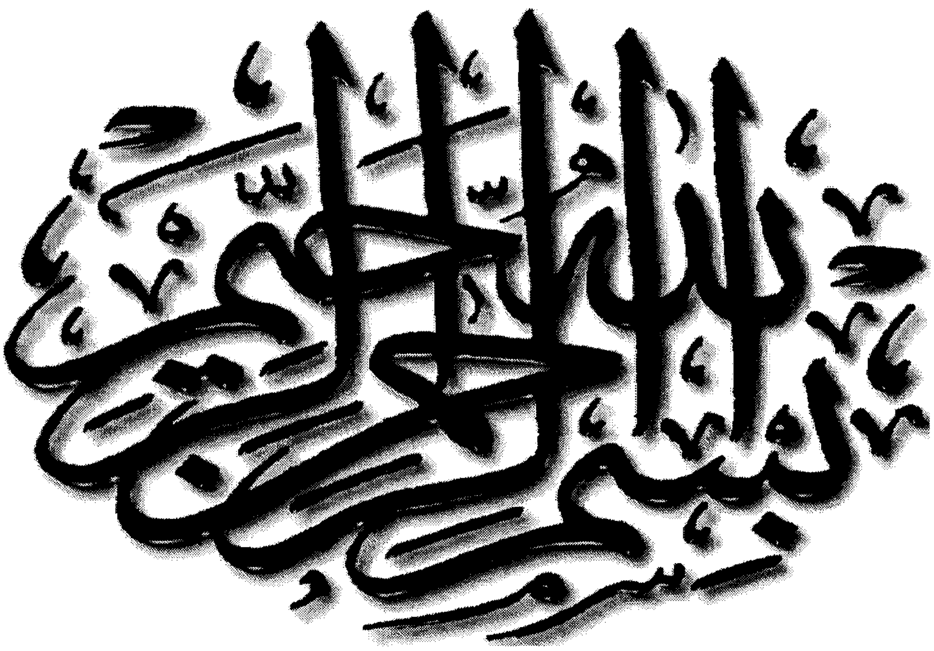
BY

BADER S. AL-MANTHARI

A thesis submitted to the School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
September 2005

Your file Votre référence
ISBN: 978-0-494-30239-2
Our file Notre référence
ISBN: 978-0-494-30239-2



IN THE NAME OF ALLAH,
THE MOST GRACIOUS, THE MOST MERCIFUL

"O my Lord, increase me in knowledge"[1]

[1] The Holy Qur'an, surat "Taha", verse: 114.

To My Parents,
My Family,
And My Dear Friends

Abstract

Future mobile communication forecasts anticipate that there will be an increasing demand for high data rates beyond what 3rd generation wireless cellular systems can offer. To boost the support for such high data rates, a new technology denominated High Speed Downlink Packet Access (HSDPA) has been proposed by the 3rd Generation Partnership Project (3GPP). In HSDPA, a high-speed downlink data channel is shared by multiple users within the same cell to offer peak rates exceeding 10 Mbps. HSDPA relies on many new concepts to achieve such high data rates among which is packet scheduling. The purpose of packet scheduling is to distribute the radio resources among the mobile users in a fair and efficient way to maximize the system throughput. In order to achieve the highest possible throughput, the scheduling algorithm should assign the radio resources to the mobile users with good channel conditions. However, this raises the issue of fairness as those with less favorable channel conditions may not get served and hence suffer from starvation.

In this thesis, we propose a novel Medium Access Control Packet Scheduler (MAC-PS) for HSDPA. The MAC-PS aims at increasing the customers' satisfactions as perceived by the service provider by expressing them by a utility function. It also expresses the service provider's satisfactions by an opportunity cost function; hence providing the freedom to control the throughput-fairness tradeoff. We mathematically show that our MAC-PS converges to the most well known scheduling algorithms, which are the Max CIR and Proportional Fairness (PF); thus giving the service provider the flexibly to choose between different scheduling strategies.

A simulation model is developed to investigate the performance of the MAC-PS algorithm. Our results reveal that the MAC-PS outperforms the Max CIR and PF algorithms in terms of fairness, ability to provide user satisfaction and flexibility in terms of allowing the service provider to control the degree of fairness of the scheduling algorithm and hence the system throughput according to its needs.

Keywords: UMTS, HSDPA, packet scheduling, utility function, opportunity cost and relative fairness.

Acknowledgements

First of all, I thank ALLAH (God) for His mercy, help and guidance, without which this document would not be made possible.

I am grateful to my supervisor professor Hossam Hassanein for giving me the opportunity to be a member of the Telecommunications Research group and to do my Masters under his supervision. Without this expert knowledge, excellent supervision and kindness I could not have finished this challenging work on time.

I am also grateful to my co-supervisor Dr. Nidal Naser from University of Guelph. His experience, knowledge, wonderful personality and most of all motivation and support were a great source of help throughout the course of my research.

I would like also to extend my gratitude to the members of my examination committee, including Dr. Roger Browse, Dr. Nick Graham and Dr. Glen Takahara. Their thorough review of this manuscript, constructive comments and queries and discussions are greatly appreciated.

I would like to express my deepest gratitude to my parents, brothers and sisters, and my friends. Mom and Dad thank you very much for your sincere love. Without your continuous support and prayers this work would not have been accomplished. My sister Rayya, you cannot imagine how I am grateful to you for your unbelievable support, invaluable advice, and standing by me at all times.

I owe Hanady Abdulsalam special thanks for her humongous unconditional help, caring and support. I am also very grateful to my friend Haider Al-Lawati who has generously devoted to me much of his knowledge and time. He has also encouraged me in many occasions and I deeply acknowledge him for that.

I would like to thank all members of the Telecommunications Research Lab and the School of Computing at Queen's University for their support and friendship.

Special thanks to Queen's University and the Ministry of Higher Education in the Sultanate of Oman for their financial support.

Contents

List of Figures

List of Tables

List of Acronyms

2G	2^{nd} Generation Wireless Communication Systems
3G	3^{rd} Generation Wireless Communication Systems
3GPP	3^{rd} Generation Partnership Project
AMC	Adaptive Modulation and Coding
APF	Adaptive Proportional Fariness
CAC	Call Admission Control
CC	Chase Combining
CN	Core Network
CQI	Channel Quality Index
CTCH	Common Traffic Channel
DRC	Data Rate Control
DTCH	Dedicated Traffic Channel
EDGE	Enhanced Data Rates for Global Evolution
EURANE	Enhanced UMTS Radio Access Network Extension for NS2
FC	Fixed Channel
FFQ	Fluid Fair Queuing
FFT	Fast Fair Throughput
FT	Fair Throughput
GSM	Global System for Mobile Communications

HARQ	Hybrid Automatic Repeat Request
HSDPA	High Speed Downlink Packet Access
HS-DPCC	High Speed Dedicated Physical Channel
HS-DSCH	High Speed Downlink Shared Channel
HS-PDSCH	High Speed Physical Downlink Shared Channel
HS-SCCH	High Speed Shared Control Channel
IR	Incremental Redundancy
MAC	Medium Access Control
Max CIR	Maximum Carrier to Interference Ratio
M-LWDF	Modified Largest Weighted Delay First
MPF	Modified Proportional Fairness
MT	Mobile Terminal
MTA-ISIR	Minimum Throughput Assured Instantaneous SIR
NRT	Non Real Time
NS2	Network Simulator 2
OC	Opportunity Cost
Ped A	Pedestrian A
PF	Proportional Fairness
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RLC	Radio Link Controller
RNC	Radio Network Controller
RNS	Radio Network System
RR	Round Robin
RRM	Radio Resource Management
RT	Real Time
SB	Score Based
SIR	Signal to Interference Ratio
SNR	Signal to Noise Ratio
TE	Terminal Equipment

TTI	Time Transmission Interval
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
USIM	UMTS Subscribers Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
Veh A	Vehicle A
WFQ	Weighted Fair Queuing
WFS	Wireless Fair Scheduling

Chapter 1

Introduction

The evolution of today's wireless technology began in the early 1980's with the introduction of the first mobile phones. These systems utilized analog interface technology and supported voice-only capabilities. This technology is still used in some parts of the world; however, it is limited in bandwidth and is low in quality. With the high demand for cell phones and the increased need for enhanced quality and more features, the second generation (2G) was introduced. 2G is primarily voice only, but it does provide higher bandwidth, better voice quality and limited data services that use packet data technology. Furthermore, the continuous success of mobile communication systems as well as its consequences in terms of the need of better Quality of Service (QoS), more efficient systems and more services have led to the development of the third generation (3G) of mobile telecommunications system: Universal Mobile Telecommunication System (UMTS). UMTS is the standard version of 3G mobile systems in Europe [1, 2]. It promises a transmission rate of up to 2 Mbps, which makes it possible to provide a wide range of multimedia services including video telephony, paging, messaging, Internet access and broadband data.

However, it is expected that there will be a strong demand for multimedia applications which require higher data rates above 2 Mbps in cellular systems, especially in the downlink, where mobile users[1] will enjoy high-speed Internet access and broadcast services. In order to offer such broadband packet data transmission services, High Speed Downlink Packet Access (HSDPA) has been introduced in release 5 of UMTS [3]. HSDPA is expected to achieve higher performance with a peak data rate that is about 10 Mbps greater than that of 3G/UMTS systems. In HSDPA, high-speed packet transmission is possible by time-sharing a commonly used data channel among access users, called the High Speed Downlink Shared Channel (HS-DSCH). UMTS already includes a downlink-shared channel but HSDPA extends this concept to provide significantly higher throughput and hence more efficient use of the radio spectrum. The high transmission rate of HSDPA will also allow UMTS to support new high-data rate services and improve the QoS of already existing ones. HSDPA relies on new technologies that make it possible to achieve such high data rate. These new technologies include: Adaptive Modulation and Coding (AMC), Hybrid Automatic Repeat Request (HARQ), and fast packet scheduling.

---

[1] The terms mobile users and users are used interchangeably throughout this document.

Packet scheduling[1] is one of the key design features of HSDPA. A Packet scheduler controls the allocation of channels to users within the coverage area of the system by deciding which user should transmit during a given time interval. Therefore, to a large extent, it determines the overall behavior of the system. One important factor that has been added to the scheduling problem in HSDPA is the channel condition of the mobile users. Mobile users experience varying channel conditions that affect their supportable data rates (i.e. how much data rates they can obtain from the Node B (base station)) from time to time. For example, the further away the user from the Node B, the worse his channel condition (i.e. the lower the data rate he can receive from the Node B). This is because the signal weakens with distance. Other factors that affect the channel condition of a user include mobility, interference caused by other users in the system and obstacles between the user and the Node B that might block the signal. The packet scheduler in HSDPA should track the instantaneous channel conditions of the users and select for transmission those who are experiencing good channel conditions in order to maximize the system throughput. However, this raises the issue of fairness, as the users who have bad channel conditions may not get served and thus they end up having very low average throughputs. Therefore, the design of the packet scheduler should take into account not only the channel conditions of the users, but also their average throughputs by giving the ones who are having very low average throughputs more priorities to increase their chance of getting served and avoid the problem of starvation. That is, the packet scheduler should balance the tradeoff between throughput and fairness.

[1] The terms packet scheduling, scheduling algorithm and packet scheduler are used interchangeably throughout this document.

The rest of this chapter is organized as follows: Section 1.1 outlines the motivations behind our work. The contributions of this thesis are summarized in Section 1.2. Finally, Section 1.3 provides a road map for the rest of the document.

1.1 Motivations

Several scheduling algorithms for HSDPA are already proposed in the literature such as Max CIR [4], Proportional Fairness (PF) [5] and Fair Throughput (FT) [6]. The first two schemes are based on a priori knowledge of the radio channel condition information of each user. They belong to the class of algorithms that prioritize users based on their channel conditions. These types of algorithms tend to maximize the system throughput by assigning all the available radio resources to the users with the best radio channel conditions and therefore they suffer from the fairness problem. Fairness is a major issue in HSDPA since prioritizing the users with best channel conditions may lead to the starvation of those who have less favorable channel conditions. This may greatly affect the revenues of the service provider as a fraction of the total number of users may quit using the service or tend to use it for a short time after they receive a poor service due to their bad channel conditions. Therefore, it may not be beneficial from a revenue point of view to the service provider to implement such algorithms since only a fraction of users (i.e. those with good channel conditions) enjoy the services provided.

FT scheme solves the problem of fairness by ensuring that all users get equal average throughputs. It belongs to the class of the algorithms that prioritize users based on their average throughputs. These algorithms do not make use of the information about the channel conditions of the users and, therefore, even though they have a high degree of fairness, their performance in terms of throughput is very poor. Implementing these algorithms may not also be beneficial to the service provider. This is because packet scheduling in HSDPA depends on the use of the instantaneous channel conditions information of the users in order to achieve the high data rates that it promises. Excluding this information, leads to very low data rates and, hence, nullifying the usefulness of HSDPA, as the system throughput is not increased. Furthermore, implementing such algorithms may not be beneficial from a revenue point of view. Since these types of algorithms try to be fair by ignoring the channel conditions of the users, users tend to receive lower average throughputs. When users receive very low average throughputs, they may stop using the service since they spend longer time in calls/connections and therefore pay more for the service. Furthermore, when users tend to use the service for a longer time due to their low average throughputs, they hold the radio resources with them, which prevents other users that want to access the service from using them. This results in low revenues as less number of users ends up using the offered services. Therefore, implementing such algorithms may not be possible as the cost of deploying HSDPA may be higher than the revenues that the service provider gets in return.

Existing scheduling algorithms do not adapt to combine the channel conditions of the users and their average throughputs in their scheduling decisions. Furthermore, existing algorithms are static in terms of fairness. This means that once the algorithm is implemented, its degree of fairness cannot change. Providing the service provider the flexibility to choose the degree of fairness of the scheduling algorithm (and hence the control of the throughput-fairness tradeoff) is very important to satisfy as many of its customers as possible and hence maximizing its profits. This is because the level to which the degree of fairness of the scheduling algorithm should be such that the maximum number of customers are satisfied and hence the maximum revenues are obtained, depends on the users' locations, age, wealth, education, etc. For example, older customers may use the Internet to check email, read news and other things that do not require high throughput and hence fair scheduling algorithms are most suitable to satisfy as many of them as possible. On the other hand, younger customers tend to download high quality movies that require high throughput and, hence, channel dependent scheduling algorithms though not fair, may be more preferred. Therefore, the service provider may use a fair scheduling algorithm in the locations where older customers live such as urban and suburban areas and a channel dependent scheduling algorithm in the locations where younger customers live such as the places near colleges and universities.

As a result, there is a need to design a packet scheduler that not only combines the channel conditions of the users and their average throughputs in its scheduling decisions but also gives the service provider the flexibility to choose its degree of fairness to satisfy the preferences of the customers from one area to another without having to implement different scheduling algorithms at different places. Such a packet scheduler would not only maximize the customer preferences, but also maximize the profits of the service provider as customers tend to pay more for the services that satisfy them. The design of such a packet scheduler is the focus of this thesis.

1.2 Thesis Contributions

The main contributions of this thesis are:

- Designing a novel generic packet scheduler for HSDPA that takes into account the users satisfactions as perceived by the service provider as well as their channel conditions and their average throughputs in the scheduling decisions.
- Introducing the concepts of utility functions and opportunity cost from economics to express the user's satisfactions as perceived by the service provider and give the service provider the flexibility to choose the degree of fairness of the scheduler based on the user's preferences.
- Formulating the scheduling in HSDPA as an optimization problem involving a utility function subject to opportunity cost constraints.
- Providing a mathematical proof showing that our proposed MAC-PS converges to the Max CIR and Proportional Fairness (PF) schemes as special cases, thus giving the service provider more flexibility to choose between different scheduling strategies.
- Designing and implementing a simulation model to evaluate the performance of our proposed MAC-PS and compare it with the Max CIR and PF algorithms.

The key point that makes our work significantly different from existing schemes in the literature is the formulation of the scheduling problem to meet four objectives: efficiency in terms of using the information about the instantaneous channel conditions of the users in the scheduling decisions, fairness in terms of the distribution of the users' average throughputs, user satisfaction in terms of providing minimum throughput guarantees and flexibility in terms of allowing the service provider to change the degree of fairness of the scheduler. In addition, the way we tackle the scheduling problem distinguishes our work from others'. We believe that this work pioneers new research directions by expressing customer levels of satisfaction as perceived by the service provider with a utility function and by maximizing the service provider profits through the use of an opportunity cost function that allows it to choose the degree of fairness of the scheduler.

1.3 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 provides an overview of UMTS, HSDPA, some of the technologies that HSDPA relies upon to achieve the high data rates and a survey of the current packet scheduling algorithms that are proposed in the literature for HSDPA. Chapter 3 presents our generic Medium Access Control Packet Scheduler (MAC-PS) for HSDPA. Chapter 4 presents the simulation model and the simulation results. We compare the performance of our MAC-PS with that of Max CIR and PF. Chapter 5 concludes this work by summarizing the main contributions, discussing the importance and viability of our approach and suggesting some future research directions.

Chapter 2

Background

Before explaining the design of our novel Medium Access Control Packet Scheduler (MAC-PS) for HSDPA, it is important to provide the background material to help the reader understand the rest of this thesis. Section 2.1 presents an overview of UMTS including the services it offers, its cell structure and the design of its network. Section 2.2 discusses HSDPA and the technologies it relies upon. Comparisons between different packet scheduling algorithms in the literature are discussed in Section 2.3. A summary of this chapter is presented in Section 2.4.

2.1 Universal Mobile Telecommunication Systems

The great success of wireless cellular systems and the increased popularity of packet-switched networks have led to the release of third generation cellular systems. Universal Mobile Telecommunication Systems (UMTS) is the European implementation of 3G cellular networks. The main goal of UMTS is not only to support traditional circuit switched services, but also to support new multimedia services that require stringent Quality of Service (QoS) requirements such as video conferencing, video on demand and online gaming. UMTS can also support other types of services that require less stringent QoS requirements such as web browsing, e-mail access, etc. UMTS is able to support these new applications because of the high data transmission rate that it can offer. Figure 2.1 compares the data rate of UMTS to those of different mobile technologies that are currently in use. As we can see, UMTS can support up to 2Mbps, four times larger than the data rate offered by Enhanced Data Rates for Global Evolution (EDGE) and more than two hundred times larger than the data rate offered by the most widely used mobile technology, Global System for Mobile Communications (GSM).

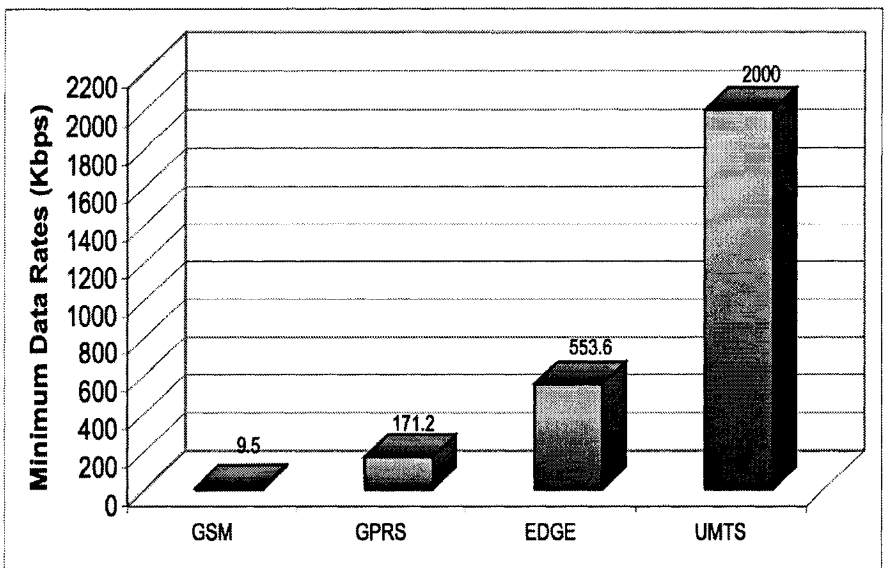


Figure 2.1: Comparisons of Data Rates Offered by UMTS and Other Technologies

This section provides an overview of UMTS. Section 2.1.1 outlines the QoS classes supported by UMTS. The UMTS cell structure is explained in Section 2.1.2. Section

2.1.3 investigates the network architecture of UMTS. Section 2.1.4 outlines the radio interface protocol of UMTS.

2.1.1 UMTS QoS Classes

Since different types of applications have different QoS requirements, UMTS has specified four QoS classes for different types of traffic; namely:

- Conversational class
- Streaming class
- Interactive class
- Background class

The conversational class has the most stringent QoS requirements and the background class has the least. The conversation class is meant for delay-sensitive applications such as voice, video telephony and video gaming. Streaming applications are more delay-tolerant than the first class but they require the end-to-end delay variations (delay jitter) to be within a certain limit. These applications usually delay the received stream (by buffering it) before decoding it and delivering it to the application program in order to cope with delay variations of the stream. Examples of streaming applications are audio streaming, video-on-demand and webcasting. The conversational and streaming classes are referred to as real-time classes. The interactive class represents non-real-time bursty traffic such as in web browsing, database access and e-mail services where the service response time is all that matters. In these applications, the users request some data from a certain server and they expect to receive it within a certain time. The background class is meant for machine-to-machine communications such as server-to-server communications like two mail servers communicating with each other. This class has the least demanding QoS requirements where end-to-end delay could be even in minutes but within a certain limit. The interactive and background applications are often referred to as non-real-time applications.

2.1.2 Cell Structure in UMTS

UMTS was designed not only to support new services, but also to be able to provide global radio coverage. Therefore, a UMTS network is designed and built in a hierarchy of zones of varying coverage (Figure 2.2). Different zones cover different geographical areas. The highest zone consists of satellite systems to cover the whole world. The rest forms the UMTS Terrestrial Radio Access Network (UTRAN). These zones are called pico-, micro-, and macro-zones. Each one of these zones is divided into cells. Smaller cells can support areas where user density is high. Therefore, pico- and micro-cells are used for high user-density areas (hot spots) such as buildings and cities whereas macro-cells are used for land-wide coverage where the density of users is low such as rural areas and deserts.

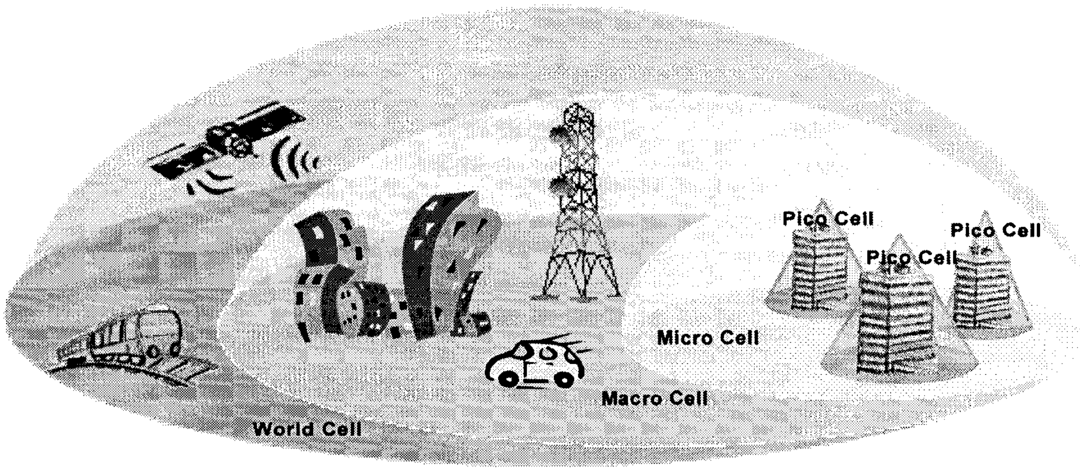


Figure 2.2: Cell Structure in UMTS

2.1.3 UMTS Network Architecture

The network architecture of UMTS is shown in Figure 2.3. It consists of three main entities [7]:

- User Equipment (UE)
- UMTS Terrestrial Radio Access Network (UTRAN)
- Core Network (CN)

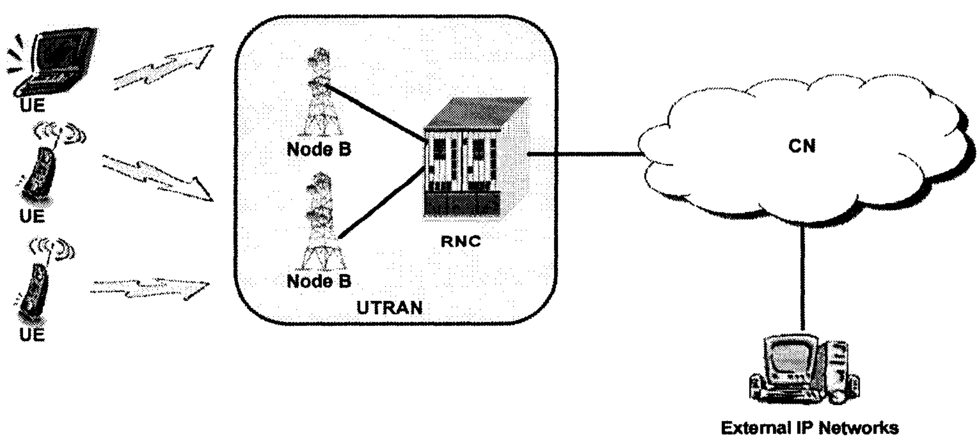


Figure 2.3: UMTS Network Architecture

The UE is the device that provides the user with a direct access to the network services. It consists of the Mobile Terminal (MT), which handles the radio transmission, the Terminal Equipment (TE) which provides the end-to-end application services to the user and the UMTS Subscriber Identity Module (USIM) which provides identification, authentication and encryption services. The UTRAN acts as a bridge between the UE and the CN, thus hiding all functionalities and overhead required to access the CN. The UTRAN is divided into individual Radio Network Systems (RNSs) where each RNS is controlled by a Radio Network Controller (RNC). Each RNC controls a set of Node Bs (base stations). Each Node B is in charge of one or more cells. The RNC implements all the Radio Resource Management (RRM) functions. The main RRM functions are:

- Handoff Management
- Call Admission Control
- Power Control
- Packet Scheduling

Handoff management algorithms decide when the control of an ongoing call should be transferred from one Node B to another as the user moves away from one cell to another cell. Call admission control algorithms decide whether a new call or handoff call should be accepted or not based on predefined factors such as the amount of available power, the number of available channels, and/or whether the current state of the cell exceeds a certain threshold. Power control algorithms determine the optimal amount of power that each user should use to transmit its signals to the Node B and the optimal amount of power that the Node B should use to transmit signals to each user [8]. Finally, packet scheduling allocates resources between users by deciding which user should transmit at any given time interval. The choice of scheduling algorithm has a great impact on the system performance in terms of throughput, packet delay, packet loss, etc. In addition to the above-mentioned functions, the RNC maintains a database, which keeps track of users, their locations, the duration of their calls and all the information that is needed for billing purposes. For a comprehensive overview of RRM functions, see [9].

Node B is in charge of controlling users that are within its coverage area. It also implements the functionalities of the Medium Access Control (MAC) and the physical layer such as packet segmentation and reassembly, modulation, channel coding, etc. Moreover, it provides the RNC with information necessary for performing resource management algorithms such as the amount of received power from each user for the power control algorithm. The CN provides access to the external networks. It consists of two domains, namely the Circuit-Switched Domain (CSD) and the Packet-Switched Domain (PSD) [10]. The CSD provides real-time services such as voice, while the PSD connects to the external data networks (e.g. provides packet access to non-real time services such as web browsing).

2.1.4 UMTS Radio Interface Protocol Architecture

Each entity of UMTS uses different protocol layers to perform its functions. In this thesis, we are concerned about the implementation of the RNC protocol layers. Figure 2.4 shows the three most important protocol layers that are implemented at the RNC: Radio Link

Control (RLC), Medium Access Control (MAC) and the Physical Layer. RLC is responsible for segmentation and reassembly of packets, sequencing, acknowledged or unacknowledged data transfer, transparent data transfer and error control over the radio interface. The MAC layer provides logical channels for the RLC layer. A set of logical channels is defined by the type of information they transfer. Logical channels can be classified as control channels and traffic channels. Control channels are used for transferring signaling messages from higher layers whereas traffic channels are used to transfer user information. There are two types of traffic channels: the Dedicated Traffic Channel (DTCH) and Common Traffic Channel (CTCH). The DTCH is dedicated only for one user whereas the CTCH is used to transfer broadcast messages (to all users) or messages shared by a group of users (common between a group of users). In addition, the MAC layer is responsible for mapping the logical channels into the transport channels, which are in turn mapped to the physical ones, at the physical layer. It is also responsible for scheduling and prioritization of packets originating from different users or from different services of the same user. In order to support the above-mentioned functionalities, the MAC layer is subdivided into the following sub-layers:

- MAC-b to support broadcast channels
- MAC-c/sh to support common and shared channels
- MAC-d to support dedicated transport channels

The physical layer is responsible for the actual transmission of the segmented packets, which are called transport blocks. It does so by mapping the transport channels from the

MAC layer into the corresponding physical ones. The physical channels are structured into radio frames each of which consists of 15 slots where each time slot is 0.667 ms. It then converts the transport blocks into signals, which are then transferred through the antenna. Transport blocks are transmitted every Time to Transmit Interval (TTI) which equals to 10 ms in UMTS.

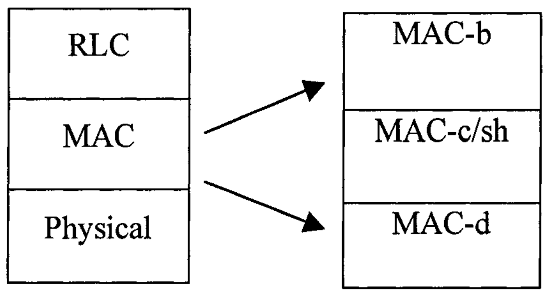


Figure 2.4: The Protocol Layer at The RNC

2.2 High Speed Downlink Packet Access (HSDPA)

The deployment of UMTS in Japan and in some parts of Europe has been a great success. So far, the number of 3G/UMTS subscribers has exceeded 6 million [11]. Figure 2.5 shows the number of UMTS subscribers compared to the world's most famous and widely deployed mobile system, which is GSM. As we can see, the number of UMTS subscribers exceeded that of GSM during the first 10 quarters since first launch. In addition, it is expected that there will be a strong demand for multimedia applications which require higher data rates above 2 Mbps in cellular systems, especially in the downlink, where users will require high-speed Internet access and broadcast services. In order to offer such broadband packet data transmission services, High Speed Downlink Packet Access (HSDPA) was introduced in release 5 of UMTS [3]. HSDPA is expected to achieve higher performance with a peak data rate that is about 10 Mbps greater than that of 3G cellular systems. In HSDPA, high-speed packet transmission is possible by time-sharing a commonly used data channel among access users, called the High-Speed Downlink Shared Channel (HS-DSCH). UMTS already includes a downlink-shared channel but, HSDPA extends this concept to provide significantly higher throughput and hence more efficient use of the radio spectrum. The high transmission rate of HSDPA will also allow UMTS to support new high data rate services and improve the QoS of existing ones. HSDPA relies on new technologies that make it possible to achieve such high data rate. These new technologies include: Adaptive Modulation and Coding (AMC), Hybrid Automatic Repeat Request (HARQ), and fast packet scheduling.

This section investigates some of the most important changes to UMTS made in order to implement HSDPA. Section 2.2.1 highlights the new logical and physical channels that are added to support HSDPA. AMC and HARQ are explained in Sections 2.2.2 and 2.2.3 respectively. Section 2.2.4 discusses the MAC-hs sub layer, which is added to the protocol stack of the Node B in order to support fast packet scheduling. Packet scheduling and its importance in HSDPA is discussed in Section 2.3. For more information about HSDPA, see the 3rd Generation Partnership Project (3GPP) Technical Specification in [3] and overviews of HSDPA in [12] and [13].

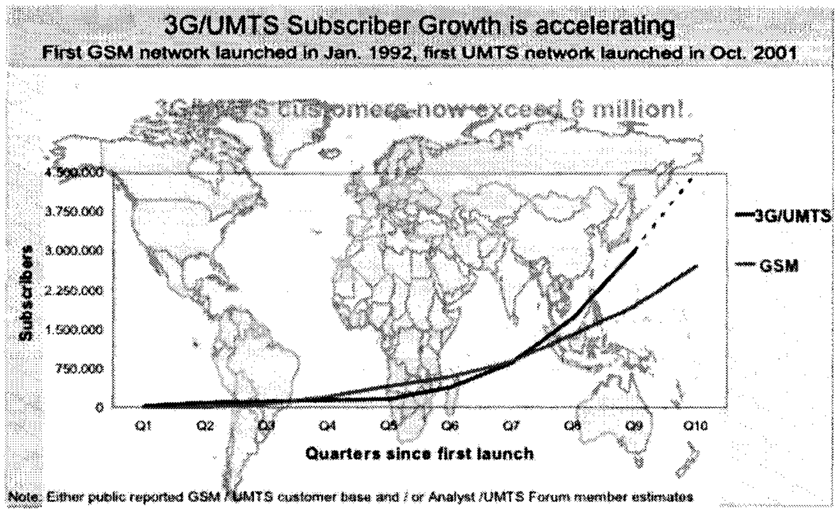


Figure 2.5: Number of UMTS and GSM Subscribers During 10 Quarters Since First Launch [11]

2.2.1 New Logical and Physical Channels in HSDPA

To support HSDPA, new logical and physical channels have been added to the UMTS specification. The logical channels include:

- High Speed Downlink Shared Channel (HS-DSCH)
- High Speed Shared Control Channel (HS-SCCH)

HS-DSCH provides the logical transport mechanism for data transfer. Unlike previous logical channels belonging to release 99 of UMTS (prior to HSDPA), which are located at the RNC, HSDPA logical channels are located at the MAC-hs of the Node B. The reason for this is to allow for quick acquisition of the instantaneous channel quality of HSDPA users and to make fast scheduling decisions based on this, as we shall discuss in Section 2.3. The purpose of HS-SCCH is to provide timing and coding information to the

UE which allows it to listen to the HS-DSCH at the right time (since it is time-shared between multiple UEs) and to use the correct codes for successful decoding of data.

The physical channels that are added to release 5 of UMTS in order to support HSDPA are:

- High Speed Physical Downlink Shared Channel (HS-PDSCH)
- High Speed Dedicated Physical Control Channel (HS-DPCCH)

HS-PDSCH is the transport mechanism for HSDPA logical channels. This channel can be time and/or code shared between HSDPA users. On the other hand, HS-DPCCH is an uplink channel (i.e. for the UE to the Node B) that is used for HS-DSCH signaling. This channel carries the packet acknowledgment signaling for the HARQ process. In addition, it also carries the value of the Channel Quality Indicator (CQI), which is used as an indicator of the channel quality of the user. CQI plays a very important role in packet scheduling (see Section 2.3).

2.2.2 Adaptive Modulation and Coding (AMC)

In previous releases of UMTS, a fast power control was implemented. The main goal of fast power control is to stabilize the quality of the received signal determined by the Signal to Noise Ratio (SNR)[1] by adjusting the transmission power so that more transmission power is used for users with bad channel conditions (weak

[1] The signal strength relative to the background noise, which is used as a measure of the quality of the signal.

received signals). However, this technique increases the overall transmitted power and, therefore, increases the level of interference, which in turn decreases the system's capacity. Therefore, this technique is excluded in HSDPA. Instead, HSDPA adapts the transmitted signal parameters to the continuously varying channel condition by applying a technique known as Adaptive Modulation[1] and Coding[2] (AMC). In HSDPA, every user regularly informs the Node B of his channel quality condition by sending a report known as a Channel Quality Indicator (CQI) in the uplink to the Node B. The CQI contains information about the instantaneous channel quality of the user. This information includes the size of the transport block that the Node B should send to the user, the number of simultaneous channel codes and the type of modulation and coding schemes that the user can support. Node B then would send data to the user at the specified rates. The user is able to measure his current channel conditions by measuring the power of the received signal from the Node B and then using a set of models, viz., the ones described in Chapter 4, to determine his current supportable data rates (i.e., the rates that he can receive data from the Node B given his current channel condition). Therefore, users with good channel conditions will enjoy potentially higher supportable data rates by using higher modulation and coding rates, whereas users with bad channel conditions will experience lower data rates instead of adjusting their transmission power. Release 99 of UMTS already includes a modulation method known as Quadrature Phase Shift Keying (QPSK). In addition to QPSK, HSDPA includes a further modulation technique known as Quadrature Amplitude Modulation (16QAM), which provides higher data rates than QPSK and is used for users with good channel conditions [12].

[1] Superimposing the information bits on the carrier frequency.
[2] Adding controllable redundant information to a block of data to protect it against errors.

2.2.3 Hybrid Automatic Repeat Request (HARQ)

Automatic Repeat Request (ARQ) is a method in which the receiving Node B sends an acknowledgment message (ACK) to the sender for a successfully received data packet. The Node B is able to detect that a packet is received correctly by comparing the checksum that is included in the packet with the one that the Node B calculates. If they do not match, the Node B discards the erroneously received packet and sends a negative acknowledgment (NAK) to the sender. The sender then retransmits the erroneous data packet. This scheme has been modified in HSDPA in order to minimize the retransmission time by using either of two strategies: Chase Combining or Incremental Redundancy (IR) [14].

In the Chase Combining strategy [15], if a packet is detected to be erroneous, the Node B will store it instead of discarding it and it will send a NAK to the sender. If the next packet (the retransmitted packet) happens also to contain errors, the previous packet and the current one are combined in an attempt to recover from the data errors in the original packet. IR is similar to Chase Combining except that additional redundant information is added to the retransmitted packet to increase the chance that the retransmitted packet will be received without errors or with minor errors so that when combined with the previous packet(s), the original packet can be recovered. The redundant information is added incrementally every time a packet is detected to be erroneous for the same reason mentioned above [16].

2.2.4 The Medium Access Control-high speed (MAC-hs) Sub-layer

A significant functionality has been moved to Node B in release 5 of UMTS in order to support the above-mentioned features. A new sub-layer called Medium Access Control-high speed (MAC-hs) has been added to the MAC layer at Node B as shown in Figure 2.6. This new sub-layer is responsible for flow control, distributing the HS-DSCH resources between all users by means of scheduling, enabling execution of the HARQ mechanism, implementation of AMC mechanisms and selection of appropriate transport format and resources. The main reason for moving all the above-mentioned functionality from the RNC to Node B is to reduce the retransmission delays for the HARQ process and to quickly get up-to-date estimates of the channel quality of HSDPA users which will be used in the scheduling decisions as well as in choosing an appropriate modulation and coding scheme for each user. It is worth mentioning that the new functions that are added to Node B will not replace those existing in the RNC. Those functions will still need to exist in the RNC in order to provide non-HSDPA services.

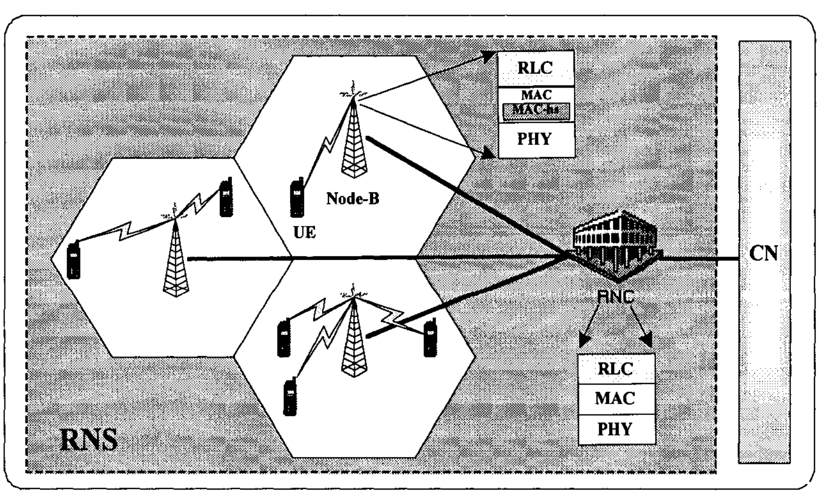


Figure 2.6: MAC-hs at the Node B

2.3 Packet Scheduling in HSDPA

One of the most important features of HSDPA is packet scheduling. Packet scheduling plays an extremely significant role in HSDPA and, therefore, is investigated in more depth in this section. The main goal of packet scheduling is to maximize the system throughput while satisfying the QoS requirements of the users. The packet scheduler determines to which user the shared channel transmission should be given at a given time. In HSDPA, the packet scheduler can exploit the short-term variations in the radio conditions of different users by selecting those with favorable instantaneous channel conditions for transmission, which is illustrated in Figure 2.7. This idea is based on the fact that good channel conditions allow for higher data rates (R) by using a higher order modulation and coding scheme and thus result in increasing of the system throughput.

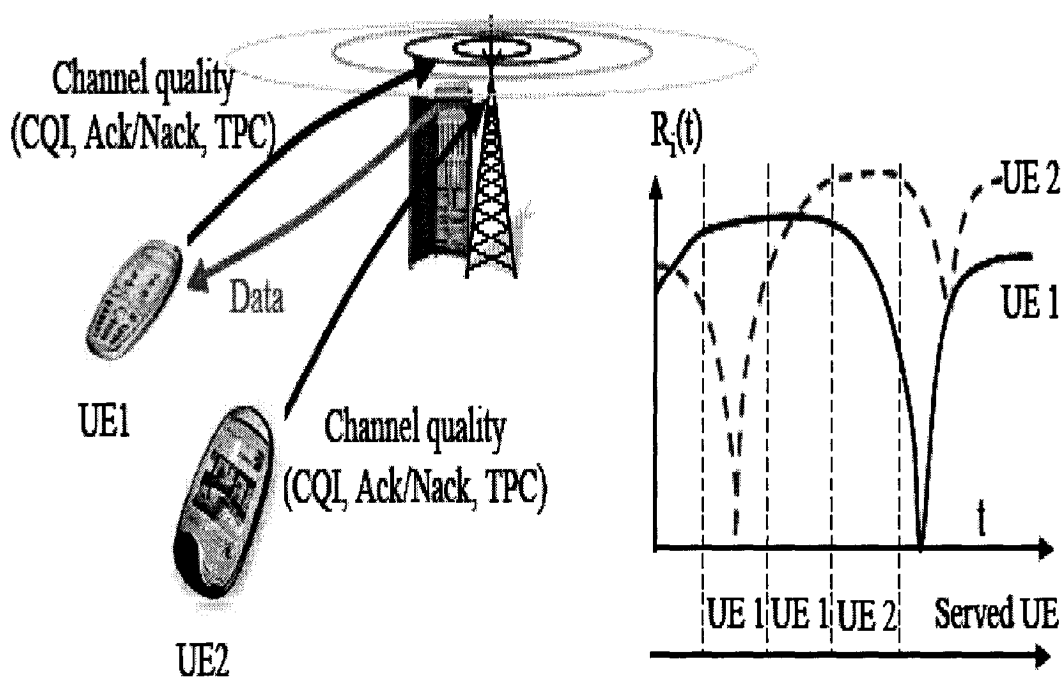


Figure 2.7: Exploiting the User Channel Quality for Scheduling Decisions [14]

In order to quickly obtain up-to-date information on the channel conditions of different users, the functionality of the packet scheduler has been moved from the RNC to the MAC-hs sub layer at the Node B [14]. In addition, the TTI has been reduced from 10 ms in UMTS release 99 to 2 ms in release 5 that includes HSDPA. This is because it allows the packet scheduler to better exploit the varying channel conditions of different users in its scheduling decisions and to increase the granularity in the scheduling process. It should be noted that favoring users with good channel conditions may prevent those with bad channel conditions from being served and may, therefore, result in starvation. A good design of a scheduling algorithm should take into account not only maximization of the system throughput, but also being fair to users who use the same service and pay the same amount of money. That is, scheduling algorithms should balance the trade off between maximizing throughput and fairness.

HSDPA is designed to support the interactive and background applications (non-real-time) and also to some extent streaming applications (real-time). Since real-time applications have different QoS constraints than non-real-time applications, the design of scheduling algorithms for real time applications should be different from the design of scheduling algorithms for non-real-time applications. Therefore, scheduling algorithms can be classified into two groups: Real-Time (RT) scheduling algorithms and Non-Real-Time (NRT) scheduling algorithms. In addition, scheduling algorithms within each group can be characterized by three factors [12]:

- **Scheduling frequency**: the rate at which users are scheduled. The scheduling algorithms that make use of the channel conditions of the users need to make decisions every TTI to better exploit the fast variation of the channel conditions and are therefore called fast scheduling algorithms. Other scheduling algorithms that do not make a decision every TTI are called slow scheduling algorithms.
- **Service order:** the order in which users are served. For example, some scheduling algorithms schedule users based on their channel conditions, whereas others schedule them randomly.
- **Allocation method:** the method for allocating resources. For instance, some scheduling algorithms provide the same data amount for all users per allocation interval, while the others give all users the same (time, code or power) per allocation interval.

The following notations are used to describe various scheduling algorithms:

- UE_i = User equipment i, where $0 < i \leq n$ and n is the total number of users in the cell.
- $S_i(t)$= current average throughput of UE_i up to time t. The current average throughput is the number of delivered bits to UE_i divided by the time from the first arrival of UE_i to the system, t_i, to the current time t. That is, $S_i(t) = B_i(t)/$ (t-t_i) where t > t_i and $B_i(t)$= the amount of successfully transmitted bits to UE_i during the period (t_i, t). The overall average throughput for UE_i is calculated by setting t equal to the time that UE_i is finished using the service. Other averaging methods are used for calculating $S_i(t)$ such as stochastic approximation [17].

- $R_i(t)$ = current supportable data rate for UE_i at time t, i.e., how much data could be transmitted given the current channel conditions of UE_i.
- CQI = Channel Quality Indicator, which is used as an indication of the channel quality for different users. As aforementioned, the Node B is able to calculate CQI from the control information that is exchanged between the Node B and the UEs through the HS-DPCCH. Higher values for CQI mean better channel conditions.
- $P_i(t)$ = priority of flow belonging to UE_i at time t.

NRT and RT scheduling algorithms are reviewed in Section 2.3.1 and Section 2.3.2, respectively. A comparison between different scheduling algorithms is shown in Table 2.1 at the end of the chapter.

2.3.1 Non-Real-Time Scheduling Algorithms for Data Applications in HSDPA

The time-shared nature of the HS-DSCH in HSDPA makes it very well suited for data traffic (i.e., interactive and background), since NRT applications do not require QoS guarantees compared to RT services (even though they expect some form of QoS treatment, those QoS requirements do not need to be guaranteed compared to the RT services in order to provide more flexibility to the conveying network and increase its efficiency). Next, we discuss some of the scheduling algorithms that are proposed for NRT applications. These algorithms have been organized based on their scheduling frequency (fast or slow).

2.3.1.1 Fast Scheduling Algorithms

Fast scheduling algorithms exploit the instantaneous channel quality of the users and try to increase the system throughput by favoring those with good channel conditions. Their name comes from the fact that they need to get the CQI of users every TTI (i.e., 2ms) for computing their priorities to better exploit the channel quality condition. In this section we will explore some of the fast scheduling algorithms for NRT traffic in HSDPA.

Maximum Carrier-to-Interference Power Ratio (Max CIR)

This algorithm tends to maximize the system throughput by serving, in every TTI, the user with the best channel quality (i.e., maximum current supportable data rate) [4]. It can be seen that this algorithm provides high system throughput since only those with high current supportable data rates get served. However, this algorithm has an obvious draw-back in that it ignores those users with bad channel conditions, which may lead to starvation. The unfairness issue in this algorithm has led to many proposals for scheduling algorithms that try to distribute resources evenly among the users while still using the channel condition for user selection for transmission.

Proportional Fairness (PF)

The PF algorithm [5] tries to increase the degree of fairness among users by selecting those with the largest relative channel quality. That is, the priority of user i is calculated as follows:

$$P_i(t) = R_i(t) / S_i(t)$$

This algorithm tends to serve users who have favorable instantaneous radio channel conditions relative to their average throughputs. Therefore, this algorithm considers not only those users with good channel conditions but also those with low average throughputs by giving them higher priority. However, recent studies ([18] and [19]) showed that this algorithm gives more priority to the users with high variance in their channel conditions who thus get more average throughputs than others. However, it provides a better degree of fairness in terms of the distribution of user throughput compared to Max CIR.

Data Rate Control (DRC) Exponent

To solve the problem of unfairness in the PF algorithm, a modified version of the PF algorithm has been proposed in [18]. This algorithm, which is known as Data Rate Control (DRC) Exponent, adds a fixed exponent term C to $R_i(t)$ in order to control the current supportable data rates of the users with different channel conditions. Hence, the priority for each user is calculated as follows:

$$P_i(t) = R_i(t)^c / S_i(t)$$

If C is equal to 1, then this algorithm is the same as the PF algorithm. Higher values for C can be chosen to emphasize the rule of the channel condition in the user selection. However, as stated in [20], there are two main problems with using a fixed value for C. First, having a fixed control parameter prevents it from being adaptable to the instantaneous channel condition of each user. Second, it is not possible to ensure fairness among all users at the same time using a common value C for all users.

Adaptive Proportional Fairness (APF)

This algorithm enhances the DRC algorithm by assigning an independent exponent parameter C_i to each user i instead of using a fixed and common one [20]. APF consists of two modules: a short-term module, which enhances the DRC algorithm, and a long-term monitoring module, which updates the exponent parameter C_i according to variations in the channel conditions in order to ensure fairness among users. The priority of each user i is calculated in this algorithm as follows:

$$P_i(t) = R_i(t)^{c_i} / S_i(t)$$

The scheduling is done every TTI whereas the update of C_i takes place at longer intervals (e.g. 100ms). The update works as follows, If the difference between the relative channel quality of user i (i.e., $R_i(t)\ /S_i(t)$) and the average value over all users, is not within an acceptable range defined by the interval $[-\varepsilon, \varepsilon]$, then C_i along with other parameters are updated according the formulas explained in [20] in order to bring the proportional data rate for user i back to the acceptable range. Simulation results show that this algorithm performs better than the PF in terms of the distribution of user throughput despite the varying channel conditions of different users.

Fast Fair Throughput (FFT)

The purpose of this algorithm is to provide a fair throughput distribution while using the information about the instantaneous channel quality conditions for the selection criterion. This algorithm assigns scheduling priorities as follows [21]:

$$P_i(t) = (R_i(t)/S_i(t))[\max_j\{avg(R_j(t))\}/avg(R_i(t))]$$

where $\max_j\{avg(R_j(t))\}$ is the maximum average supportable data rate from all users in the cell at time t, $avg(R_i(t))$ is the average supportable data rate of user i. FFT tends to equalize the users throughputs in the long run because the $avg(R_i(t))$ term compensates those with less favorable channel conditions by giving them more priority and therefore it distributes the throughput evenly among users [14].

Score-Based (SB)

According to Bonald in [22], the PF scheduler is fair (in terms of the distribution of the user throughput) only in ideal cases where users experience similar channel conditions. However, PF becomes unfair and unable to exploit multi-user diversity[1] in more realistic situations where users usually experience different channel conditions. Therefore, Bonald proposed a Score-Based algorithm in order to overcome this problem. The SB algorithm selects user i with best score:

$$i = \operatorname{argmin}_i L_i(t)$$

Where $L_i(t)$ is the rank at time t of user's i is current supportable data rate $R_i(t)$ among the past supportable data rates observed over a window of size W. That is, $L_i(t)$ is the rank of $R_i(t)$ among the values $\{R_i(t), R_i(t-1), \ldots, R_i(t-W+1)\}$. Therefore, this algorithm selects the user whose current supportable data rate is high relative to his own rate statistics instead of selecting the one whose current supportable data rate is high relative to his average throughput as in the PF algorithm. Simulation results show that the SB algorithm solves the problem of the PF algorithm. However, SB is slightly more complex in terms of implementation than the PF algorithm. In addition, choosing the size

[1] Exploiting the variations of the channel conditions of the users by serving those with more favorable channel conditions for the benefit of user and/or system throughput.

of W might be another problem. Small values for W might not be appropriate to track the distribution of the user supportable data rates while large values for W will increase the time it takes to find the rank.

Minimum Throughput Assured Instantaneous SIR (MTA-ISIR)

This method aims to maintain a minimum throughput for each user by adjusting his SIR if it falls below a certain threshold (Signal-to-Interference power ratio SIR can be used to compute the CQI of the users) [23]. Similar to the Max CIR algorithm, this method chooses the user with the highest CQI (or SIR). However, the CQI of user i is computed as follows:

$$CQI = \begin{cases} CQI & avg(CQI) \geq W \\ CQI \cdot (W / avg(CQI)) & avg(CQI) < W \end{cases}$$

where CQI is the instantaneous Channel Quality Indicator (or SIR at time t), $avg(CQI)$ is the average CQI for user i up to time t and W is a predefined threshold for CQI. Therefore, if the average CQI of user i is less than W, then the CQI is multiplied by a factor so that the $avg(CQI)$ would equal W in order to increase the users chance to get selected for transmission. Bringing the CQI of users with low channel quality to W makes the packet assignment opportunity among users more uniform, thus increasing the degree of fairness. The authors of [23] claim through simulation results that this method resolves the unfairness problem of the Max CIR algorithm and outperforms the PF algorithm in terms of the user throughput.

2.3.1.2 Slow Scheduling Algorithms

Slow scheduling algorithms usually tend to increase fairness among users and hence they do not need to compute the CQI of users every TTI (some of them do use the CQI in their computations but they only compute it every T seconds where T >>TTI). Therefore, these algorithms are known to be slow. Most of these algorithms were proposed for UMTS release 99 prior to HSDPA and some of them were proposed for other systems. In this section, some slow packet scheduling algorithms will be explored.

Round Robin (RR)

Users in this algorithm are served in a round robin (i.e., cyclic order) manner. This algorithm does not make use of the information about the channel quality of users and therefore it may offer low system throughput compared to the fast scheduling algorithms. In fact, simulation results in [14] and [6] show that RR results in low system throughput compared to the Max CIR and the PF algorithms. However, it is a very simple algorithm. In addition, it is fair in that it ensures that all users in the system get equal opportunity of transmission regardless of their channel quality conditions.

Fair Throughput (FT)

The goal of this algorithm is to ensure an equal number of bits received by each user in the system regardless of their channel quality conditions [6]. A possible implementation of FT would be to select at each TTI the user with the lowest average throughput [14]. As in RR, the CQI of each user is not used in the scheduling decision. This algorithm is fair in terms of the distribution of user throughput since each user gets the same amount of throughput regardless of his channel condition. Similar to RR, it suffers from lower system throughout compared to fast scheduling algorithms.

Average CQI

This algorithm serves the user with the largest average CQI every TTI. In this algorithm average CQI is computed over a window of length T where T >>TTI. T= 100ms is used in [14] which makes it a slow algorithm. Note that if T=TTI then this algorithm will be equivalent to Max CIR.

2.3.2 Real-Time Scheduling Algorithms for Streaming Applications in HSDPA

Streaming applications impose strict constraints on the network in order to satisfy their QoS requirements. UMTS uses dedicated channels to support streaming applications as well as conversational services. However, since HSDPA offers a high transmission rate, as well as high spectral efficiency over dedicated channels [24], there has been some research on using the shared channel (HS-DSCH) of HSDPA to support streaming services. Nonetheless, supporting these kinds of services in HSDPA presents many challenges. One of these challenges is that any packet scheduling algorithm must be able to guarantee QoS requirements for streaming users as well as exploiting the information about their instantaneous channel conditions in its scheduling decisions. Guaranteeing the QoS requirements for streaming users is a challenging task especially when the traffic load in the cell is high. Therefore, a Call Admission Control (CAC) mechanism is needed to determine the level of acceptable traffic load such that the QoS of streaming users is guaranteed [14]. In addition, the QoS constraints must be maintained not only in each cell but also during the handoff process.

In this section, we discuss some scheduling algorithms for streaming services. All the scheduling algorithms discussed in this section are fast algorithms. CAC mechanisms that are required to guarantee QoS requirements for streaming users are out of the scope of this thesis. It is assumed that there already exist some CAC mechanisms so that the QoS requirements can be met.

Wireless Fair Scheduling (WFS)

This algorithm first computes what is called the Transmission Candidate Set (TCS) where

$$TCS = \left\{ i; \frac{R_i(t)}{S_i(t)} \geq \gamma, \quad \gamma \textit{ is a given threshold } \right\}$$

That is, only the flows belonging to users where the ratio of their current supportable data rates to their average throughputs (i.e., relative channel quality) at time t is greater than a certain threshold will be considered for transmission [25]. Then the flows in the TCS are served using the Weighted Fair Queuing (WFQ) algorithm. The WFQ is a practical implementation of the Fluid Fair Queuing (FFQ) algorithm where each packet flow is treated as a fluid flow and every non-empty queue is served with a rate that is equal to:

$$\frac{W_i}{\sum_{\forall j} W_j} \cdot C$$

where W_i is the weight of flow i, and C is the total link capacity. WFQ emulates the FFQ, but it does not treat the flows as being fluids but rather with packets with finite granularity. For more information on WFQ, refer to [26].

WFS provides a high degree of fairness since it uses the WFQ algorithm, which is widely used in Integrated Services [26] because of the high degree of fairness it offers. However, it might not be good for streaming services because it does not take into account multi-user diversity as only those who belong to the TCS are served. Moreover, WFS does not take into account the user's packet delay because it schedules flows based on their average throughputs, and not according to their deadlines. This may result in packet discarding [14].

Modified Proportional Fairness (MPF)

The well-known algorithm Proportional Fairness (PF) is modified in this algorithm to accommodate real-time packets. In this algorithm, if the delay for user i is below a certain threshold, then the priority for this user is computed using the PF algorithm; otherwise, the priority is computed using the Fast Fair Throughput (FFT) [21]. Therefore, the priority for user i is computed as follows:

$$P_i(t) = \begin{cases} R_i(t)/S_i(t), & D_i(t) < \tau \\ R_i(t)/S_i(t)[\max_j\{avg(R_j(t))\}/avg(R_i(t))], & D_i(t) \geq \tau \end{cases}$$

where $\max_j\{avg(R_j(t))\}$ is the maximum average supportable data rate from all users in the cell at time t, $avg(R_i(t))$ is the average supportable data rate of user i, $D_i(t)$ is the delay of the packet in the head of the queue of user i at time t, τ is a predefined threshold (usually τ= o.6T where T is the packet due time). Thus, MPF gives more priority for users with packets close to their deadlines in order to prevent them from being discarded. In addition, MFP also increases the priority of those who have low average throughputs by using the PF algorithm.

Modified Largest Weighted Delay First (M-LWDF)

The priority of user i is computed in this algorithm as follows [27]:

$$P_i(t) = \log(\delta_i)(R_i(t)/S_i(t))(D_i(t)/T_i(t))$$

where $\log(\delta_i)$ is a QoS differentiation term to give different priorities for users with different demands on QoS requirements, $D_i(t)$ is the delay of the packet in the head of the queue of user i at time t, $T_i(t)$ is the packet deadline (packet discard time). The term $D_i(t)/T_i(t)$ ranges from 0 to 1. It reaches 1 when the packet is close to its deadline thus giving it more priority. As we can see, this algorithm considers the channel conditions of the users in the cell (represented by $R_i(t)$), $S_i(t)$, and the packet delay, which makes it a good choice for streaming packets. However, if two flows with the same packet delay have different supportable data rates, then they would be assigned different priorities, which may be unfair. To increase the level of fairness in this algorithm, the Fast Fair Throughput (FFT) algorithm should be used instead of the Proportional Fairness (PF) algorithm [14]. That is

$$P_i(t) = (R_i(t)/S_i(t))[\max_j\{avg(R_j(t))\}/avg(R_i(t))](D_i(t)/T_i(t))$$

where $\max_j\{avg(R_j(t))\}$ is the maximum average supportable data rate from all users in the cell at time t, $avg(R_i(t))$ is the average supportable data rate of user i. The FFT algorithm tends to equalize the average user throughput regardless of the average supportable data rate of the user, which increases the degree of fairness.

A comparison of the scheduling algorithms discussed is provided in Table 2.2.

Scheduling method	Services supported	Scheduling frequency	Service order	Allocation Method
Max CIR	NRT	Fast	Highest channel quality	Same resources (time, code, or power)
Proportional Fairness (PF)	NRT	Fast	Highest relative channel quality	Same resources (time, code, or power)
Data Rate Control (DRC)	NRT	Fast	Highest relative channel quality and a fixed exponent term	Same resources (time, code, or power)
Adaptive PF	NRT	Fast	Highest relative channel quality and a dynamic exponent term	Same resources (time, code, or power)
Fast Fair Throughput (FFT)	NRT	Fast	Highest relative channel quality and average bit rate	Same resources (time, code, or power)
Score Based (SB)	NRT	Fast	Highest rank of the current data rate among the past data rates	Same resources (time, code, or power)
MTA-ISIR	NRT	Fast	Highest current and average channel quality	Same resources (time, code, or power)
Round Robin (RR)	NRT	Slow	Round Robin manner	Same resources (time, code, or power)
Fair Throughout (FT)	NRT	Slow	Lowest average throughput	Same data amount
Average CQI	NRT	Slow	Highest average CQI	Same resources (time, code, or power)
Wireless Fair Scheduling	RT	Fast	Relative channel quality exceeding a certain threshold	Resources according to the WFQ
Modified Proportional Fairness (MPF)	RT	Fast	Highest relative channel quality	Same resources (time, code, or power)
Modified Largest Weighted Delay First (M-LWDF)	RT	Fast	Highest relative channel quality and packet due time	Same resources (time, code, or power)

Table 2.1: Comparison of Different Scheduling Algorithms for HSDPA

2.4 Summary

The main goal of UMTS is to offer sufficient capacity and broadband capabilities in order to support new multimedia services such as video conferencing, online gaming, web browsing, etc. UMTS specifies four different classes for four different types of traffic, namely, the conversational class, the streaming class, the interactive class and the background class. These classes have different QoS requirements in terms of end-to-end packet delay, delay jitter, service response time, etc. However, it is expected that the demand for data services will increase in the near future causing significant pressure on the underlying network. Therefore, HSDPA was introduced in release 5 of UMTS. HSDPA offers a peak data rate of up to 10 Mbps compared to 2Mbps in release 99 of UMTS. HSDPA is able to support new services as well as improve the ones already existing in UMTS. New technologies have been added to UMTS in order to support HSDPA. These technologies include: Adaptive Modulation and Coding (AMC), Hybrid Automatic Repeat Request (HARQ) and fast packet scheduling. Packet scheduling plays a very important role in HSDPA since it is responsible for distributing resources to the users in a fair and efficient way and therefore it determines the overall behavior of the system. Packet scheduling in HSDPA is able to take advantage of the information about the instantaneous channel quality of the users in order to maximize the system throughput by serving users only with favorable channel conditions. However, this raises the issue of fairness, as users with bad channel conditions might not get access to radio channels. Therefore, the design of any packet scheduling algorithm should take into account not only maximizing the system throughput but also distributing the resources to the users in a fair way. In this chapter, we explored UMTS and HSDPA. We also explored some of the new technologies that are introduced to support HSDPA with emphasis on packet scheduling in HSDPA. We surveyed some of the packet scheduling algorithms that are proposed for the streaming class as well as those proposed for the interactive and/or background classes.

Chapter 3

Generalized Packet Scheduler

In Chapter 2, we reviewed the packet scheduling problem for the HSDPA system and defined its design objectives and identified the system constraints. We remark that the packet scheduler for HSDPA should prioritize users within the coverage area not only based on their instantaneous channel quality but also based on their average throughputs, hence giving more priority to those with very low average throughputs. Also, the design should be flexible enough to allow the service provider to choose the degree of fairness of the scheduler depending on the environment of operation, the cost versus the gain of deployment, and the user's willingness to pay for the services offered by the service provider based on the quality of those services, etc. In this chapter, we propose a novel Medium Access Control Packet Scheduler (MAC-PS) to provide priority scheduling among users based on their instantaneous channel quality as well as their average throughputs. The proposed MAC-PS uses a utility function to express the user satisfactions as perceived by the service provider. We also introduce the concept of opportunity cost from economics to the scheduling decision to provide the service provider the flexibility to choose the degree of fairness of the scheduler and hence control the throughput-fairness tradeoff. We mathematically show that our MAC-PS converges to the Max CIR or Proportional Fairness (PF) algorithms as special cases, thus giving the service provider flexibility to choose between different scheduling strategies.

The rest of this chapter is organized as follows. Section 3.1 presents the system model and the problem definition. Section 3.2 presents some economics concepts, which are adopted in this thesis, followed by a formulation of the scheduling problem. Section 3.3 illustrates the versatility of our proposed algorithm. Section 3.4 summarizes the chapter.

3.1 System Model and Problem Definition

The system model as shown in Figure 3.1 is composed of a Core Network (CN) and UMTS Terrestrial Radio Access Network (UTRAN), which consists of two components: Radio Network Controller (RNC) and Node B. The RNC can connect to several Node Bs (only one is shown in Figure 3.1).

The CN connects the UMTS network with other external networks such as the Internet and the Public Switched Telephone Network (PSTN). The RNC is in charge of the Radio Resource Management (RRM) as described in Chapter 2. Within the coverage area of a Node B there exists several mobile users that utilize the High Speed Downlink Shared Channel (HS-DSCH) for data transfer. Those users are connected on the uplink by a High Speed Dedicated Physical Control Channel (HS-DPCCH). Every TTI, each user determines his current channel condition and sends the Channel Quality Indicator (CQI) report to the Node B through the HS-DPCCH as we explained in Chapter 2. In Chapter 4, we will explain the models that are used to determine the data rate of the user giving his current channel conditions.

The packet scheduler model is defined as follows. We assume that the Node B serves n mobile users and selects one user for transmission in a slot of some fixed time duration. Also, and without loss of generality, we assume that each user has one connection request. Thus, the Node B maintains one queue for every user. In addition, we consider only non-real-time traffic (e.g., FTP service).

Upon call arrival, the Radio Link Controller (RLC) layer receives traffic in the form of IP packets from higher layers, which are segmented into fixed size Protocol Data Units (PDUs). These PDU's are stored in the transmission queue of the corresponding user in a first-in first-out fashion. Subsequently, the PDUs are transmitted to the appropriate mobile user according to the adopted scheduling discipline.

Let i be the index for user $i, 1 \le i \le n$, where n is the total number of mobile users that are within the coverage area of Node B. Then the scheduling problem can be defined as follows:

Given a set of n mobile users and their corresponding channel quality reports (one CQI for each user) and a set of downlink radio resources at Node B (time slots, codes or power), the scheduling problem is then how to allocate the shared downlink radio resources among the mobile users (i.e., how to select a user from the n data users for transmission) such that the following objectives are met:

- Efficiency
- Fairness
- User Satisfaction
- Flexibility

It should be noted that efficiency here means that the scheduler should take into account the varying channel conditions of different mobile users in the scheduling decisions to maximize the system's throughput. Fairness may have several definitions. For example, it could mean that all users get equal throughputs or all of them get equal chance of transmission (as in the Round Robin algorithm). In this thesis, we define fairness to be in terms of the distribution of users' average throughputs, that is, how the average throughputs of different users are distributed over the real interval. For instance, scheduler A is considered fairer than scheduler B if the average throughputs of the users in A are distributed over a smaller interval than those of B. Moreover, user satisfaction may also have many definitions, including being fair. That is, the fairer the scheduler is, the more satisfied the users are. In this thesis, we define user satisfaction as the percentage of users whose average throughputs exceed a predefined value. This definition is different from our definition of fairness because a scheduler could be fair but may cause many users to have average throughputs that are below a predefined value, and the users therefore are not satisfied. Finally, flexibility means that the scheduler should give the service provider the flexibility to choose the degree of fairness as discussed earlier.

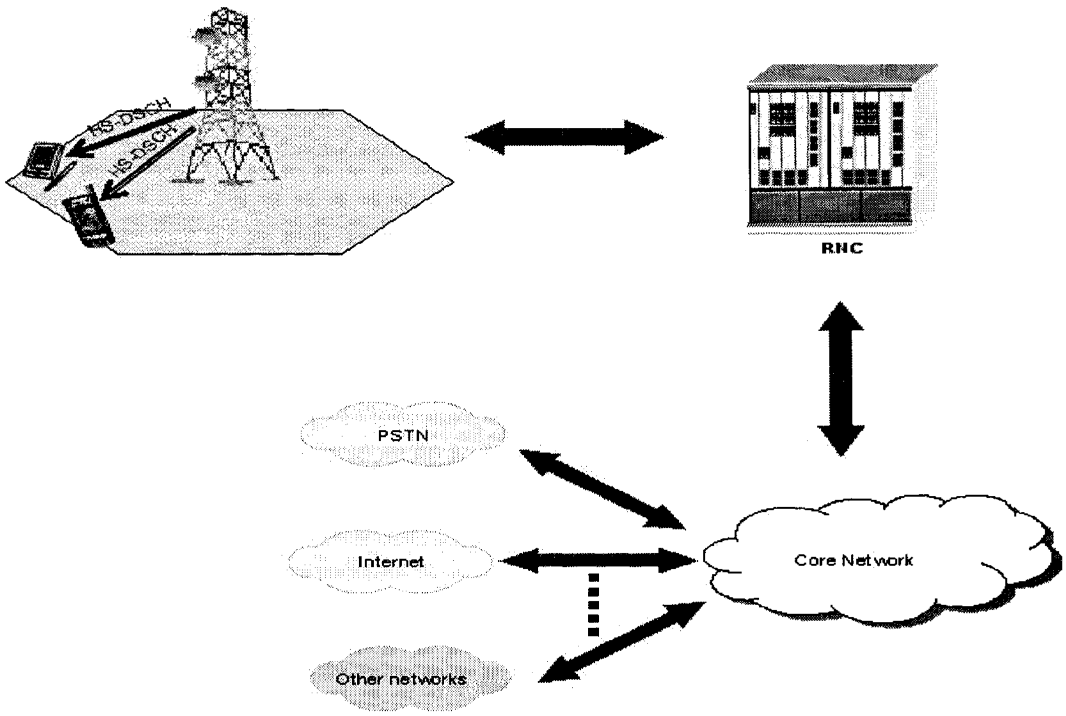


Figure 3.1: The System Model

3.2 Scheduling Algorithm Formulation

In this section, we present and formulate our scheduling algorithm that considers all the objectives discussed in Section 3.1. Before we proceed, we first present some relevant economics concepts, which are adopted to formulate our scheduling algorithm. The rest of this section is organized as follows. Section 3.2.1 provides some background about utility functions and opportunity costs. Our proposed scheduling algorithm and its formulation is explained in Section 3.2.2.

3.2.1 Utility Functions

In this section, we describe the use of a utility function from economics to express a user's preferences for some goods. A utility function U is a way to express consumer preferences by assigning larger numbers to the bundles that that consumer likes more [28]. The bundle $(X_1, X_2, \ldots, X_n)$ is said to be preferred to bundle $(Y_1, Y_2, \ldots, Y_n)$ if and only if the utility of $(X_1, X_2, \ldots, X_n)$ is larger than the utility of $(Y_1, Y_2, \ldots, Y_n)$ (i.e. $(X_1, X_2, \ldots, X_n) > (Y_1, Y_2, \ldots, Y_n)$ if and only if $U(X_1, X_2, \ldots, X_n) > U(Y_1, Y_2, \ldots, Y_n)$). It should be noted that the magnitude of the assignment is not important, what is important is how U orders the bundles of goods. For example, if $U(X_1, X_2) = 5$ and $U(Y_1, Y_2) = 3$, then what matters here is the order of the relation (i.e. 5 > 3), not the values 5 and 3. Therefore, we would have the same preferences if $U(X_1, X_2) = 10$ and $U(Y_1, Y_2) = -4$, since 10 > -4.

3.2.1.1 A Utility Function Example: Perfect Substitutes

Assume that a consumer wants to buy 20 pens and there are only two colors of pens: red and blue, but the customer does not care about the color of the pens (the customer only cares about getting 20 pens). These two goods (red and blue pens) are called perfect substitutes because each one perfectly substitutes the other [28]. The utility function can be written as $U(X_1, X_2) = X_1 + X_2 = 20$ where X_1 represents red pens and X_2 represents blue pens. If the customer likes red pens twice as much as blue pens then a possible expression of the utility function can be $U(X_1, X_2) = 2X_1 + X_2 = 20$.

3.2.1.2 Opportunity Cost

The opportunity cost for a good is defined as the value of any other goods or services that a person must give up in order to produce or get that good [28]. The concept of opportunity cost in economics is used to express other alternatives that exist and which we give up in order to produce or get some good. For example, when going to the movie theater, then the opportunity cost is how much money could have been made if one works for the same amount of time he spent watching a movie. Note that in opportunity cost there is no exact definition for the other alternatives that exist. That is, the alternative could have been any thing that could be done instead of going to the movie theater. In economics, goods are produced as long as the opportunity cost of producing them is less than or equal to the cost of giving them up. This cost could be materialistic such as money, or non-monetary such as happiness. Another example of opportunity cost is buying government bonds. The opportunity cost in this case is buying stocks from a company or saving the money in the bank to get some interest on it.

3.2.2 The Medium Access Control Packet Scheduler (MAC-PS)

In this section, we propose a novel Medium Access Control Packet Scheduler (MAC-PS) that is characterized by the following features: efficiency, fairness, user satisfaction and flexibility. The MAC-PS is located at the Medium Access Control-high speed (Mac-hs) sub-layer at the Node B in conformance with release 5 of UMTS that contains HSDPA.

3.2.2.1 The MAC-PS Formulation

The user satisfactions as perceived by the service provider in HSDPA can be expressed by a utility function U ($X_{i1}(t), X_{i2}(t)$) where U ($X_{i1}(t), X_{i2}(t)$) is the utility function of user i at time t, $X_{i1}(t)$ a quantitative measure of the user satisfactions in this system such as the average throughput, current data rate, average delay, etc, and $X_{i2}(t)$ is a fairness measure that represents how fair the scheduling algorithm is to the user. If we assume that the utility is additive, i.e., the aggregate utility of the system $=\sum_{i}^{n} U(X_{i1}(t), X_{i2}(t))$, where n is the total number of users in the system, then the scheduling algorithm can be formulated to be an optimization problem:

$$\text{Maximize} \sum_{i=1}^{n} \text{U}(X_{i1}(t), X_{i2}(t))$$

$$\text{Subject to OC} \leq \text{K}$$

where OC is a cost function and K is a predefined value. That is, we need to find a user that if served, then the system's utility is maximized. Since the number of users varies with time, and OC may be varying with time, then the optimal solution will also be varying with time. What we can do is that at each TTI (scheduling decision) we can find the user that would maximize the system utility at this TTI. We also introduce the opportunity cost concept to the cost function. The concept of opportunity cost arises here because of the tradeoff between throughput and fairness as described in Chapter 2. If the scheduler tries to be fair, then the system's throughput will decrease. Therefore there is an opportunity cost of fairness. In our case, since we want to maximize $X_{i1}(t)$ (e.g., average throughput) given some fairness measure, then the opportunity cost of fairness would be how much would the system throughput be using this algorithm if we had served the user with the best channel condition given that we served instead another user just to satisfy the fairness measure. For example, let's say that the current data rate for user i, given his current channel condition, is 6 Mbps while the current data rate for user j, given his current channel condition, is 2 Mbps. Assume that after solving U($X_{i1}(t), X_{i2}(t)$), we find that the local maxima is achieved by serving user j, then the opportunity cost is 6-2 = 4 Mbps. Therefore, in order to be fair to users and also increase the system's throughput, then we would maximize U $(X_{i1}(t), X_{i2}(t))$ so that the opportunity cost $\leq$ K Mbps. This means that we have to find a solution such that the opportunity cost does not exceed a certain value, which in our formulation is equal to K.

The definition of the utility function that expresses users' satisfaction (from service provider point of view), the fairness measure that is used in the utility function and the opportunity cost function are presented next.

3.2.2.2 Cobb-Douglas Utility Function for HSDPA Scheduling

We adopt the Cobb-Douglas utility function [28] from economics in our scheduling algorithm. Cobb-Douglas utility function is expressed as U $(X_1, X_2) = X_1^c \cdot X_2^c$, where $c, d \geq 0$. Note that we can apply a monotonic transformation to this function to produce another function that describes exactly the same preferences. For example we could apply the natural log to the Cobb-Douglas function and we get:

$$U(X_1, X_2) = \ln(X_1^c \cdot X_2^c) = c \ln(X_1) + d \ln(X_2).$$

We adopt the Cobb-Douglas utility function as follows. Let X_1 be any performance metric that the service provider wants to optimize such as the average user throughput or average delay. Let X_2 be a fairness measure that increases as the user's or system's perception of fairness increases which results in an increase in U. Then we can express the preferences of user i at time t, $1 \leq i \leq n$, where n is the total number of users in the system, by U $(X_{i1}(t), X_{i2}(t))$ = $X_{i1}{}^{c}(t) \cdot X_{i2}{}^{d}(t)$. In case $X_{i1}(t)$ is high (e.g. the user' average throughput is high) and $X_{i2}(t)$ is low (the system is unfair, for example, only a few users are served) then U will be lower than if both $X_{i1}(t)$ and $X_{i2}(t)$ are high. Therefore, in order to maximize the system's overall utility, we need to achieve the highest possible values for $X_{i1}(t)$ and $X_{i2}(t)$ for all users. However, it is not possible to achieve high values for both $X_{i1}(t)$ and $X_{i2}(t)$ for all users because of the tradeoff between throughput and fairness as we mentioned earlier. In this case, what we can do is to find a user that if served, the system's utility will be maximized. We provide our definition of $X_{i1}(t)$ and $X_{i2}(t)$ in the next section.

We define our fairness measure for user i as follows:

Let S_i (t)= average throughput for user i up to time t.

$\max_j$ S_j(t) = maximum average throughput achieved among all users up to time t.

a = a predefined constant ≥ 0.

Then our fairness measure for user i at time t, α_i (t), is defined as:

$$\alpha_i(t) = S_i(t) / (\max_j S_j(t))^a$$

That is, our fairness measure for user i is defined as the ratio between the average throughput of user i and the maximum throughput achieved among all users in the system. We call this fairness measure "relative fairness". a is a binary variable that is used along with other constants to increase the flexibility of our proposed algorithm as we shall see in Section 3.3. It should be noted that our definition of relative fairness agrees with the definition of fairness we provided at the beginning of this chapter where we defined fairness to be in terms of the distribution of users' average throughputs. This is because if all users achieve close relative fairness values, then this means that their average throughputs are distributed over a smaller interval.

Finally, we define the opportunity cost of serving user i at time t (i.e. the opportunity cost of fairness) as follows:

$$\mathrm{OC}\,(i,t) = \left(\max_j R_j(t)\right) - R_i(t)$$

where R_i (t) = current data rate for user i at time t, and $\max_j R_j$ (t)= the maximum current data rate of all users at time t. Therefore, we define the opportunity cost of fairness to be the difference between the current data rate of user i and the maximum current data rate of all users if user i is selected for transmission. That is, the opportunity cost is how much data rate the system would compromise if user i is selected for transmission. If there is a user j whose current data rate is maximum and the scheduling algorithm chooses user j for transmission, then the Node B would be sending at the maximum data rate and thus the system's throughput would be maximized. However, this would be unfair for other users who have small data rates (i.e., bad channel conditions) and, therefore, the system may choose user i to be served at the expense of high opportunity cost in order to increase fairness between users. The service provider can choose the value of opportunity cost it can tolerate, as we elaborate in the next section.

3.2.2.3 Generalized MAC-PS Algorithm Formulation

In this section, we provide precise definitions of $X_{i1}(t)$ and $X_{i2}(t)$ that will be used in the Cobb-Douglas utility function. We then provide an explanation of this utility function given our definition of $X_{i1}(t)$ and $X_{i2}(t)$. Let:

S_i (t) = average throughput for user i up to time t in Kbps.

$\max_j S_j$ (t) = maximum average throughput achieved among all users up to time t.

R_i (t) = current supportable data rate for user i at time t in Kbps.

α_i (t) = S_i (t) / ($\max_j S_j$ (t))^a, is a fairness measure for user i at time t and a is a binary variable to increase the flexibility of the MAC-PS (see Section 3.3).

OC (i, t) $= \left(\max_j R_j(t)\right) - R_i(t)$, the opportunity cost of serving user i at time t in Kbps.

$\lambda, \sigma, \gamma, b \equiv$ predefined real constants. These constants are used to determine the shape of X_2 in the Cobb-Douglas utility function and also to increase the flexibility of our MAC-PS as we shall see in Section 3.3

K $\equiv$ predefined value for the opportunity cost function in Kbps (i.e., max tolerable opportunity cost).

$c \equiv$ Cobb-Douglas utility function's constant where $c \geq 0$. The value of this constant determines the weight on $X_{i1}(t)$ in the Cobb-Douglas utility function.

$d \equiv$ Cobb-Douglas utility function's constant where $d \geq 1$. The value of this constant determines the weight on $X_{i2}(t)$ in the Cobb-Douglas utility function. The value of this constant must be an odd integer because $X_{i2}(t)$ in our adopted Cobb-Douglas utility function could be negative representing the unhappiness of the user and therefore d must be odd to preserve this.

$n \equiv$ total number of users in the system .

ChC $\equiv$ Channel Capacity.

$X_{i1}(t)$ = R $_i$ (t), the current data rate of user i at time t. The utility of user i being served increases as R $_i$ (t) increases.

$X_{i2}(t)$ = a function of α_i (t)= $\lambda \cdot \ln(\alpha_i(t)) + (1/\sigma^{(1-\alpha_i(t))})^b + \gamma \cdot (1-\alpha_i(t))$,the fairness measure for user i increases as the user's relative fairness increases which increases his utility function. This measure is used to ensure fairness among the mobile users.

Therefore, we can express the utility of user i at time t as follows:

$$\mathrm{U}\,(X_{i1}(t), X_{i2}(t)) = X_{i1}^{\ c}(t) \cdot X_{i2}^{\ d}(t)$$

$$= (\mathrm{R}_i\,(\mathrm{t}))^{c}\,(\lambda \cdot \ln(\alpha_i(t)) + (1/\sigma^{(1-\alpha_i(t))})^b + \gamma \cdot (1-\alpha_i(t))\,)^{d}$$

Assuming that the utility function is additive, then the aggregate utility of the system is:

$$\sum_{i=1}^{n} (\mathrm{R}_i\,(\mathrm{t}))^{c}\,(\lambda \cdot \ln(\alpha_i(t)) + (1/\sigma^{(1-\alpha_i(t))})^b + \gamma \cdot (1-\alpha_i(t))\,)^{d} \qquad (3.1)$$

Given our opportunity cost and the channel capacity constraints (i.e., the sum of all users' average throughputs should not exceed the channel capacity), then at each TTI, we want to find the user that would maximize the following objective function:

$$\text{Maximize} \sum_{i=1}^{n} (R_i(t))^{c} (\lambda \cdot \ln(\alpha_i(t)) + (1/\sigma^{(1-\alpha_i(t))})^b + \gamma \cdot (1-\alpha_i(t)))^{d}$$

$$\text{Subject to} \sum_{i=1}^{n} S_i(t) \leq ChC, \; OC(i,t) \leq K$$

$$\forall_i \text{ where } 1 \leq i \leq n \tag{3.2}$$

It is important to note that at each TTI, the current supportable data rate ($R_i(t)$) of every user i is known and the average throughput $S_i(t)$ (and hence the relative fairness $\alpha_i(t)$) can be calculated by using any throughput averaging method. Therefore, a solution of Eq. 3.2 can found by computing the aggregate utility of the system if user i is scheduled (Eq. 3.1) $\forall i$ and then finding the user with the highest aggregate utility. In other words, a solution of Eq. 3.2 can be found by choosing user i for transmission such that:

$$i = \text{argmax}_i \, [(R_i(t))^{c} (\lambda \cdot \ln(\alpha_i(t)) + (1/\sigma^{(1-\alpha_i(t))})^b + \gamma \cdot (1-\alpha_i(t)))^{d}$$

$$+ \sum_{j=1, j\neq i}^{n} (R_j(t))^{c} (\lambda \cdot \ln(\alpha_j(t)) + (1/\sigma^{(1-\alpha_j(t))})^b + \gamma \cdot (1-\alpha_j(t)))^{d}] \tag{3.3}$$

where user i is selected for transmission and all other users j, $j \neq i$, are not selected for transmission[1] (If user i is selected for transmission then $\alpha_i(t)$ will increase and $\alpha_j(t)$ $\forall j \neq i$ will decrease as other users are not served. See Section 3.2.2.4 for numerical examples).

Two important factors affect the scheduling decision (i.e., the choice of user i). The first factor is the user current supportable data rate ($R_i(t)$). As it can be seen from the

[1] Only one user is scheduled for transmission at each TTI.

function, as the user's current data rate increases, his utility function increases, and in turn his chance of being the one that maximizes the objective function and, therefore, being selected for transmission. The degree to which the current supportable data rate affects the scheduling decision depends on the parameter c. Larger values for c give the current supportable data rate more weight. Note that the user's current data rate is not affected by the choice of the user since it depends on his channel condition and, therefore, is assumed to be intact in Equations 3.1, 3.2, and 3.3 even if the user is not served (i.e., it is not set to 0 if the user is not selected for transmission; hence $X_{i1}(t)$ is not set to zero).

The second factor is the user relative fairness ($\alpha(t)$). In our utility function, $X_{i2}(t)$ is a function of $\alpha_i(t)$ (i.e., $X_{i2}(t) = f(\alpha_i(t))$), and by choosing appropriate values for the constants in the objective function, α can have significant impacts on the scheduling decision. α measures how the user's average throughput is different from the maximum average throughput that is achieved among all the users in the system. When the ratio between the user's average throughput and the maximum average throughput is low then $X_{i2}(t)$ decreases very rapidly, which reflects the fact that the user is unhappy, since he perceives the system as unfair (i.e., at least one user in the system is receiving very high throughput while this user is not). This causes the utility of the user to decrease very rapidly such that if someone with high average throughput is served, though his utility will increase, the overall utility will not be maximized due to the rapid decrease of the utilities of those with low average throughputs. Therefore, the scheduler will be forced to serve those with low average throughputs, since if served, their utility function will sharply increase which results in a maximization of the system's utility even though those users may not have the best channel conditions. Figure 3.2 plots $X_{i2}(t)$ for $0 \leq \alpha \leq 1$ (i.e., from worst relative fairness $\alpha = 0$ to the best relative fairness $\alpha = 1$) for $a=1, b=1$, $\lambda=3$, $\sigma=1.2$, $\gamma=2.5$. As we can see, $X_{i2}(t)$ decreases at a much faster rate as the user's relative fairness decreases from 1 to 0 and it approaches $-\infty$ as the user's relative fairness approaches 0. Therefore, if the algorithm serves a user with high average throughput (e.g., α close to one) then $X_{i2}(t)$ will increase at a much slower rate than if a user with low average throughput (e.g., α close to zero) is served. As a result, the scheduler may choose the user with low average throughput since if served, the increase in the overall utility function will be larger than if a user with good average throughput is served (due to the high rate of increase of $X_{i2}(t)$). The effect that $X_{i2}(t)$ has on the scheduling decision depends on the value of d where larger values of d give $X_{i2}(t)$ more weight.

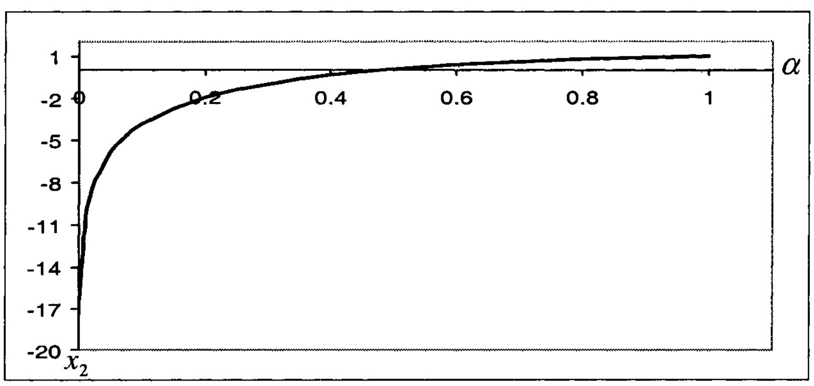


Figure 3.2: A plot of X_2 for $0 \leq \alpha \leq 1$

A flowchart of the MAC-PS algorithm is shown in Figure 3.3. At each time slot t, every mobile user i sends his estimate of his current channel condition to the Node B. This estimate includes the data rate ($R_i(t)$) that this mobile user is able to support giving his current channel condition as explained in Chapter 2. Then, the Node B passes this information to the MAC-PS. The MAC-PS calculates the maximum data rate ($\max_j R_j(t)$) of all mobile users. Next, the MAC-PS calculates the aggregate utility of the system (i.e., Eq. 3.2) if mobile user i is selected for transmission and all other users are not (only one user is selected for transmission at each TTI). This computation is done for every user i in the system. After that, the MAC-PS finds the user with the highest aggregate utility of the system (i.e., Eq. 3.3). Provided that the opportunity cost of serving this user does not exceed the predefined value K (i.e., $(\max_j R_j(t)) - R_i(t) \le K$)), this user is selected by the MAC-PS for transmission at this time (slot t). Otherwise, the MAC-PS selects the user with the next highest aggregate utility.

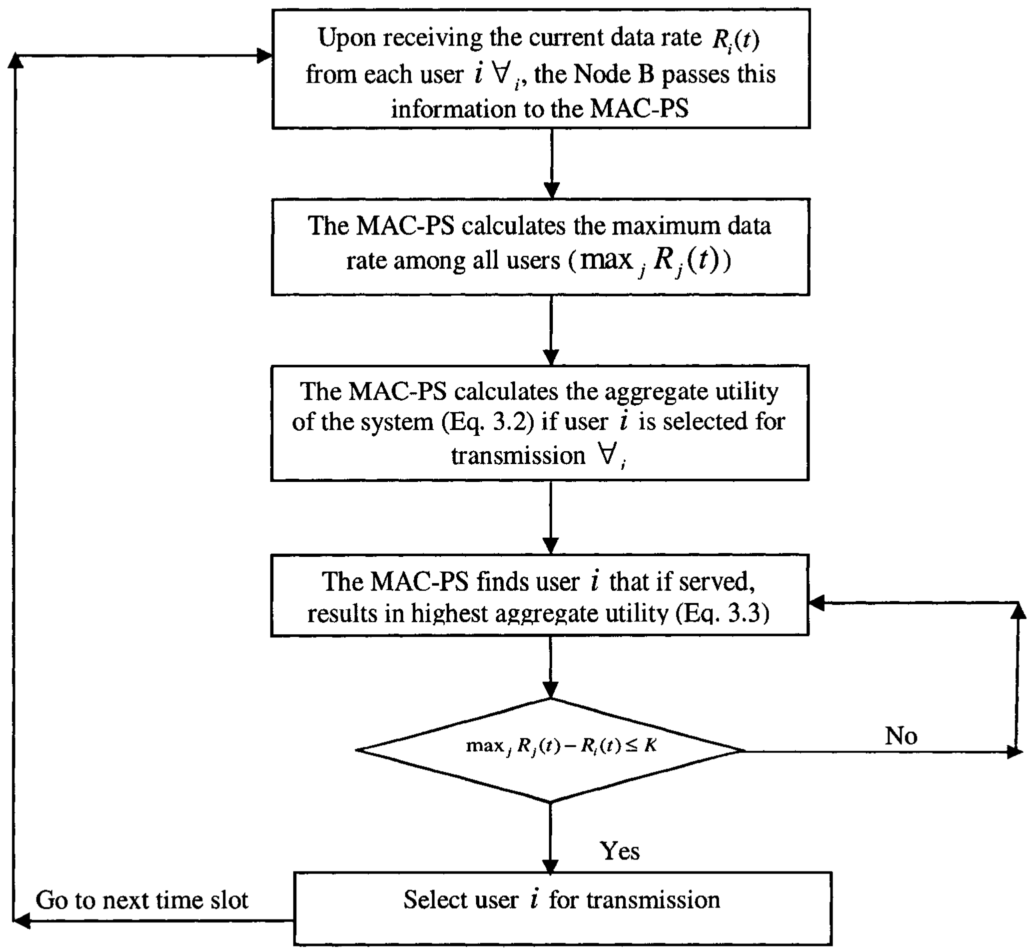


Figure 3.3: The MAC-PS Algorithm

3.2.2.4 Illustrations of the MAC-PS Algorithm

In this section, we illustrate by numerical examples how our MAC-PS algorithm schedules HSDPA users. We show the effects of relative fairness, opportunity cost and current data rates in choosing a user for transmission at any given time interval. We compare our proposed algorithm with the Max CIR and Proportional Fairness (PF) algorithms. Detailed explanations for the results shown in this section are provided in Appendix A.

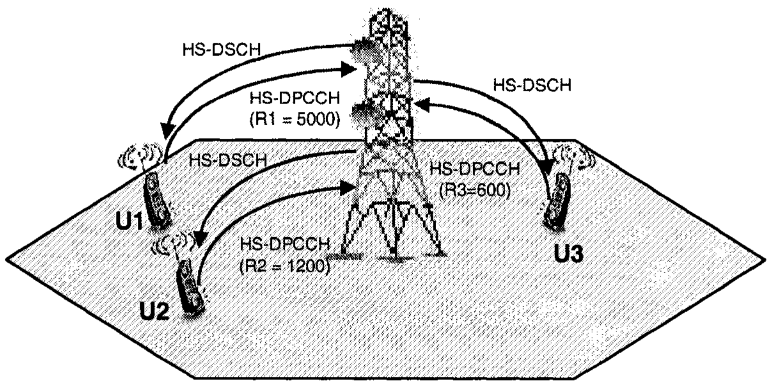
3.2.2.4.1 The Effect of Relative Fairness and Opportunity Cost

Suppose at time t there are three users connected to the Node B as shown in Figure 3.4: U1, U2 and U3. These users report their current data rates (R (t)) to the Node B through the HS-DPCCH as aforementioned. For example, the current data rate (R_1(t)) of U1 at time t is 5000 Kbps. This information is passed to the MAC-PS algorithm, which is located at the Medium Access Control-high speed (MAC-hs) sub layer at the Node B. Assume that the average throughputs for U1, U2 and U3 are 4000 Kbps, 3000 Kbps and 500 Kbps, respectively.

In the case of Max CIR, the user with the highest channel quality (i.e., highest data rate) is selected for transmission, which is in this case U1. With PF, the user with the highest relative channel quality ($R_i(t)/S_i(t)$), which is also in this case U1 (5000/ 4000) is scheduled. However, the MAC-PS uses the information about the average throughput and current data rate for each user to calculate his relative fairness. It then calculates the utility for each user i if selected for transmission and the utility for all other users $j \neq i$

given that i is scheduled according to Eq. 3.2. The MAC-PS then chooses the user with the highest aggregate utility. Assuming K = 5000 Kbps, the aggregate utilities of U1, U2 and U3 if each one of them is served are 3902.44, 3904.26 and 3917.76, respectively (see Tables A.1, A.2 and A.3 in Appendix A). Therefore, U3 is selected for transmission. This user is selected for transmission because his relative fairness ($\alpha_3(t)$) at time t is very low (500/4000=0.125) and, therefore, if he is not served, this results in a large decrease in his utility function (from – 1899.41 if served to -1939.36 if U1 is served or -1918.96 if U2 is served) causing the aggregate utility of the system to decrease. However, the opposite happens if this user is served, i.e., the sharp increase in his utility function causes the aggregate utility of the system to increase rapidly. As a result, the effect of relative fairness causes U3 to be scheduled despite his low current data rate (600 Kbps) compared to U1 and U2 (5000 and 1200 Kbps, respectively). Therefore, we can conclude that our algorithm is fairer than Max CIR and PF since the user with low relative fairness is scheduled whereas in Max CIR and PF, U1 is selected for transmission despite the fact U3 is experiencing low average throughput.

With K=5000, all users are considered for transmission since the opportunity cost of serving them is less than or equal to K. However, suppose we decrease K to 4000. Then only users 1 and 2 are considered for transmission and, therefore user 2 is scheduled since he is the one with the highest aggregate utility (U3 is not considered for transmission since the opportunity cost of serving his is larger than K (i.e., $(\max_j R_j(t)) - R_3(t) =$ (5000-600) > 4000)). If we set K to 1000, then only user 1 is considered for transmission since now the opportunity cost of serving users 2 and 3 is larger than 1000. This shows that by decreasing the values of K, fewer users are considered for transmission, namely, those with good channel conditions (i.e., high current data rates). Therefore, this gives the service provider the flexibility to choose the degree of fairness of the scheduler.



	R_i
U1	5000
U2	1200
U3	600

MAC-hs

MAC-PS

	S_i	R_i
U1	4000	5000
U2	3000	1200
U3	500	600

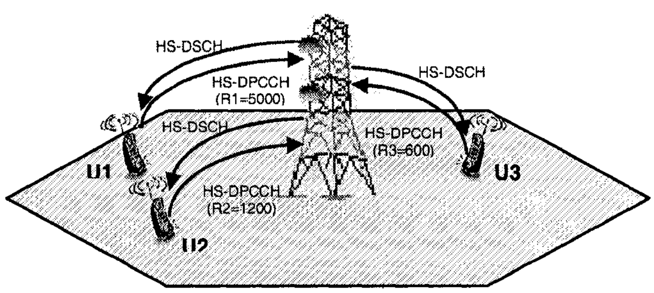
Aggregate utility of U1	3902.44
Aggregate utility of U2	3904.26
Aggregate utility of U3	3917.76

U3 is scheduled

Figure 3.4: The Effect of Relative Fairness

3.2.2.4.2 The Effect of Current Data Rates

Suppose that we have the same settings as in Section 3.2.2.4.1. We saw that the MAC-PS schedules U3 for transmission. The selection of U3 will continue until his relative fairness improves (the value of his relative fairness at which the scheduler will stop choosing his depends on the values of the scheduler parameters). Suppose that after another time, the scheduler has the same information as in the previous section except that the average throughput for U3 is now 1000 Kbps (see Figure 3.5). Assuming K=5000, then U3 is still scheduled for transmission (his aggregate utility = 4978.90) to bring his average throughput closer to the maximum one which is in this case the average throughput of U1 (see Tables A.4 and A.5 in Appendix A).



	R_i
U1	5000
U2	1200
U3	600

MAC-hs

MAC-PS

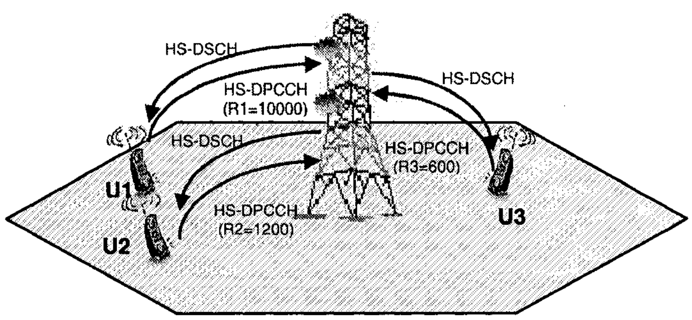
	S_i	R_i
U1	4000	5000
U2	3000	1200
U3	1000	600

Aggregate utility of U1	4976.58
Aggregate utility of U2	4976.22
Aggregate utility of U3	4978.90

U3 is scheduled

Figure 3.5: The Effects of Current Data Rates (R_1(t)= 5000 Kbps)

However, if the current data rate of U1 is changed from 5000 to 10000 Kbps as shown in Figure 3.6, then U1 now has the highest aggregate utility (9939.40) and, therefore he is scheduled for transmission (see Tables A.6 and A.7 in Appendix A). This clearly shows the effect of the current data rates of the user on our MAC-PS. Our proposed algorithm considers both the current data rate of each user and each user's relative fairness. The higher the current data rate of a user, the higher his chance of being selected by our MAC-PS. In addition, the lower the relative fairness of the user, the higher his chance of being scheduled to increase his relative fairness and bring his average throughput closer to the maximum average throughput



	R_i
U1	10000
U2	1200
U3	600

MAC-hs

MAC-PS

	S_i	R_i
U1	4000	10000
U2	3000	1200
U3	1000	600

Aggregate utility of U1	9939.40
Aggregate utility of U2	9842.71
Aggregate utility of U3	9845.40

U1 is scheduled

Figure 3.6: The Effects of Current Data Rates ($R_1(t)$= 10000 Kbps)

3.2.2.5 Discussion

In Section 3.1, we have identified the main objectives that a successful packet scheduling algorithm for HSDPA should meet. In this section, we investigate the extent to which our MAC-PS algorithm satisfies the aforementioned objectives.

- **Efficiency:** As can be seen, the proposed algorithm takes into account the instantaneous channel conditions of users (through their current supportable data rate) and gives more chance to users with good channel conditions to be selected for transmission and, therefore, it should be efficient.
- **Fairness:** The proposed algorithm considers not only the instantaneous channel condition of the users, but also their average throughputs compared to the maximum average throughput and it tries to use both in the scheduling decision. Therefore, it is not just those with good channel conditions that have a better chance of getting served but also those with low relative fairness since if they are served, we get a rapid increase in $X_{i2}(t)$, which may maximize the objective function. As a result, those with low average throughputs have a better chance of getting served even though they might have bad channel conditions.
- **User satisfaction:** User satisfaction as perceived by the service provider is taken into account by using both the instantaneous channel condition and the user relative fairness. Exploiting the information about the channel condition results in achieving high user average throughputs and this is what users want. In addition, this algorithm prevents users with bad channel conditions from achieving low throughputs by taking into account their average throughputs relative to the maximum average throughput among all users (relative fairness) and, therefore,

does not suffer from the problem of serving only those users with good channel conditions while ignoring the rest.

- **Flexibility:** Introducing the concept of opportunity cost to our proposed algorithm gives it a high degree of flexibility. This gives the service provider the flexibility to choose the degree of fairness and, therefore, control the throughput-fairness tradeoff. In our algorithm, the opportunity cost function is defined as OC (i, t) = $((\max_j R_j(t)) - R_i(t))$. That is, we define the opportunity cost as the loss of throughput if the user with the maximum data rate is not served. Therefore, the smaller the value K, the higher the opportunity cost, the higher the system throughput and the lower the degree of fairness (because only those whose current supportable data rates are close (depending on K) to the maximum one are chosen for transmission). The service provider may configure the algorithm to set K based on some conditions such as the number of users or the system current average throughput exceeding a certain limit. It is totally up to the service provider to determine how fair the algorithm should be depending on the factors explained in Chapter 1. In addition, the service provider may choose the appropriate values for K (for each Node B) to correspond to a certain degree of fairness.

It is important to note that it is only possible to confirm whether our MAC-PS algorithm meets these objectives by testing its actual performance. The performance evaluation of our MAC-PS algorithm is the subject of the next chapter.

3.3 Special Cases

In this section, we prove that our algorithm converges to the Max CIR or the PF algorithms as special cases by setting its parameters to some specific values.

***Lemma 1*:** The MAC-PS can be equivalent to the Max CIR algorithm.

Proof:

If we set K=0, then the MAC-PS will find user i for which OC $(i$, t)= (max $_j$ R $_j$ (t)) - R $_i$ (t), $\leq$ 0. The only user that satisfies this is the user with the maximum current supportable data rate (i.e., max $_j$ R $_j$ (t)). Therefore, the MAC-PS will choose the user with the highest current supportable data rate (i.e. best channel quality) which is equivalent to Max CIR.

***Lemma 2*:** The MAC-PS can be equivalent to the Proportional Fairness algorithm

Proof:

If we set $a = 0$, $b = 0$, c =0, d=1, λ =1, $\gamma = 0$, and we ignore the opportunity cost (i.e. set K to the highest data rate that the system can support (10 Mbps) so that all users are considered for transmission). Then the utility function in the MAC-PS becomes ln($S_i(t)$)+1 but since 1 is common to every user, we can then take it off. Therefore, the MAC-PS will find user i such that:

$$\text{Maximize} \sum_{i=1}^{n} \ln(S_i(t))$$

$$\text{Subject to} \sum_{i=1}^{n} S_i(t) \leq \text{ChC}$$

$$\forall_i, 1 \leq i \leq \text{n}$$

Maximizing the aggregate utility of the system is equivalent to maximizing the objective function F, where F is a function of $\vec{S}(t)$, and $\vec{S}(t)$ is a vector of the users' average throughputs at time t. That is, if we find a vector $\vec{S}(t)$ that maximizes F, then the aggregate utility function will also be maximized. In other words, we want to choose some values for the users average throughputs, such that they maximize F, and hence they maximize the aggregate utility of the system. In our case, choosing such values is done by choosing a user for transmission that results in the vector $\vec{S}(t)$ maximizing the objective function F. Therefore the problem can be formulated as follows:

$$\text{Maximize F}(\vec{S}(t)) \equiv \sum_{i=1}^{n} \ln(S_i(t))$$

$$\text{Subject to} \sum_{i=1}^{n} S_i(t) \leq \text{ChC}$$

$$\forall_i, 1 \leq i \leq \text{n} \tag{3.4}$$

where S_i (t+1) is the average throughput for user i at slot t+1 and can be calculated by using an exponentially smoothed filter as follows [14]:

$$S_i(t+1) = \begin{cases} (1-1/t_c)S_i(t) + 1/t_c \cdot R_i(t) & \textit{if user i is served} \\ (1-1/t_c)S_i(t) & \text{otherwise} \end{cases} \tag{3.5}$$

where t_c is the time constant of the filter and $R_i(t)$ is the current supportable data rate of user i. It should be noted that according to this equation, if the user is not served at time t, then his current data rate is set to zero in the calculation of his average throughput. This makes sense since the user's average throughput depends on his current data rate and if he is not served at time t then his effective (actual) data rate is zero. However, as aforementioned, if the user is not selected for transmission, then his current data rate in his utility function (Eq 3.1, 3.2, and 3.3) is not set to zero, since the current data rate depends on his channel condition not on his being served or not. In other words, if a user is not served at time t, then his current data rate is set to zero in the calculation of his average throughput but not in the calculation of his utility function (i.e. $X_{i2}(t)$ is only affected whereas $X_{i1}(t)$ stays intact).

Since ln ($S_i(t)$) is strictly concave and is differentiable then so is the objective function F. Also since the feasible region is bounded, then an optimal solution exits. Furthermore, the solution is unique and we can use a gradient ascent method to find it. However, a global optimal solution cannot be found since the number of users and the channel capacity are varying with time. Nevertheless, we can look for a locally optimal solution. That is, at each time slot, schedule the user that would result in a movement towards the optimal solution (i.e. a movement along the maximum objective function F gradient direction).

Let:

$F_i'(\vec{S}(t))$= gradient of the objective function in the direction of serving user i.

We would like to find the value of i with the largest gradient and moving to the maximal point along that direction. Since we know what the user's average throughput would be if served or not, then the optimization problem can be reduced to finding the maximum gradient in the direction of serving user i (i.e. maximize $F_i'(\vec{\mathrm{S}}\ (\mathrm{t}))$). We first find the gradient in the direction of serving user i. We can do this by parameterizing the movement along the ray in the direction of serving user i by μ, and then F_i can be written as a function of μ as follows:

$$F_i(\mu) = \sum_{i=1}^{n} \ln\left(S_i(t) + \mu(S_i(t+1) - S_i(t))\right) \tag{3.6}$$

Taking the derivative with respect to μ and evaluating it at $\mu = 0$ (to find the critical point, in this case maxima), we get

$$F_i'(\mu) = \sum_{i=1}^{n} \frac{1}{((S_i(t) + \mu S_i(t+1) - S_i(t)))} \cdot ((S_i(t+1) - S_i(t))) \qquad \text{(by the chain rule)}$$

$$= \frac{1}{((S_i(t) + \mu S_i(t+1) - S_i(t)))} \cdot ((S_i(t+1) - S_i(t)))$$

$$+ \sum_{j=1, j\neq i}^{n} \frac{1}{\left((S_j(t) + \mu S_j(t+1) - S_j(t))\right)} \cdot \left((S_j(t+1) - S_j(t))\right)$$

$$\therefore F_i'(0) = \frac{1}{S_i(t)} \cdot ((S_i(t+1) - S_i(t))) + \sum_{j=1, j\neq i}^{n} \frac{1}{S_j(t)} \cdot \left((S_j(t+1) - S_j(t))\right)$$

$$= \frac{1}{S_i(t)} \cdot \left(\frac{R_i(t)}{t_c} - \frac{S_i(t)}{t_c}\right), \qquad (S_i(t) \text{ is updated by Eq. 3.5 (user is } i \text{ served))}$$

$$+ \sum_{j=1, j\neq i}^{n} \frac{1}{S_j(t)} \cdot \left(-\frac{S_j(t)}{t_c}\right), \qquad (S_j(t) \text{is updated by Eq. 3.5 (user } j \text{ is not served))}$$

$$= \frac{1}{S_i(t)} \cdot \left(\frac{R_i(t)}{t_c}\right) - \frac{1}{S_i(t)} \cdot \left(\frac{S_i(t)}{t_c}\right) - \sum_{j=1, j\neq i}^{n} \frac{1}{S_j(t)} \cdot \left(\frac{S_j(t)}{t_c}\right)$$

Therefore, the gradient in the direction of serving user i can be written as:

$$\frac{1}{S_i(t)} \cdot \left(\frac{R_i(t)}{t_c} \right) - \sum_{j=1}^{n} \frac{1}{S_j(t)} \cdot \left(\frac{S_j(t)}{t_c} \right) \quad (3.7)$$

Since the summation term is common for all users, we can take it out, and, therefore, the maximum gradient direction (i.e. the user that would result in a movement along the maximum gradient direction) is:

argmax $_i$ $(F_i'(\vec{S}(t)))$ = argmax $_i$ $\frac{R_i(t)}{S_i(t)}$, which is the PF algorithm (t_c is constant and common to all users and, therefore, can be taken off).

Table 3.1 summarizes the settings of the parameters in our proposed algorithm in order to obtain the Max CIR and PF algorithms.

Algorithm	Parameters
Max CIR	$K = 0$
PF	$a = 0,\ b = 0,\ c = 0,\ d = 1,\ \lambda = 1,\ \gamma = 0$

Table 3.1: Specific Parameter Values for Convergence to the Max CIR and PF algorithms

3.4 Summary

In this chapter we proposed a novel Medium Access Control Packet Scheduler (MAC-PS) for HSDPA. This scheduler adopts the Cobb-Douglas utility function to express user satisfactions as perceived by the service provider. We have also proposed the use of opportunity cost in the scheduling decision to express the cost of fairness in the overall system's throughput. The proposed scheduler not only considers the instantaneous channel conditions of the users but also their average throughputs compared to the maximum average throughput of all of them (relative fairness). Therefore, our proposed scheduler serves users with good channel conditions while keeping track of those with low average throughputs, hence giving them more priority for transmission. The proposed MAC-PS was designed to meet four objectives: efficiency, fairness, user satisfaction and flexibility. It is efficient because it exploits the variations of users' channel conditions by giving more chances to those with good channel conditions for transmission. Its fairness comes from the fact that it keeps track of the users' relative fairness, and increases the chance of those with bad relative fairness for being selected for transmission. User satisfaction is considered by using both the channel condition and relative fairness. The proposed scheduler is flexible because it uses the concept of opportunity cost, which allows the service provider to choose and control the degree of fairness of the scheduler and, therefore, control the throughput-fairness tradeoff. Finally, we have mathematically shown that our proposed scheduler converges to the Max CIR and PF algorithms as special cases.

Chapter 4

Performance Evaluation

In this chapter we evaluate the performance of the MAC-PS algorithm. Section 4.1 describes the simulation model and the performance metrics used for the evaluation. The performance of the MAC-PS algorithm under various traffic and mobility conditions is analyzed in Section 4.2. We compare the performance of our proposed algorithm to that of the Max CIR and PF algorithms. A summary of observations is given in Section 4.3.

4.1 Simulation Model and Performance Metrics

In this section, we fist explain the performance evaluation method which we have used to evaluate our MAC-PS. Then we describe the basic models of the designed simulation tool. Next we explain how the average throughput is estimated at each time slot (this is needed for the MAC-PS and PF algorithms). Finally, we list the performance metrics.

4.1.1 Performance Evaluation Method

The performance evaluation of the MAC-PS is based on a discrete event simulation using Network Simulator 2 (NS2) [29]. NS2 is an open source simulation tool that simulates many network topologies and protocols. However, NS2 by itself does not support UMTS and HSDPA. Therefore, an extension to NS2 has been used in the evaluation process. Enhanced UMTS Radio Access Network Extension for NS2 (EURANE) is an extension to NS2 that was developed by the European Commission 5th framework project to support UMTS and HSDPA [30]. EURANE has been used for the performance evaluation of the MAC-PS algorithm, as well as the Max CIR and PF algorithms. All the relevant simulation parameters are included in Appendix B.

4.1.2 Simulation Model

The simulation model consists of one cell where Node B is located at the center of the cell (Figure 4.1). The user is connected to Node B on the downlink by the HS-PDSCH that is shared among all users and by the HS-DPCCH on the uplink, which is dedicated for each user. The HS-DPCCH is used to send the current estimate of the channel condition of the user. Each user calculates the Signal-to-Noise Ratio (SNR), which is the strength of the signal received from the Node B relative to the background noise. The SNR is then mapped to a value between 0 and 30 to quantitatively represent the current status of the channel condition of the user. The mapping procedure is explained in the next section. Next, this value is used as an index in the tables provided by the HSDPA specifications to determine the data rate that the user could support given his channel current channel condition. Finally, the user sends the data rate that he can support to the Node B through the HS-DPCCH.

The Node B is connected to the RNC by a duplex link of 622 Mbps bandwidth and 15 ms delay. The RNC is connected to the Internet by a duplex link of 10 Mbps bandwidth and 15 ms delay. There is also an FTP server that is connected to the Internet by a duplex link of 10 Mbps and 35 ms delay.

Three different environments are used in this simulation. Two of them are recommended by the 3rd Generation Partnership Project (3GPP): Pedestrian A (Ped A) and Vehicle A (Veh A) environments [31]. Pedestrian A simulates pedestrians walking in a pico cell at 3 km/hr. The cell diameter for this environment is 450 m. Vehicle A simulates vehicles traveling at a speed of 60 km/hr in a micro cell of 1000 m in diameter. The channel model for these two environments is described in the next section.

The third environment is created to test the scheduling algorithms under an extreme condition. In this environment, the channel conditions of all users are fixed at different values. Users are grouped according to their channel conditions. The main reason for creating this environment is to show how the evaluated scheduling algorithms serve users with different channel conditions in order to test the degree of fairness. Such tests cannot be made with Ped A and Veh A because the channel conditions of the users are changing with time according to their channel models.

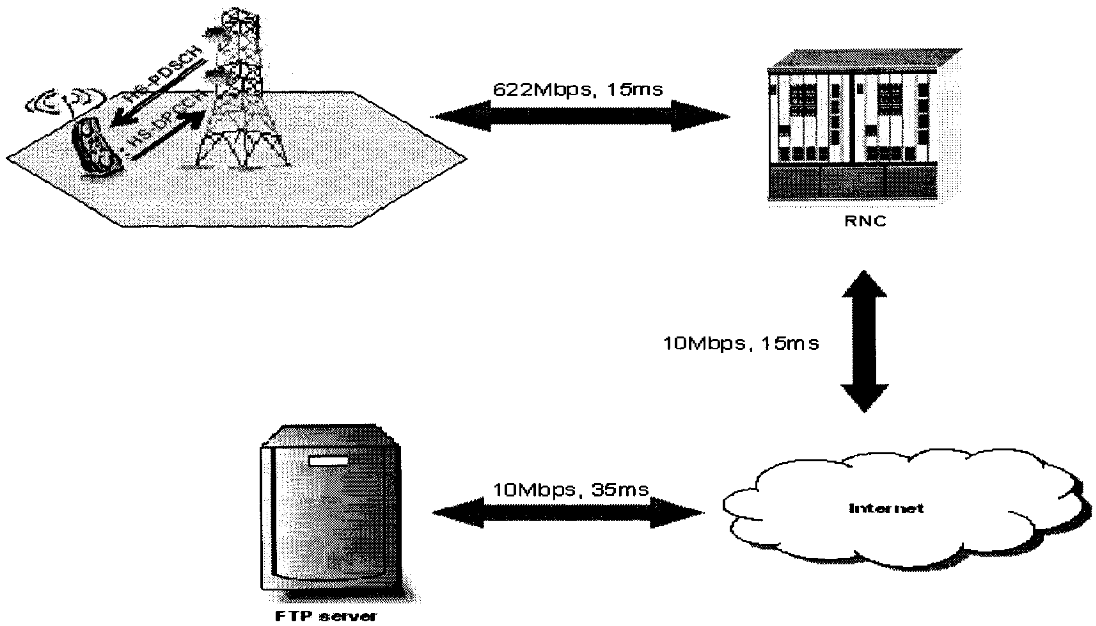


Figure 4.1: The Simulation Model

4.1.3 Propagation Model

The propagation model describes how much the radio signal attenuates on its way from the Node B to the user and, therefore, it describes how the channel condition of the user changes with time depending on the environment of the user and his mobility. In EURANE, the propagation model consists of five parts: path loss, shadowing, multi-path fading, intra-cell interference and inter-cell interference. Path loss is the attenuation in the signal due to the distance from the sender to the receiver. It increases as the distance increases. In addition, obstacles between the sender and the receiver cause shadow fading. The mobile is often shielded from the Node B but due to reflection and diffraction of the radio signal, it still can reach the Node B. Shadow fading causes slow variation of the signal received at the receiver. Moreover, as the signal travels from the sender to the receiver, it gets diffracted at different obstacles on its way, which results in several copies of the same signal coming from different paths to the receiver. Therefore, the received signal is the sum of those copies. This is called multi-path fading. The inter-cell interference is caused by the activities in the neighboring cell whereas the intra-cell interference is caused by the other users within the same cell.

In EURANE, each one of these parts is considered independent and is expressed in dB. The path loss is calculated as follows:

$$L(d) = 137.4 + 10 \cdot \beta \log_{10}(d)$$

where d is the distance from the UE to the Node B in Kilometers, β is the path loss exponent and is equal to 3.52. Shadowing is modeled through a lognormal distribution with a mean value of 0 dB and a correlation distance. The values of the standard deviation and the correlation distance depend on the environment of the user. The multi-path fading in EURANE corresponds to 3GPP channel models for Pedestrian A and Vehicle A environments [31]. The intra-cell interference and inter-cell interference are assumed to be constants and are set equal to 30 and -70 dBm respectively. Then at the user side, the Signal-to-Noise Ratio (SNR) is extracted from the received signal from the Node B to determine how strong the signal is according to the following formula:

$$SNR = P_{tx} - L_{Total} - 10 \cdot \log_{10}(10^{\frac{I_{intra} - L_{Total}}{10}} + 10^{\frac{I_{inter}}{10}})$$

$$= P_{tx} - 10 \cdot \log_{10}(10^{\frac{I_{intra}}{10}} + 10^{\frac{I_{inter} + L_{Total}}{10}})$$

where P_{tx} is the transmitted code power in dBm, L_{Total} is the sum of the path loss, shadowing, and multipath fading in dB, I_{intra} and I_{inter} are the inter and intra cell interference respectively in dBm.

The SNR is then mapped to Channel Quality Indicator (CQI) that is used to determine the rate at which the user can support from the Node B according to the 3GPP standard as follows [32]:

$$CQI = \begin{cases} 0 & SNR \leq -16 \\ \left\lfloor \dfrac{SNR}{1.02} + 16.62 \right\rfloor & -16 < SNR < 14 \\ 30 & 14 \leq SNR \end{cases}$$

After getting the CQI of each user, these values are used to determine the data rates that each user can support. The HSDPA specification comes with tables that determine the data rates for each combination of CQI and channel codes used. These tables are used in our simulation and can be found at [32]. The rates that the users can accept from the Node B will vary in time depending on their location, speed and environment. For example, the further the user is from the Node B, the larger the path loss and the lower the SNR, which results in lower data rates. Also the higher the speed of the mobile user is, the larger the effect of multipath fading and, therefore, the lower the SNR, which means lower data rates.

More information on the channel model and the calculation of the SNR can be found in [31].

4.1.4 Traffic Model

The traffic model consists of FTP sessions. Each user sends a request for one FTP file from the FTP server. The file size is 0.5 MB. The user's connection terminates after the download is complete. Users are modeled as a Poisson process with a mean value of one second. Users are uniformly distributed in the cell.

4.1.5 Throughput Averaging

The MAC-PS and PF algorithms require the average throughput of each user for their scheduling computations. In our simulation, the average throughput for user i is computed every TTI by using an exponential smoothing filter as follows [14]:

$$S_i(t) = \begin{cases} \left(1 - \frac{1}{t_c}\right) \cdot S_i(t-1) + \frac{1}{t_c} \cdot R_i(t) & \textit{if user i is served} \\ \left(1 - \frac{1}{t_c}\right) \cdot S_i(t-1) & \textit{otherwise} \end{cases}$$

where $S_i(t)$ is the average throughput approximation of user i at time slot t, $R_i(t)$ is the current supportable data rate of user i at time slot t, t_c is the time constant of the exponential smoothing filter and is set to 100 in most recent studies [22]. It should be noted that when the user does not have backlogged packets to transmit then the averaging procedure is not carried out to avoid the reduction of the estimated average throughput during the traffic inactivity periods.

4.1.6 Performance Metrics

The following performance metrics are used in this study:

- Average user throughput: calculated as the total number of successfully delivered bits divided by the duration of the user.
- Cell throughput: calculated as the average of users throughputs.

- Percentage of packet loss: number of dropped packets divided by the total number of packets.
- Fairness level: in terms of the distribution of the users average throughputs.
- User satisfaction level: a user is satisfied, if the user's average throughput is larger than or equal to a predefined value (e.g. 64 Kbps).

4.2 Simulation Results

In this section, we show simulation results of the evaluated algorithms for the three environments that we discussed earlier. The simulation results obtained in all experiments have a 95% confidence level with 5% confidence intervals (see Appendix C).

4.2.1 Pedestrian A (Ped A)

Figure 4.2 compares the cell throughput of our proposed algorithm, Max CIR, PF and Round Robin (RR) for the Pedestrian A environment with 25 users. The figure shows that the Max CIR algorithm achieves the highest cell throughput (2.1 Mbps). This is expected since the Max CIR algorithm only serves users at their best channel conditions at the expense of ignoring those with bad channel conditions. The lowest cell throughput is achieved by RR (0.9 Mbps), less than half the cell throughput achieved by the Max CIR algorithm. This is because RR does not make use of the channel conditions of users and gives all users the same number of time slots regardless of their instantaneous channel conditions. The cell throughput of the PF algorithm is between the Max CIR and RR (1.56 Mbps) because it tries to balance the users' average throughputs and their channel conditions.

The cell throughput achieved by our proposed algorithm (MAC-PS) with K= 7.3 Mbps is a little bit lower than the PF algorithm (1.4 Mbps). The reason for this is that our algorithm serves the users with low average throughputs even more than the PF algorithm by giving them more time slots in order to increase their relative fairness and maximize the overall utility of the system. However, as we decrease the K to 3 Mbps (according to our definition of the OC(i,t), the lower values of K, the lower the fairness), the cell throughput increases from 1.4 Mbps to 1.85 Mbps. This is because when K=3 Mbps only the users who if served, the opportunity cost of serving them is less than or equal 3 Mbps (i.e., their instantaneous channel conditions are good enough such that the opportunity cost of serving them does not exceed 3 Mbps).

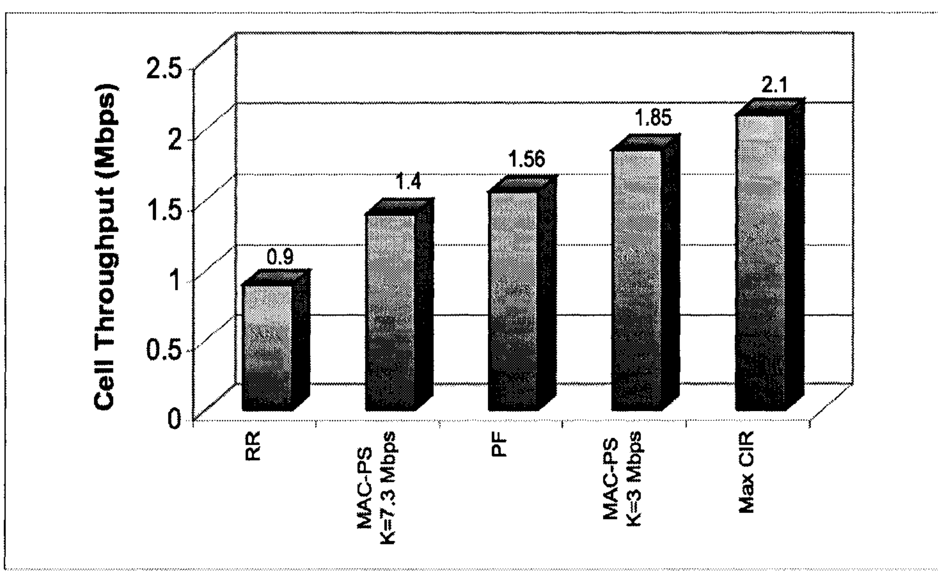


Figure 4.2: The Cell Throughput for Ped A (25 Users)

The effect of changing the values of K on the cell throughput is clearer in Figure 4.3. This figure shows the cell throughput of the proposed algorithm with K= 7.3, 4.5, 3, and

1 Mbps. Clearly, as K decreases, the cell throughput rapidly increases, since the users who are experiencing bad channel conditions are not selected for transmission for lower values of K. Even if a user with bad channel condition might maximize the utility function, the opportunity cost of serving them is larger than the minimum accepted values (i.e., 7.3, 4.5, 3, or 1 Mbps) and, therefore, they are not scheduled for transmission. The figure shows that when the values for K are 7.3, 4.5, 3,and 1 Mbps, the cell throughputs are 1.4, 1.62, 1.85, and 2.1 Mbps, respectively.

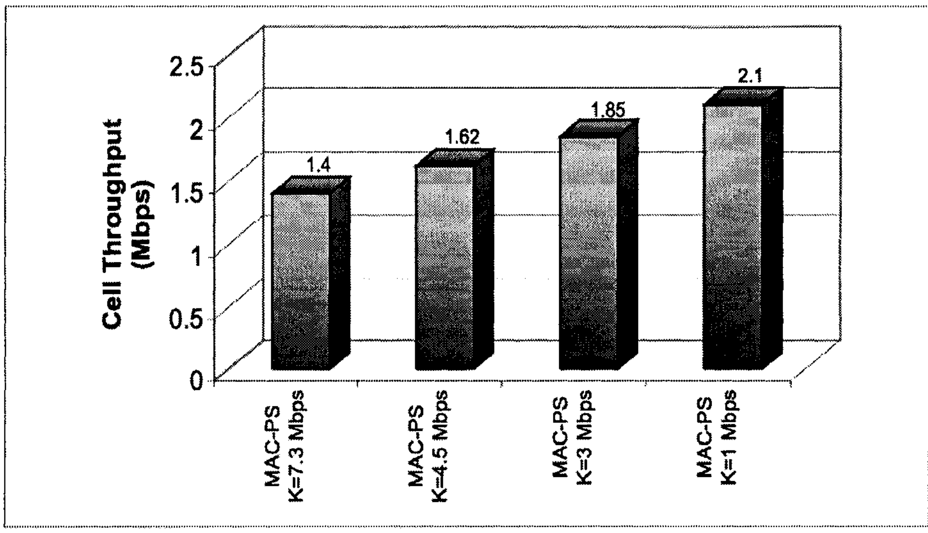


Figure 4.3: The Cell Throughput for Ped A (25 Users) for Different Values of K

Note here that when K =1 Mbps, then the performance of our algorithm is the same as the Max CIR. This holds for all the graphs that we show later. The reason for this is that the only users that are selected for transmission are the ones that have the best channel conditions (hence Max CIR).

Figure 4.4 shows the Cumulative Distribution Function (CDF) of the users average throughputs for the evaluated algorithms (for 25 users). Note that slope of the CDF of the users average throughputs determines the fairness of the scheduling algorithm because the steeper the slope is, the more compact the distribution of users average throughputs, which means that the algorithm is fairer. As shown in the figure, the CDF of RR has the steepest slope and the CDF of Max CIR has the flattest. Even though Max CIR has a very high throughput compared to RR, only few numbers of users enjoy that. Users who have good channel conditions get served more than those with bad channel conditions and therefore this results in an unfair distribution of throughput in Max CIR. RR gives an equal chance (in terms of time slots) for every user to transmit and, therefore, it is more fair in terms of the distribution of the user's average throughput. However, it is not completely fair because those with good channel conditions better utilize their resources by using higher order modulation and coding schemes and, therefore, can attain higher average throughputs. This explains why the CDF curve for RR is not completely vertical. The CDF curve of our proposed algorithm with K =7.3 Mbps has a steeper slope than those of the Max CIR and PF algorithms indicating that our algorithm is fairer. The effect of relative fairness in the definition of $X_{i2}(t)$ in our utility function keeps the values of the users average throughputs relatively closer to each other than the Max CIR and PF algorithms and results in a steeper CDF slope. The figure shows that when we decrease the value of K to 3 Mbps, the slope becomes flatter than the PF. This clearly shows the tradeoff between throughput and fairness. The effect of changing the values of K is further shown in Figure 4.5. Clearly as we decrease K (from 7.3 to 4.5,3, and 1 Mbps) the slope of the CDF of our algorithm becomes flatter confirming the tradeoff between throughput and fairness.

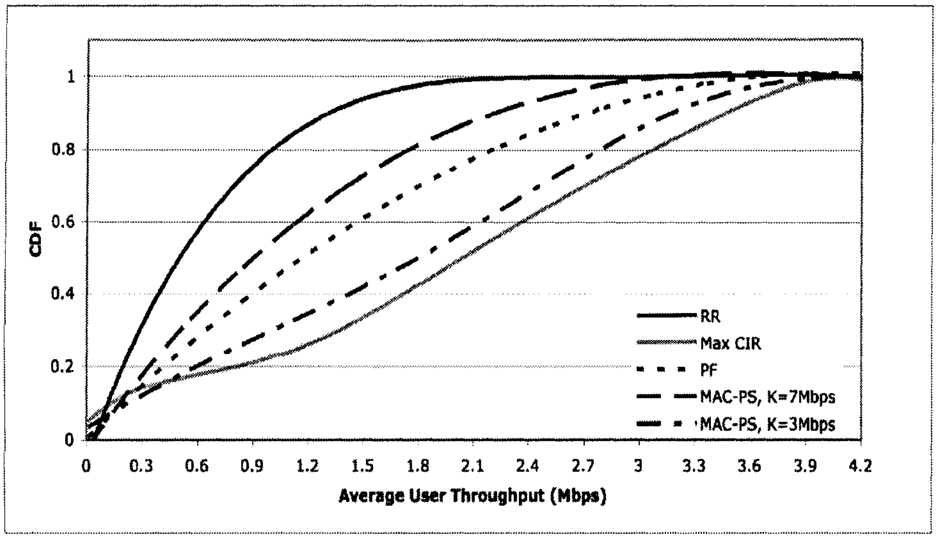


Figure 4.4: The Distribution of the Users' Average Throughputs for Ped A (25 Users)

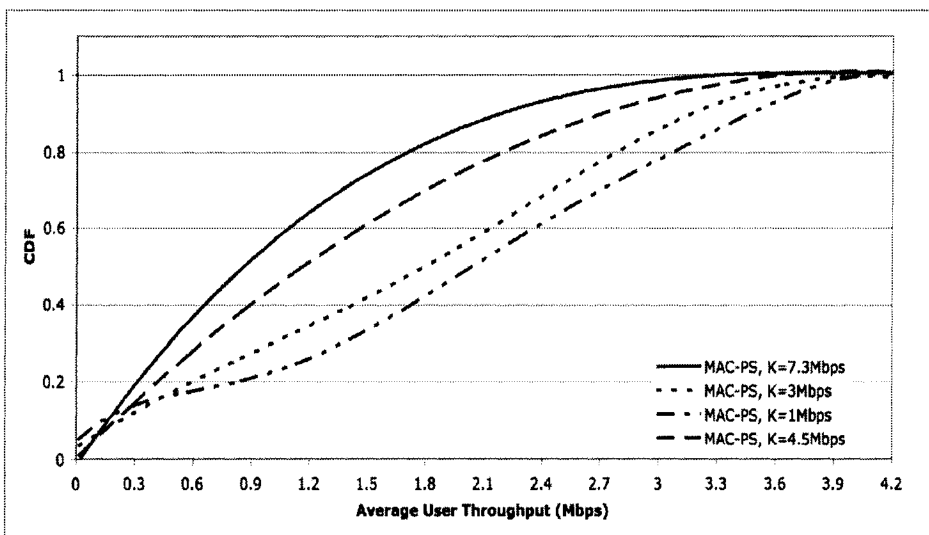


Figure 4.5: The Distribution of the Users' Average Throughputs for Ped A(25 Users) for Different K

Figure 4.6 shows the percentage of satisfied users with 128 Kbps (i.e., users are satisfied if their average throughputs are greater than or equal 128 Kbps). The MAC-PS with K=7.3 Mbps achieves the best performance in terms of user satisfaction. This is because of the use of the current supportable data rate and relative fairness together in the scheduling decision. Using the current supportable data rates of the users in the scheduling decision allows Node B to send at higher rates and thus increases the users average throughputs. In addition, the use of relative fairness prevents those with the best channel conditions from monopolizing the radio resources and gives more chances to the ones with bad channel conditions for transmission. The figure shows that the Max CIR algorithm achieves the worst user satisfaction (e.g., for 60 users, Max CIR achieves 66% while our algorithm achieves 78%).

The effect of different values of K is depicted in Figure 4.7. As we decrease K from 7.3 Mbps to 4.5, 3 and 1 Mbps, the percentage of satisfied users drops accordingly. As the opportunity cost is increased, a fewer number of users are served (i.e., those with good channel conditions) and, therefore, more users are left without being served; thus their levels of satisfaction drop. With K = 1 Mbps, the behavior of our algorithm is exactly the same as that of Max CIR, since only those with the best channel conditions are served.

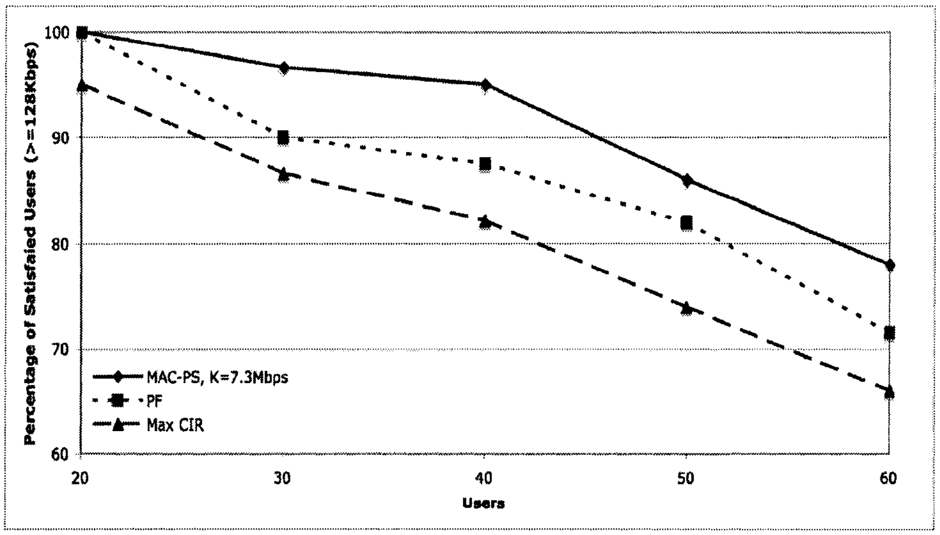


Figure 4.6: The Percentage of Satisfied Users (≥128Kbps)

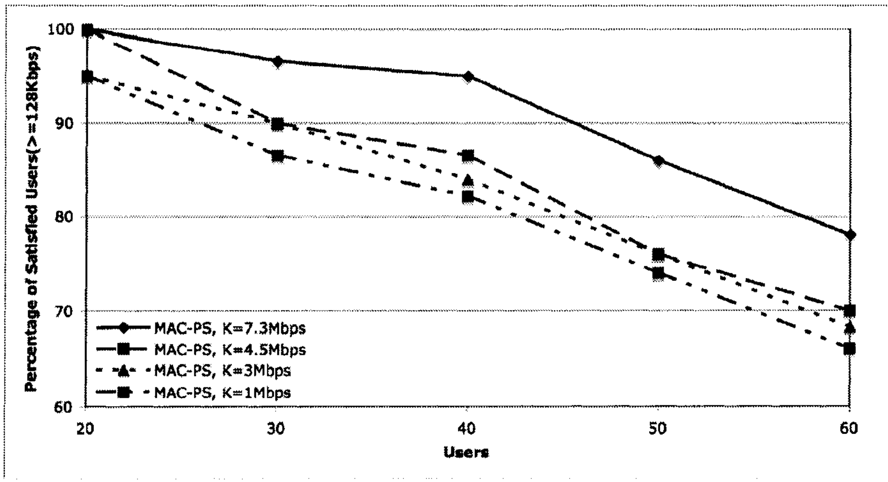


Figure 4.7: The Percentage of Satisfied Users (≥128Kbps) with Different K Values

Figures 4.8 and 4.9 show the percentage of satisfied users with 356 Kbps. The results look very similar to the case with 128 Kbps except that the percentages of satisfied users are lower. For example, the percentage of satisfied users in the Max CIR is 31% with 60 users compared to 66% with the same number of users in the case of 128Kbps. These two figures clearly show the superiority of our MAC-PS in terms of user satisfaction. In addition, they also show that increasing the system throughput by favoring those with good channel conditions comes at the expense of higher percentages of unsatisfied users.

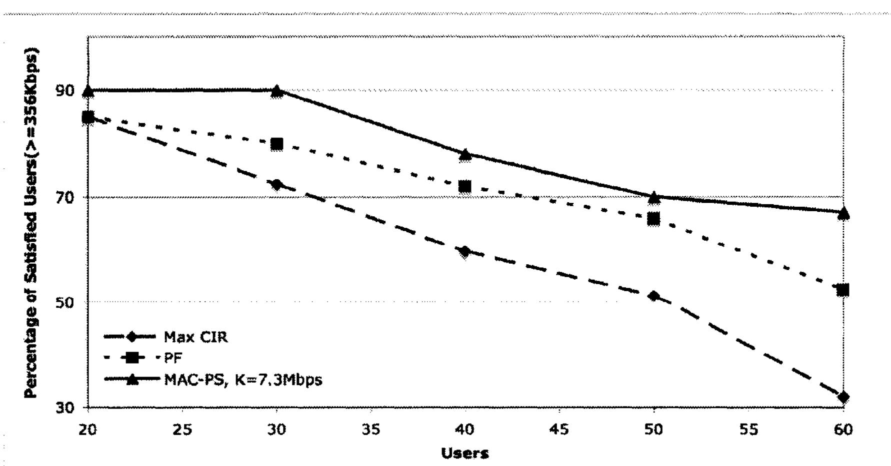


Figure 4.8: The Percentage of Satisfied Users (≥356Kbps)

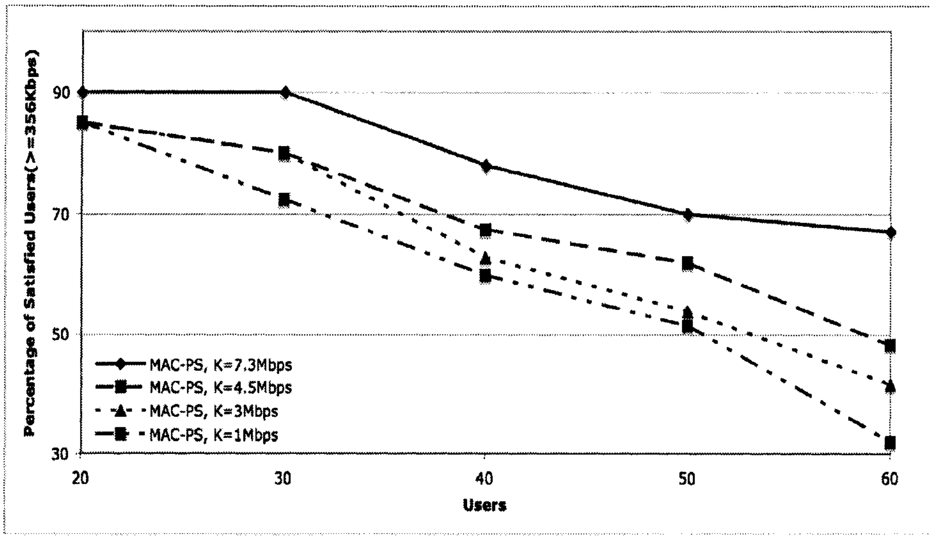


Figure 4.9: The Percentage of Satisfied Users (≥356Kbps) with Different K Values

4.2.2 Vehicle A (Veh A)

Figure 4.10 shows the cell throughput of the MAC-PS, Max CIR, PF and RR algorithms for the Vehicle A environment with 25 users. The figure shows similar behavior to that observed earlier in the Pedestrian A environment for the evaluated algorithms in terms of cell throughput. However, the overall cell throughput for all algorithms is lower than that in the Pedestrian A environment because the path loss and multi-path fading effects are higher in the Vehicle A environment, since the cell size and the mobile speed are larger than those of the Pedestrian A environment (see Appendix B). The cell throughput of our algorithm with K=7.3Mbps, K=3Mbps, RR, PF and Max CIR are 0.82, 1.18, 0.51,0.94 and 1.37 Mbps, respectively. The reason for this is already explained in the analysis of the Pedestrian A environment.

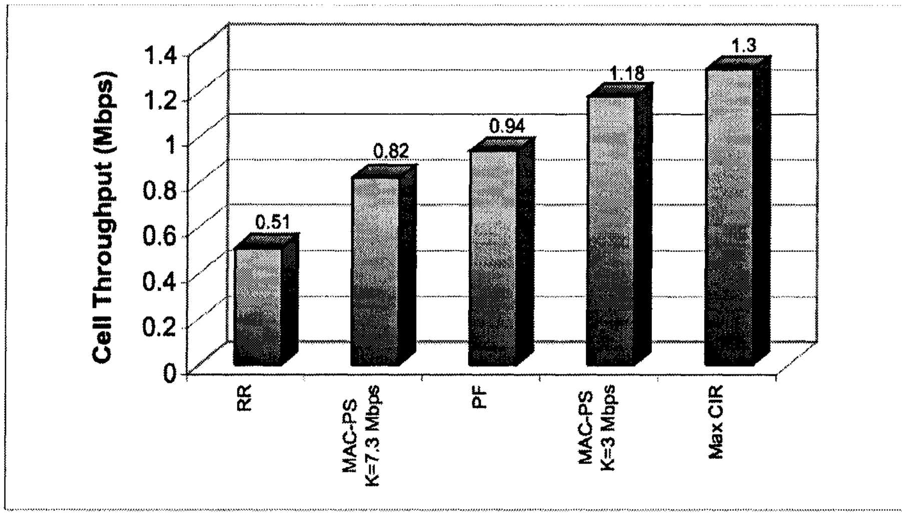


Figure 4.10: The Cell Throughput for Veh A (25 Users)

Figure 4.11 depicts the effect of different values of K. Note that when K =4.5 Mbps, the cell throughput of the MAC-PS is close to PF (1.06 Mbps). Also, as is the case with the Ped A environment, when K =1 Mbps, MAC-PS behaves exactly the same as Max CIR. Results clearly show that when the values of K are decreased, the cell throughput increases as only the users with favorable channel conditions are considered in the scheduling decisions.

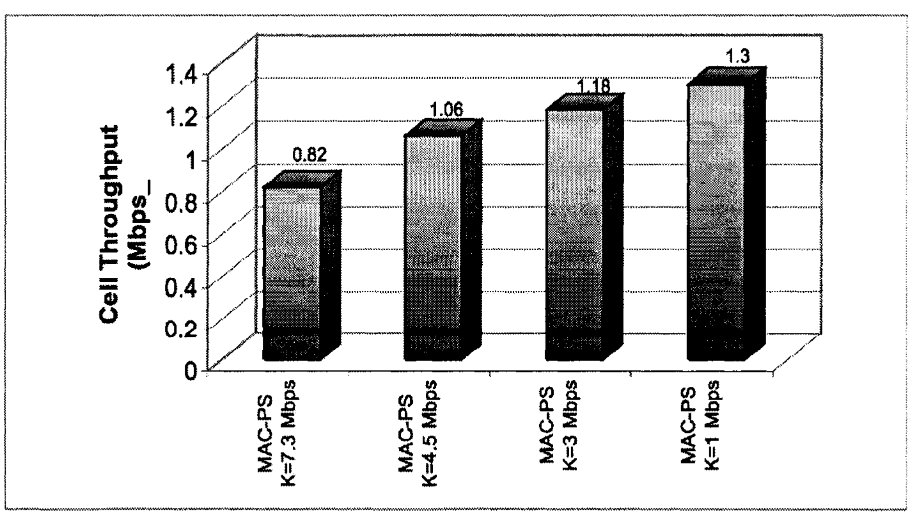


Figure 4.11: The Cell Throughput for Veh A (25 Users) for Different Values of K

Figure 4.12 shows the CDFs of the users' average throughputs for the evaluated algorithms. Following the same analysis as in the Pedestrian A environment, we can conclude that the RR algorithm outperforms our MAC-PS, Max CIR and PF in terms of the distribution of the user average throughput and hence it is more fair. The CDF of Max CIR has the flattest slope indicating that it is the worst algorithm in terms of fairness since the users' average throughputs are distributed over a larger range of values. The slope of the CDF for the PF algorithm is flatter than that of the RR algorithm and steeper than the slope of the CDF for the Max CIR algorithm. This shows that the PF algorithm outperforms the Max CIR algorithm in terms of fairness because it considers the user' average throughput in calculating the priorities of users. The slope of the CDF curve of the MAC-PS with K=7.3 Mbps is steeper than the PF algorithm because of the relative fairness effect as explained before. Therefore, our algorithm is fairer than the PF algorithm. However, as K is decreased to 3Mbps, the slope becomes flatter than that of the PF demonstrating once again the tradeoff between fairness and throughput. The effect of decreasing the values of K is shown in Figure 4.13. As K decreases, the cell throughput increases but at the expense of leaving the users with bad channel conditions unserved and, therefore, the slope of the CDF curve becomes flatter showing that the the degree of fairness of our algorithm is decreased.

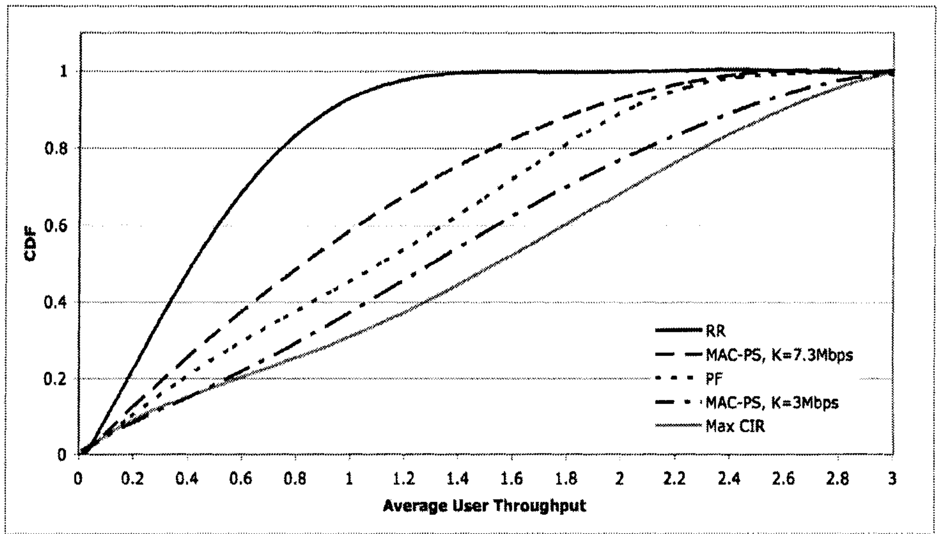


Figure 4.12: The Distribution of the Users' Average Throughputs for Veh A (25 Users)

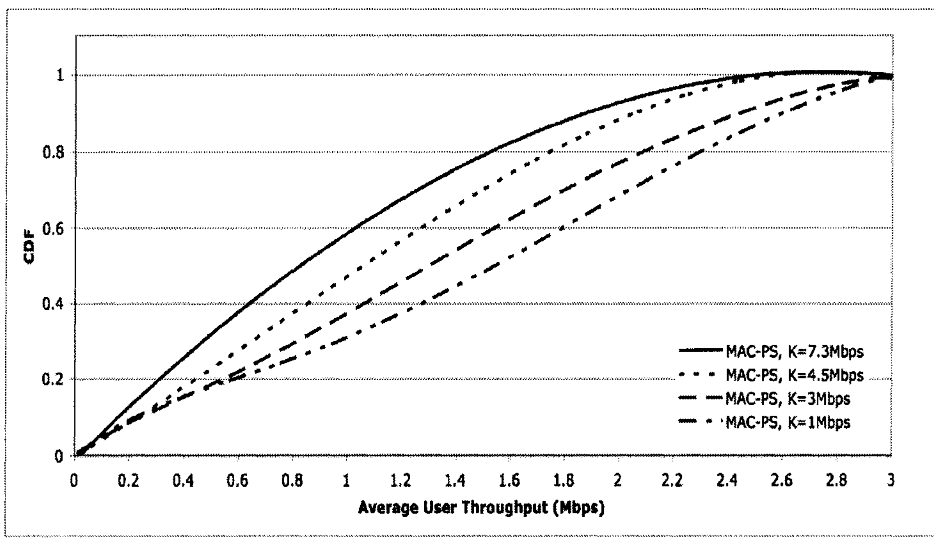


Figure 4.13: The Distribution of the Users' Average Throughput for Veh A(25 Users)for Different K

The percentage of satisfied users with 64 Kbps is depicted in Figure 4.14. With K=7.3 Mbps, the MAC-PS achieves the best performance in terms of user satisfaction. The worst performance is achieved by the Max CIR algorithm. The effect of relative fairness in our algorithm helps the users who are experiencing bad channel conditions and low average throughputs to get served and that is why its performance is better than the Max CIR and PF algorithms.

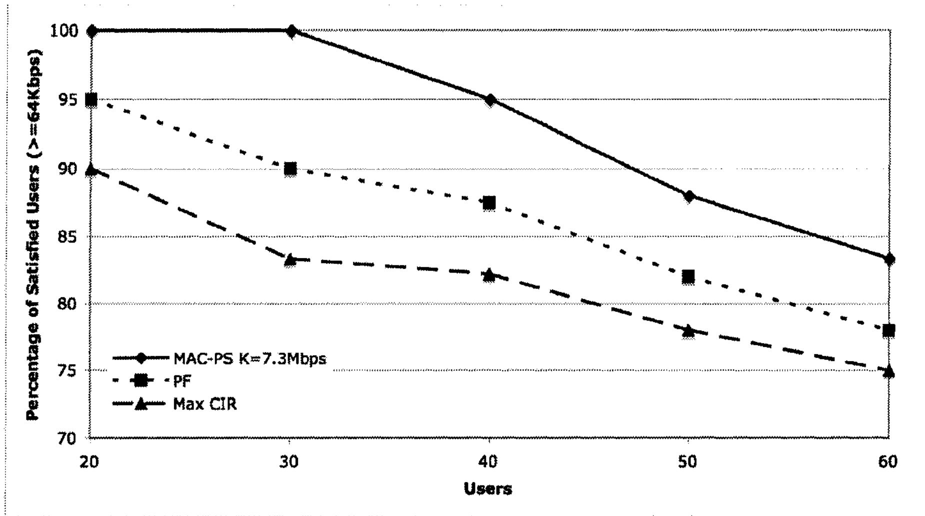


Figure 4.14: The Percentage of Satisfied Users (≥64Kbps)

Figure 4.15 shows the performance of our algorithm in terms of user satisfaction with 64 Kbps with different values of K. As K decreases, the percentage of satisfied users decreases indicating that fewer users are served, namely, those with good channel conditions.

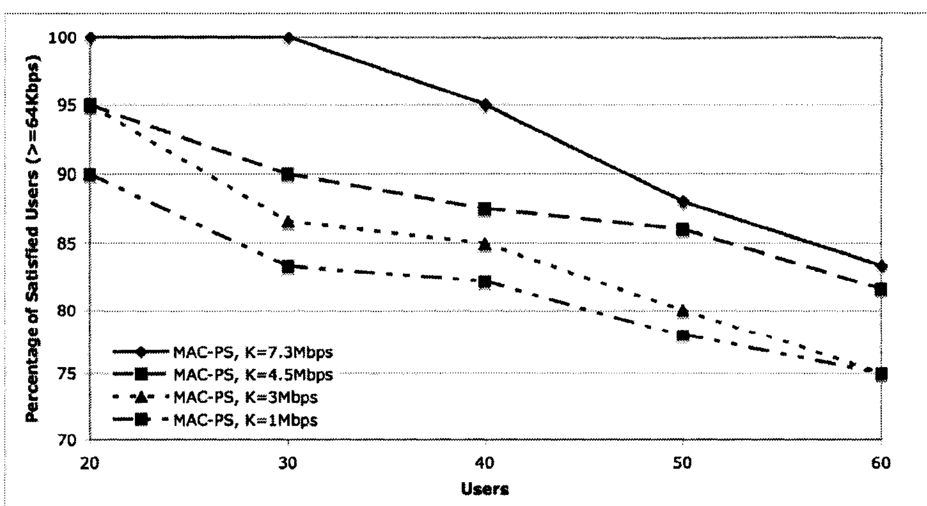


Figure 4.15: The Percentage of Satisfied Users (≥64Kbps) with Different K Values

Figures 4.16 and 4.17 show the performance of the evaluated scheduling algorithms in terms of user satisfaction with 128 Kbps. The performance of the algorithms look very similar to the case with 64 Kbps except that the percentages of satisfied users are less, since it is harder to achieve a minimum average throughput of 128 Kbps than 64 Kbps.

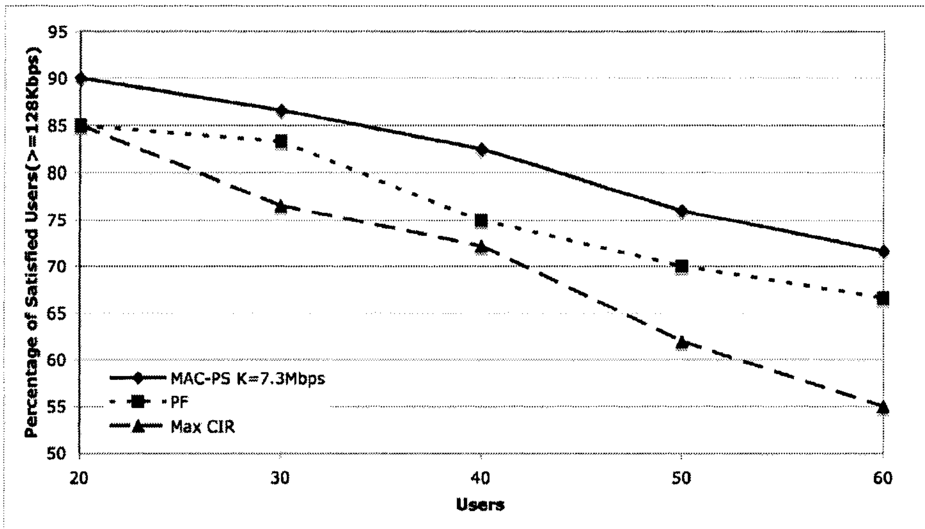


Figure 4.16: The Percentage of Satisfied Users (≥128Kbps)

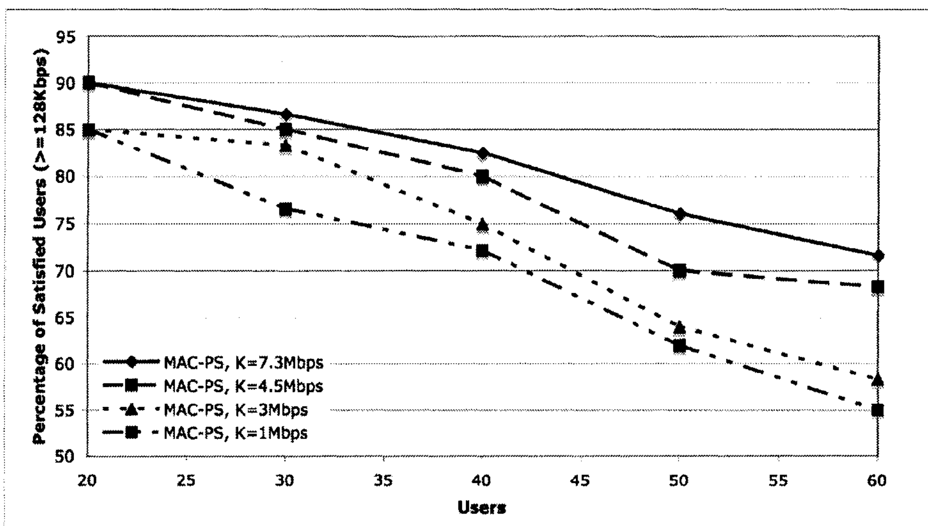


Figure 4.17: The Percentage of Satisfied Users (≥128bps) with Different K Values

4.2.3 Fixed Channel (FC)

The scheduling algorithms are evaluated in this environment based on average user throughput and percentage of packet loss. Seven values are used for the SNR (channel condition of the users): -7, -4, -1, 2, 5, 8, and 11 dB. For each SNR value, there are 10 users (a total of 70 users in the cell). Results for each group of 10 users based on their SNR are collected. For example, the average throughput is computed for users with SNR=-7 separately from users with SNR =-4, etc. This demonstrates how the scheduling algorithms serve users with different channel conditions.

Figure 4.18 shows the average user throughput for different SNR values. The unfairness of the Max CIR is obviously shown in this figure. The users with SNR =11 achieve the highest average throughput (2.03Mbps) whereas those with low SNR values achieve very little average throughput such as users with SNR=-4 (44 Kbps). As a matter of fact, users with SNR=-7 do not get served at all and get 0 average throughput. This is because the Max CIR is busy serving those with good channel conditions (e.g. SNR=11, and 8) and there is no chance to serve those with bad channel conditions (e.g. SNR=-7). The performance of the MAC-PS with K=7.3 Mbps and the PF algorithm in terms of average user throughput is modest for users with SNR=11, and 8 and superior to the Max CIR for users with SNR = 5, 2, -1, -4, and -7. Users with SNR=11 get 604 and 677 Kbps with the MAC-PS and the PF algorithm, respectively. The MAC-PS achieves higher user throughputs with SNR=-7, -4, and -1 (63,154, and188 Kbps) than the PF algorithm (45, 93, and 147 Kbps). However, the users' average throughputs for the PF algorithm (286, 390, 512, and 677 Kbps) are higher than the MAC-PS (243, 352, 444, and 604 Kbps) for users with SNR=2,5,8,and 11 respectively. This is because that the MAC-PS gives users with SNR =-7, -4, and -1 more time slots than the PF algorithm because of the relative fairness factor in the utility function. This indicates that our algorithm is fairer with users who are having bad channel conditions.

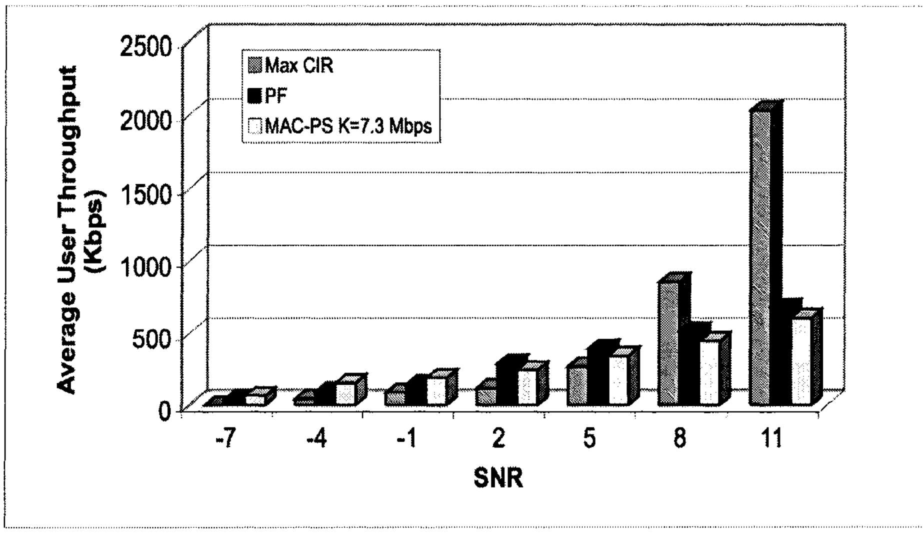


Figure 4.18: Average User Throughput of Users with Different SNR

The percentage of packet loss is depicted in Figure 4.19. This figure confirms the unfairness behavior of the Max CIR algorithm because while the users with good channel conditions achieve very low packet loss (e.g. 6.2% with SNR = 11), those users with bad channel conditions achieve very high packet loss (e.g. 100% with SNR = -7). The favoritism that the Max CIR provides for the users with the best channel conditions is at the expense of excluding those with less favorable channel conditions from being selected for transmission and, therefore, results in very low average throughput and high packet loss. The PF algorithm achieves higher packet loss for users with SNR = -7, -4 and -1 (66.2 %, 52.1% and 45.32%, respectively) than the MAC-PS algorithm (59.3 %, 46.2 % and 42.33%, respectively). The MAC-PS algorithms gives the users with bad channel conditions more time slots than the PF algorithm to increase their relative fairness and, hence, they achieve better packet loss. However, for users with SNR= 2, 5, 8 and 11, the performance of the PF algorithm in terms of the percentage of packet loss (34.1%, 30.9%, 26.7% and 23%, respectively) is slightly better than the MAC-PS algorithm (36.2%, 32.04%, 28.7% and 24.5%, respectively). The reason for this is because the MAC-PS gives fewer time slots to those users than the PF algorithm in order to serve those with bad channel conditions.

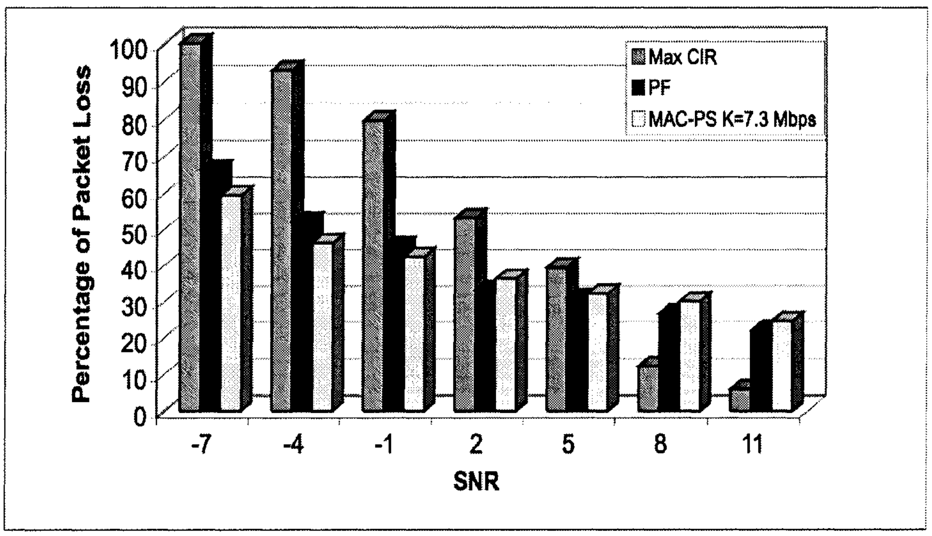


Figure 4.19: Percentage of Packet Loss of Users with Different SNR

4.3 Summary

Simulation results show that MAC-PS is fairer than the Max CIR and PF algorithms in terms of the distribution of the users' average throughputs. The relative fairness factor in the utility function of our MAC-PS prevents the users with less favorable channel conditions from being relegated to low average throughputs. This is because of the sharp increase in the utility function of those users if they are served; hence they end up maximizing the utility function despite their bad channel conditions. The simulation evaluation of the MAC-PS also shows that it achieves better packet loss for users with bad channel conditions than the Max CIR and PF algorithms. Moreover, the MAC-PS achieves a better user satisfaction level than the Max CIR and PF algorithms. In addition, simulation results clearly show that by decreasing the value of K (i.e., increasing the opportunity cost of fairness), the cell throughput increases at the expense of decreasing the degree of fairness, thus giving the service provider the flexibility to choose the degree of fairness and hence control the throughput-fairness tradeoff. In addition, simulation results confirm that our MAC-PS is superior to the Max CIR and PF algorithms in terms of fairness, user satisfaction and flexibility. Moreover, our MAC-PS is efficient, as it exploits the channel conditions of the users in the cell. Therefore, we can conclude that our MAC-PS is superior to other existing algorithms because of its unique features: efficiency, fairness, user satisfaction and flexibility.

Chapter 5

Conclusions and Future Work

In this thesis, we proposed a novel Medium Access Control Packet Scheduler (MAC-PS) for High Speed Downlink Packet Access (HSDPA). The scheduler expresses the users' satisfactions as perceived by the service provider by using a utility function. We have adopted the Cobb-Douglas utility function to express the users satisfactions in HSDPA. In our scheduler, two parameters affect the users' satisfactions: their current data rates and their relative fairness (i.e., the ratio of their average throughputs to the maximum average throughput achieved among them). The higher their data rates, the higher their satisfactions and also the fairer the system is, the more satisfied the users are. These two parameters try to balance the tradeoff between fairness and throughput. The higher the data rates for users, the higher their chances of getting selected by the scheduler for transmission. However, the scheduler prevents users with good channel conditions from monopolizing the radio resources because as the average throughputs of the users with bad channel conditions get very low compared to the other users, the scheduler selects them for transmission because of the sharp increase in their utilities so that they maximize the overall utility of the system.

In addition, the scheduler takes into account the service provider preferences by using a cost function that allows it to choose the degree of fairness of the scheduler and thus increase or decrease the system throughput according to its requirements. We have adopted the concept of opportunity cost from economics by letting the cost function be the opportunity cost of fairness. Fairness has an opportunity cost in terms of the amount of the data rates that the system has to give up in order to serve the users with bad channel conditions. By using the opportunity cost, we give the service provider the flexibility to choose the tolerated level of the loss in data rates so as to maximize its profits. Furthermore, we have mathematically shown that our proposed algorithm converges to the Max CIR and Proportional Fairness (PF) algorithms as special cases by setting some parameters to certain values, thus giving the service provider flexibility to choose between different scheduling strategies depending on its needs.

Simulation results reveal that our algorithm provides a good degree of fairness in terms of the distribution of the users average' throughputs compared to the Max CIR and PF algorithms. Its performance is also superior in terms of user satisfaction. Results also show that by increasing the opportunity cost, our algorithm tends to increase the cell throughput at the expense of degrading the level of fairness. This is expected, since with high opportunity cost of fairness, fewer users are selected for transmission, namely, those with the best channel conditions.

In the future, we would like to design a Call Admission Control (CAC) scheme that can be combined with the MAC-PS algorithm to be able to serve real time users and, hence, support multiple traffic types (real- and non-real-time traffic). We would like also to extend the opportunity cost concept to other aspects of wireless networking protocols. For example, we could extend the opportunity cost function to the CAC scheme that we intend to design because there is an opportunity cost for accepting new users to the system. This is because accepting new calls may degrade the Quality of Service (QoS) of ongoing calls in the system. Expressing this cost in the objective function will give the service provider the flexibility to accept as many users as it wishes to maximize its profits while keeping the OoS of ongoing calls above a certain level.

Bibliography

[1] H. Kaaranene, et al, "UMTS Networks, Architecture, Mobility, and Services," John Wiley & Sons, 1st edition, 2001.

[2] H. Holma and A. Toskala, "WCDMA for UMTS, Radio Access for Third Generation Mobile Communications," revised edition, John Wiley & Sons, 2001.

[3] 3GPP TS 25.308, "High Speed Downlink Packet Access (HSDPA); Overall Description", Release 5, March 2003.

[4] S. Borst, " User-level Performance of Channel-aware Scheduling Algorithms in Wireless Data Networks", *Proceedings of the IEEE INFOCOM, vol. 1, pp.321-331,* March 2003.

[5] A. Jalali, R. Padovani and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Date Rate Personal Communication Wireless System", *Proceedings of the IEEE Vehicular Technology Conference (VTC), pp. 1854-1858,* May 2000.

[6] C.J. Ong, H.J. Chong and R. Kwan, "Effects of Various Packet Scheduling Algorithms on the Performance of High Speed Downlink Shared Channel in a WCDA Network", *Proceedings of the IEEE Pacific Rim Conference on Communications, Computer and Signal Processing, vol. 2, pp. 935-938,* Aug 2003.

[7] W. Perndl, "Scheduling Algorithms for UMTS FDD Downlink", PH.D. Dissertation, Vienna University of Technology, July 2001.

[8] S. Pettersson, "A Comparison for Radio Resource Management Strategies in Bunched Systems for Indoor Communication", Available: http://www.s3.kth.se/radio/Publication/Pub1999/, August 2005.

[9] N. Nasser and H. Hassanein, "Radio Resource Management Algorithms in Wireless Cellular Networks", Handbook of Algorithms for Mobile and Wireless Networking and Computing, A, Boukerch (ed.), CRC press, in press.

[10] R. Krecher, J. Placht and T. Ruedebusch, "UMTS UTRAN Signaling", Tektronix, 2005.

[11] J. Bienaime, "Looking To The Future, 3G UMTS Services and Beyond", *Tehran Telecom Summit*, July 2004.

[12] T. Kolding, et al, "High Speed Downlink Packet Access: WCDMA Evolution", *IEEE Vehicular Technology Society News, vol. 50, pp. 4-10,* February 2003.

[13] S. Parvall, et al, "Evolving WCDMA for Improved High Speed Mobile Internet", *Proceedings of the Future Telecommunications Conference*, November 2001.

[14] P. Jose "Packet Scheduling And Quality of Service in HSDPA", PH.D. Dissertation, Aalborg University, October 2003.

[15] D. Chase, "Code Combining – A Maximum – Likelihood Decoding Approach for Combining an Arbitrary Number of Noise Packets", *IEEE Transactions on Communications, vol.33, pp. 441-448,* May 1985.

[16] Motorola, "Performance Comparison of Hybrid-ARQ Schemes", TSG-RAN WG#17, Oct 2004.

[17] X. Liu, E. Chong and N. Shroff, "Opportunistic Transmission Scheduling With Resource-Sharing Constraints in Wireless Networks", *IEEE Journal on Selected Areas in Communications, vol. 19, pp. 2053-2064,* Oct 2001.

[18] H. Kim, et al, "An Efficient Algorithm for QoS in Wireless Packet Data Transmission", *Proceedings of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'2002), vol. 5, pp. 2244–2248,* September 2002.

[19] J. M. Holtzman, "Asymptotic Analysis of Proportional Fair Algorithm", *Proceedings of the 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'2001), vol. 2, pp. F33–F37,* October 2001.

[20] G. Aniba and S. Aissa, "Fast Packet Scheduling Assuring Fairness And Quality Of Service in HSDPA", *Proceedings of the Canadian Conference on Electrical and Computer Engineering, vol. 4, pp. 2243-2246,* May 2004.

[21] G. Barriac and J. Holtzman, "Introducing Delay Sensitivity into the Proportional Fair Algorithm for CDMA Downlink Scheduling", *Proceedings of the IEEE Seventh International Symposium on Spread Spectrum Techniques And Applications vol. 3, pp. 652-656,* 2002.

[22] T. Bonald, "A Score-Based Opportunistic Scheduler for Fading Radio Channels", *Proceedings of the European Wireless*, 2004.

[23] Y. Ofuji, S. Sbeta and M. Sawahashi, "Fast Packet Scheduling Algorithm Based on Instantaneous SIR with Constraint Condition Assuring Minimum Throughput in Forward Link", *IEEE Wireless Communications and Networking, vol. 2, pp. 860-865,* March 2003.

[24] J. Peisa and E. Englund, "TCP Performance over HS-DSCH", *Proceedings of the Vehicular Technology Conference, VTC Spring, vol. 2, pp. 987-991,* 2002.

[25] J.H. Rhee, et al, "A Wireless Fair Scheduling Algorithm for 1xEV-DO System", *Proceedings of the Vehicular Technology Conference, VTC Fall, 2001, pp. 743-746,* 2001.

[26] H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks", *IEEE/ACM Transactions on Networking, vol 7, No 4, pp 473-489,* August 1999.

[27] M. Andrews, et al," Providing Quality of Service over a Shared Wireless Link", *IEEE Communications Magazine, vol. 39, pp. 150-154,* Feb 2001.

[28] H. Varian, "Intermediate Microeconomics: A Modern Approach", 6th edition, W. W. Norton & Company, 2003.

[29] Network Simulator 2, Available: http://www.isi.edu/nsnam/ns/, August 2005.

[30] Enhanced UMTS Radio Access Network Extensions for NS2, Available: http://www.ti- wmc.nl/eurane/, August 2005.

[31] Deliverable D3. 2v2, "End-to-end Network Model for Enhanced UMTS", Available: http://www.ti-wmc.nl/eurane/, August 2005.

[32] 3GPP TS25.214, "Physical Layer Procedures", Release 5, version 5.5.0, June 2003.

Appendix A

Numerical Examples of MAC-PS

In this Appendix, we show by numerical examples how the MAC-PS computes the aggregate utility of each user. The MAC-PS keeps track of the average throughput of each user up to time t and updates it after every scheduling decision by using an exponential smoothing filter (see chapters 3 and 4) as follows:

$$S_i(t+1) = \begin{cases} (1-1/t_c)S_i(t) + 1/t_c \cdot R_i(t) & \textit{if user i is served} \\ (1-1/t_c)S_i(t) & \text{otherwise} \end{cases} \tag{A.1}$$

where $S_i(t+1)$ is the average throughput for user i at slot t+1, t_c is the time constant of the filter and is set to 100 [22], and $R_i(t)$ is the current data rate of user i at time t. Given the parameter settings in Table A.1 and the information about the users' average throughputs and their current data rate in Table A.2 at time t, the MAC-PS algorithm calculates the following:

1- The maximum throughput achieved among all users up to time t ($\max_j S_j(t)$) = 4000.

2- The maximum current data rate of all users at time t ($\max_j$ R_j(t)) = 5000.

3- The relative fairness of each user i at time t ($\alpha_i(t)$=S_i(t) / ($\max_j$ S_j(t))^a). In this example, the relative fairness of U1, U2, and U3 are 1 (4000/4000), 0.75 (3000/4000), and 0.125 (500/4000), respectively.

4- The opportunity cost of serving each user (($\max_j R_j$(t)) − R_i(t)). For example if K=5000, then the Opportunity Cost (OC) of serving U1 = (5000-5000) $\leq$ K and therefore U1 is considered for transmission. If K =1000, then the OC of serving U3 = (5000-600) > K and therefore U3 is not considered for transmission.

5- The utilities of each served and not served user. For example, if U1 is served then from Eq. A.1 his average throughput ($S_i(t+1)$) will be: (1-1/100) 4000+ 5000/100= 4010. This user still has the maximum average throughput and therefore his relative fairness is still 1. Then his utility if served = $(5000)^1$ (1) =5000 (see Eq. 3.1). However, if not served, then his average throughput will be: (1-1/100) 4000 = 3960 and, therefore, his relative fairness will be 3960/4010 = 0.987. Hence the utility of U1 if he is not served = $(5000)^{1}\,((3\cdot\ln(0.987)+\frac{1}{(1.2)^{(1-0.987)}}+2.5(1-0.987))^1$=4956.29. The utilities of U2 and U3 are calculated in the same way.

6- The aggregate utility of each user if served. For example, if U1 is served, then his utility will be 5000 and the utilities of U2 and U3 will be 841.80 and -

3878.72, respectively (in this case U2 and U3 are not served). Then the aggregate utility in case U1 is served = 5000+841.80 -1939.36=3902.44.

7- The user with the highest aggregate utility is served provided that the opportunity cost of serving this user is not larger than K. As a result, U3 will be served in this example. Detailed computations are listed in Table A.3.

It should be noted that some of these steps can be combined. For example, steps 3 and 4 can be combined, i.e., the opportunity cost of each user can be obtained at the same time his relative fairness is computed.

A.1 The Effect of Relative Fairness and Opportunity Cost (OC)

Tables A.1, A.2, and A.3 show the parameter settings for the MAC-PS algorithm illustrated in Section 3.2.2.4.1, the information available to the MAC-PS algorithm at time t and the results obtained by our MAC-PS using the above mentioned steps.

It should be noted that the utility of U1 is the same whether U2 or U3 is served since this user has the maximum average throughput. However, this is not the case with U2 and U3. For example, if U3 is not served and U1 is served, then the utility of U3 is -1939.36. However, if U3 is not served and instead U2 is served then the utility of U3 is -1918.96. This is because when U1 (who has the maximum throughput) is served, then his average throughput increases and so does the maximum average throughput and hence the relative fairness of U3 further decreases causing his utility function to decrease. However, when U2 is served instead, then not only the average throughput of U3 decreases, but also the average throughput of U1 and hence the maximum average throughput decreases causing the relative fairness of U3 to increase. Therefore, the utility of U3 if U2 is served, is larger than his utility if U1 served. Even though U3 is not served, if U2 is served, then his utility increases because now his average throughput is closer to the maximum one since the maximum average throughput also decreases.

Users / Info	U1	U2	U3
$S_i(t)$	4000 Kbps	3000 Kbps	500 Kbps
$R_i(t)$	5000 Kbps	1200 Kbps	600 Kbps

Table A.1: Information Available for the MAC-PS at time t

Parameter	**Value**
a	1
b	1
c	1
d	1
λ	3
σ	1.2
γ	2.5

Table A.2: The MAC-PS Parameter Settings

Parameter	Value
$\max_j S_j(t)$	4000 Kbps
$\max_j R_j(t)$	5000 Kbps
$\alpha_1(t)$	1
$\alpha_2(t)$	0.75
$\alpha_3(t)$	0.125
OC of serving U1	0
OC of serving U2	3800
OC of serving U3	4400
Utility of U1 if served	5000
Utility of U1 if not served and U2 is served	4956.29
Utility of U1 if not served and U3 is served	4956.29
Utility of U2 if served	866.93
Utility of U2 if not served and U1 is served	841.80
Utility of U2 if not served and U3 is served	860.87
Utility of U3 if served	-1899.41
Utility of U3 if not served and U1 is served	-1939.36
Utility of U3 if not served and U2 is served	-1918.96
Aggregate utility if U1 is served (Eq. 3.3)	3902.44
Aggregate utility if U2 is served (Eq. 3.3)	3904.26
Aggregate utility if U3 is served (Eq. 3.3)	3917.76

Table A.3: Computations Done by the MAC-PS

A.2 The Effect of Current Data Rates

Tables A.4, A.5 A.6 and A.7, show the information available to the MAC-PS algorithm at time t and the results obtained by our MAC-PS for the example in Section 3.2.2.4.2. Tables A.4 and A.6 contain the same information except that the current data rate of U1 is changed from 5000 Kbps in Table A.4 to 10000 Kbps in Table A.6 to show the effects of the current data rates on our MAC-PS. It is assumed that the same parameter settings are used for the MAC-PS as in Table A.2.

Users / Info	U1	U2	U3
$S_i(t)$	4000 Kbps	3000 Kbps	1900 Kbps
$R_i(t)$	5000 Kbps	1200 Kbps	600 Kbps

Table A.4: Information Available for the MAC-PS at some time

Parameter	**Value**
$\max_j S_j(t)$	4000 Kbps
$\max_j R_j(t)$	5000 Kbps
$\alpha_1(t)$	1
$\alpha_2(t)$	0.75
$\alpha_3(t)$	0.25
OC of serving U1	0
OC of serving U2	3800
OC of serving U3	4400
Utility of U1 if served	5000
Utility of U1 if not served and U2 is served	4965.29
Utility of U1 if not served and U3 is served	4965.29

Utility of U2 if served	866.93
Utility of U2 if not served and U1 is served	841.80
Utility of U2 if not served and U3 is served	860.87
Utility of U3 if served	-838.26
Utility of U3 if not served and U1 is served	-865.21
Utility of U3 if not served and U2 is served	-847.01
Aggregate utility if U1 is served (Eq. 3.3)	4976.58
Aggregate utility if U2 is served (Eq. 3.3)	4976.22
Aggregate utility if U3 is served (Eq. 3.3)	4978.90

Table A.5: Computations Done by the MAC-PS with Information in Table A.4

Users / Info	U1	U2	U3
$S_i(t)$	4000 Kbps	3000 Kbps	1000 Kbps
$R_i(t)$	10000 Kbps	1200 Kbps	600 Kbps

Table A.6: Information Available for the MAC-PS at Some Other time

Parameter	**Value**
$\max_j S_j(t)$	4000 Kbps
$\max_j R_j(t)$	10000 Kbps
$\alpha_1(t)$	1
$\alpha_2(t)$	0.75
$\alpha_3(t)$	0.25
OC of serving U1	0
OC of serving U2	8800
OC of serving U3	9400

Utility of U1 if served	10000
Utility of U1 if not served and U2 is served	9822.78
Utility of U1 if not served and U3 is served	9822.78
Utility of U2 if served	866.93
Utility of U2 if not served and U1 is served	822.65
Utility of U2 if not served and U3 is served	860.75
Utility of U3 if served	-838.26
Utility of U3 if not served and U1 is served	-883.25
Utility of U3 if not served and U2 is served	-847.01
Aggregate utility if U1 is served (Eq. 3.3)	9939.40
Aggregate utility if U2 is served (Eq. 3.3)	9842.71
Aggregate utility if U3 is served (Eq. 3.3)	9845.40

Table A.7: Computations Done by the MAC-PS with Information in Table A.6

Appendix B

Simulation Parameters

In this Appendix, we present the relevant simulation parameters used in the performance evaluation of our algorithm in Chapter 4.

B.1 The MAC-PS Algorithm Parameters

Table B.1 presents the values used for the parameters in our proposed algorithm.

Parameter	**Value**
a	1
b	1
c	1
d	1
λ	3
σ	1.2
γ	2.5

Table B.1: The MAC-PS Algorithm's Parameters

B.2 The Ped A and Veh A Environments Parameters

The parameters relevant to the Ped A and Veh A environments are shown in Table B.2. It should be noted that most of these parameters are relevant to the channel model that is used in each of these environments and therefore they are not used in the Fixed Channel environment since the channel is fixed at different SNR values.

Parameter	Value
Mobile speed for Pedestrian A	3 Km/hr
Mobile speed for Vehicle A	60 Km/hr
Cell diameter for Pedestrian A	450 m
Cell diameter for Vehicle A	1000 m
Base station transmission power	38 dBm
Path loss exponent	3.52
Shadowing	Log normal distribution with mean 0
Base station antenna gain	17 dBi
Base station height	30 m
Distance loss at 1 Km	1.374e02
Intra cell interference	30 dBm
Inter cell interference	-70 dBm

Table B.2: The Ped A and Veh A Environment Parameters

B.3 Other Parameters

The system parameters are shown in Table B.3.

Parameter	Value
Simulation Time	100s
Number of cells	1
Traffic Type	FTP
Frame period	2 ms
Time slot per frame	3
File size	0.5 Mb
Number of requests per user	1
HS-DSCH multi codes	10
Min CQI value	0
Max CQI value	30
Exponential smooth Filter constant (t_c)	100
HSDPA coverage	full
Node B buffer size	30 Mb
Call arrival	Poisson process with mean 1 sec
Spatial distribution of the users in the cell	Uniformly distributed

Table B.3: System Parameters

Appendix C

Confidence Interval

The accuracy of simulation results is normally described in terms of confidence intervals placed on the mean values of these results. The procedure to calculate the confidence intervals is described in this appendix.

Let $X_1, X_2, \ldots, X_N$ be the simulation results of the same experiment resulting from N different runs. Furthermore, let us assume that these results are statistically independent. The sample mean, $\overline{X}$, of these results is given by:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \text{(C.1)}$$

and the sample variance of the sample values, S_x^2, is defined as follows:

$$S_x^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1} \tag{C.2}$$

The standard deviation of the sample mean is given by:

$$\frac{S_x}{\sqrt{N}} \tag{C.3}$$

Under the assumption of independence and normality, the sample mean is distributed in accordance to the *t-distribution*. The upper and lower limits of the confidence interval regarding the simulation results are defined as follows:

$$\text{Lower Limit} = \overline{X} - \frac{S_x t_{\alpha/2,N-1}}{\sqrt{N}} \tag{C.4a}$$

$$\text{Upper Limit} = \overline{X} + \frac{S_x t_{\alpha/2,N-1}}{\sqrt{N}} \tag{C.4b}$$

where $t_{\alpha/2,N-1}$ is the upper $\alpha/2$ percentile of the *t-distribution* with *N-1* degrees of freedom. Its values are widely available in tabular form. The intervals thus obtained are referred to as the intervals with 100 (1-α) percent confidence and *N-1* degrees of freedom. These confidence intervals can be made as small as desired by increasing the number of runs. As the value of N increases, the *t-distribution* approaches the normal distribution. For all practical purposes, when *N* is greater than 30 or so the values of $t_{\alpha/2,N-1}$ for the *t-distribution* become identical to those of the normal distribution.

In this thesis, the 95% confidence interval for each data point was obtained based on 10 independent runs per simulation experiment. The simulation running times have been chosen long enough to ensure stability and tight confidence intervals. In all simulation results presented in this thesis, the confidence intervals rarely exceeded 5% of their corresponding mean values.