DELAY-BASED PACKET SCHEDULING FOR HIGH SPEED DOWNLINK PACKET ACCESS

By

SAMREEN RIAZ HUSAIN

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
August, 2007

Abstract

High Speed Downlink Packet Access (HSDPA) is a cellular system that was standardized by the 3rd Generation Partnership Project (3GPP). HSDPA can support data rates of up to 14.4 Mbps through the use of a shared channel. Due to its high transmission rates, the highly popular multimedia applications are converging over this network. Moreover, as the shared channel is assigned to a single user in a given time interval, the scheduling decision is considered as a crucial one. Conventional HSDPA scheduling schemes utilize the fluctuations in channel condition to maximize system throughput by selecting users with relatively good radio conditions. However, this raises the issue of fairness as users with relatively poor channel conditions might not be served and consequently may suffer from starvation. Furthermore, Real-Time (RT) applications have strict delay constraints and require that packets are transmitted within a certain delay threshold.

In this thesis, a Delay Based Scheduler (DBS) is proposed for HSDPA which aims at minimizing the average queuing delay at the packet scheduler without compromising system throughput and fairness. In addition, the scheme can balance the tradeoff between throughput maximization and the minimization of queuing delay through the attunement of a parameter, thus allowing the service provider to choose between these two metrics. The DBS maintains the delay constraints of RT applications by defining delay thresholds for each traffic class and dropping packets that exceed their delay limit. The DBS accommodates Quality of Service (QoS) prioritization by defining and utilizing desired QoS parameters in the scheduling assignment. Finally, it was mathematically shown that the DBS can converge to a Non-Real-Time (NRT) scheme known as the Max CIR algorithm, allowing the scheduler to support RT and NRT applications simultaneously. The performance of the DBS was evaluated and compared to other well known schemes. It was found that the DBS can minimize the aggregate queuing delay of the system and maintain similar throughput and fairness.

Acknowledgements

I would like to take this opportunity to thank my supervisors, Professor Hossam Hassanein and Professor Nidal Nasser, for their patience, kindness and guidance without which the completion of this thesis would not have been possible. I am grateful to Professor Hassanein for giving me an opportunity to complete my Masters program under his supervision. His expertise, knowledge and enthusiasm have made a significant contribution to my graduate experience. The co-supervision of Professor Nasser, at the University of Guelph, has provided motivation and support that were a great source of inspiration during the course of my research. I appreciate his knowledge and experience in the HSDPA area and, moreover, his suggestions and comments for this thesis.

I would like to extend my gratitude to the members of my examination committee, Dr. Randy E. Ellis, Dr. Pat Martin and Dr. Ying Zou. Their thorough review of this manuscript, discussions and constructive comments were greatly appreciated.

A very special thanks goes out to Bader Al-Manthari who is presently pursuing a Doctoral Degree at Queen’s University. Bader helped me with the simulations and other technical support that was required for the completion of this research project.

I owe special recognition to my friend Vincent Mok for his encouragement and assistance. He has truly given me a helping hand on various occasions through my undergraduate and graduate studies and I deeply acknowledge him for that. I am also thankful to my friends and fellow classmates Shruti Atre and Andy Mao for making my graduate life, at Kingston, a memorable experience. Special thanks to my brother and friend, Salman, for his constant jokes and enthusiasm that were an invaluable source of content.

I would like to express my deepest gratitude to my parents Farha Begum and Sayed Riaz Husain, without whose motivation and encouragement I would not have considered a graduate program. Over the years, I have been humbled by my mother’s perseverance and amnesty. Under my father’s tutelage I learned that anything is possible with determination. It was through their persistence and hard work that I enrolled in an engineering undergraduate program at Queen’s University in Canada and, furthermore, they encouraged me to pursue a graduate degree. My parents are truly my ideals and mentors and have taught me the biggest lessons of life. I doubt that I will ever be able to convey my appreciation fully, but I owe them my eternal gratitude.

Finally, I recognize that this research would not have been possible without the financial assistance of Queen's University.

Contents

List of Tables

List of Figures

List of Acronyms

16QAM	Quadrature Amplitude Modulation
1G	1st Generation Cellular Systems
2G	2nd Generation Cellular Systems
3G	3rd Generation Cellular Systems
3GPP	3rd Generation Partnership Project
AMC	Adaptive Modulation and Coding
AMR	Adaptive Multi-Rate
APF	Adaptive Proportional Fairness
ARQ	Automatic Repeat Request
BCH	Broadcast Channel
BLER	Block Error Rate
CAMEL	Customized Application for Mobile Enhanced Logic
CCCH	Common Control Channel
CDMA	Code Division Multiple Access
CN	Core Network
CODEC	Speech Compression/Decompression
CPCH	Uplink Common Packet Channel
CQI	Channel Quality Indicator
CTCH	Common Traffic Channel

D-AMPS	Digital AMPS
DCCH	Dedicated Control Channel
DCH	Dedicated Channel
DPCH	Dedicated Physical Channel
DRC	Data Rate Control Exponent
DSCH	Downlink Shared Channel
DTCH	Dedicated Traffic Channel
EDGE	Enhanced Data Rate for GSM Evolution
ER	Exponential Rule
ETSI	European Telecommunications Standard Institute
FACH	Forward Access Channel
FTP	File Transfer Protocol
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HARQ	Hybrid Automatic Repeat Request
HSDPA	High Speed Downlink Packet Access
HS-DPCCH	High Speed Dedicated Control Channel
HS-DSCH	High Speed Downlink Shared Channel
HS-PDSCH	High Speed Physical Downlink Shared Channel
HS-SCCH	High Speed Shared Control Channel
MAC	Medium Access Control
Max CIR	Maximum Carrier-to-Interference Ratio
M-LWDF	Modified Largest Weighted Delay First
Node B	Base Station
NRT	Non-Real-Time
PCCH	Paging Control Channel
PCH	Paging Channel
PDC	Personal Digital Communication
PDU	Protocol Data Unit
PF	Proportional Fairness

QBER	Queue-Based Exponential Rule Scheduler
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RACH	Random Access Channel
RLC	Radio Link Control
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RRC	Radio Resource Control
RRM	Radio Resource Management
RT	Real-Time
SDU	Service Data Unit
SGSN	Serving GPRS Support Node
TBS	Transport Block size
TF	Transport Format
TFRC	Transport Format Resource Combination
TGS	Throughput Guarantee Scheduling
TTI	Transmission Time Interval
UE	User Equipment
UMTS	Universal Mobile Telecommunication Systems
USIM	UMTS Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice-over-IP
WFS	Weighted Fair Scheduling

Chapter 1

Introduction

The airways were first discovered as a communication medium in the late 19^{th} century. The first mobile telephone services that were introduced in the early 1980s supported analog voice technologies and were limited to a single cell environment with low bandwidth and poor quality. The advancement in the semiconductor technology led to the introduction of smaller and lighter weight equipment which increased the popularity of wireless communication systems. Moreover, the need to improve transmission quality, system capacity and coverage led to the development of the second generation (2G) cellular system known as Global System for Mobile communications (GSM) which was standardized by the European Telecommunications Standard Institute (ETSI). GSM is the most successful and widely used wireless communication system to date. The advent of the Internet increased the demand of data services over the wireless medium which in-turn encouraged the introduction of data transmissions in cellular systems. However, GSM has a data rate of up to 9.6 kbps but the highly popular multimedia applications require greater transmission rates. Consequently, this led to the development of Universal Mobile Telecommunication System (UMTS) which is labeled as a 3rd generation (3G) wireless cellular system. UMTS supports universal mobility through the use of global roaming standards and has a data transmission rate of 2 Mbps [1].

The escalating demand for multimedia applications will require data rates that are much beyond what 2G and 3G wireless cellular systems can offer. In order to provide such high transmission rates, the 3rd Generation Partnership Project (3GPP) standardized a 3.5G wireless communication system known as High Speed Downlink Packet Access (HSDPA) which is an extension to UMTS. HSDPA can support a data rate of up to 14.4 Mbps through the use of a shared channel known as the High Speed Downlink Shared Channel (HS-DSCH) [2]. Moreover, HSDPA aims at achieving lower delays and improving overall system capacity with the help of various new features such as: Adaptive Modulation and Coding (AMC), Hybrid Automatic Repeat Request (HARQ) and channel dependent scheduling. Furthermore, the HSDPA packet scheduler is implemented at the base station (known as Node B in UMTS and HSDPA systems) which converts the network from a central to a distributed scheduling system and consequently improves the overall system throughput and radio spectral efficiency.

HSDPA assigns the shared channel resource to a particular user at a given time interval depending upon the packet scheduling algorithm. Therefore, the scheduling assignment is a crucial decision that can improve the overall system performance. Moreover, mobile users have varying channel conditions due to factors such as mobility, interference etc.

Users with good channel conditions can transmit a greater number of bits per second when compared to users with poor channel conditions. HSDPA utilizes these variations in channel conditions for the scheduling decision where the packet scheduler can maximize the system throughput by selecting users with relatively good channel conditions. However, assigning the HS-DSCH to users with good channel conditions causes an unfair distribution of system resources as users with poor channel conditions might not get access to the shared channel. Therefore, a good scheduling scheme should not only improve system performance but also distribute system resources in a fair manner as users with the same cellular services expect similar performance regardless of their channel condition.

The rest of this chapter is organized as follows: Section 1.1 discusses the motivations of this work and the research objectives. The contributions made by the thesis are summarized in Section 1.2. Finally, the remainder of the thesis is outlined in Section 1.3.

1.1 Motivations and Objectives

Numerous scheduling algorithms have already been proposed for HSDPA. These algorithms can be categorized into two groups depending upon their delay requirements viz. Non-Real-Tine (NRT) and Real-Time (RT) scheduling schemes. NRT scheduling algorithms are employed to schedule traffic classes that are delay tolerable and hence these schemes aim at maximizing system throughput by utilizing the channel condition information of users for their scheduling priority. The two most well known NRT scheduling algorithms for HSDPA are: Maximum Carrier-to-Interference Ratio (Max CIR) [3] and Proportional Fairness (PF) [4] schemes. RT scheduling schemes such as Modified Largest Weighted Delay First (M-LWDF) [5] and Exponential Rule (ER) [6] are used to schedule delay sensitive traffic such as audio and video streaming applications. These applications have strict delay requirements and hence RT schemes consider the queuing delay of a user in the scheduling assignment.

As mentioned above, users with favorable channel conditions can support a higher bit rate which allows them to enjoy faster downloads. NRT schemes utilize this fact to maximize system throughput by scheduling users with relatively good channel conditions. However, this leads to an unfair distribution of system resources as it prevents users with poor channel conditions from being assigned to the shared channel. Therefore, users paying for the same service are alienated due to their channel conditions. As a result, users with unfavorable channel conditions may switch service providers due to poor service. Hence, it is not beneficial for the service provider to implement scheduling algorithms that cater users with good channel conditions as they contribute a fraction of the total revenue that can be generated. Therefore, a good scheduling algorithm should not only maximize system throughput but also assign system resources in a fair manner. Moreover, various services are available simultaneously over a cellular network and these services are differentiated based on their QoS requirements consequently allowing users to choose a service depending on their service needs. For example, users that transfer more than 100 Mb over the cellular network on a daily basis would choose a service that caters to users with high transfer requirements whereas occasional users would choose a more basic service. The service provider may prioritize users based on their Quality of Service (QoS) requirements; however, conventional NRT scheduling algorithms for HSDPA such as the Max CIR and PF schemes do not assign any QoS parameters that may distinguish users based on their requirements. Hence, QoS prioritization in a scheduling scheme is also important for generating higher revenues as users paying for a better service expect certain service guarantees.

Recently, RT applications have become exceptionally popular and services such as video telephony are used on a day-to-day basis. Moreover, the high supportable data rate of HSDPA has caused the convergence of these applications on to this wireless communication system. However, RT applications are highly delay sensitive and hence have strict delay constraints. For this reason, RT scheduling algorithms in HSDPA-based networks prioritize users based on the queuing delay experienced by them at the Node B. For example, the M-LWDF and ER schemes try to minimize the aggregate queuing delay of the system while considering the instantaneous channel conditions of the users in an attempt to balance the tradeoff between throughput maximization and the mitigation of queuing delay. This is because, selecting users that have high queuing delays but poor channel conditions will lower the overall system throughput whereas prioritizing users based on their channel condition over their queuing delay will increase the average queuing delay of the system. Furthermore, the M-LWDF and ER schemes cannot adjust the desired level of throughput or queuing delay for a particular traffic load. Thus, a RT scheduling scheme should be able to balance the tradeoff between throughput and queuing delay through the attunement of a scheduling parameter.

As previously mentioned, RT applications are highly delay sensitive and hence these applications define certain delay thresholds based on the requirements of their traffic class. Scheduling schemes can maintain these delay thresholds by dropping packets when their queuing delay exceeds the maximum tolerable amount at the expense of a certain error rate. For instance conversational traffic such as Voice-over-IP (VoIP) applications can tolerate a certain amount of error as a conversation can be understood if certain words from a sentence are missing. However, if the response time exceeds a certain tolerable amount then the call is rendered useless. Moreover, as RT scheduling schemes try to minimize queuing delay, majority of the packets are transmitted before their expiry term which consequently minimizes the error rate. Hence, an HSDPA scheduling scheme should maintain the delay thresholds of RT traffic applications and at the same time minimize the cumulative queuing delay of the system.

Existing HSDPA schemes are capable of scheduling only one type of traffic i.e. RT or NRT. Although, RT scheduling algorithms can be used to prioritize NRT traffic by the assignment of certain threshold delays (e.g. a high delay threshold value can be assigned to NRT traffic for the M-LWDF scheme), this changes the traffic characterization of these classes as NRT applications do no have a strict response time. Therefore, a successful HSDPA scheduling scheme should be adaptable in a mixed traffic environment such that it can schedule RT traffic based on their delay requirements and NRT traffic based on the channel conditions of the users. In other words, a scheduling scheme should support RT as well as NRT applications simultaneously without changing their traffic characterization.

Based on the above discussion, it can be concluded that the objective of a successful HSDPA scheduling algorithm should not only be throughput maximization but also the distribution of system resources in a fair manner so as to prevent the starvation of users with relatively poor channel conditions. Moreover, as RT applications are highly delay sensitive and require that packets are transmitted within a certain delay threshold, a scheduling scheme designed for these applications should try to minimize the average queuing delay of the system, in order to maintain the delay requirements of their respective traffic classes. Cellular providers offer various services simultaneously and may differentiate users based on their choice of service. Hence, a scheduling scheme should be capable of prioritizing users based on their QoS requirements. Furthermore, an HSDPA scheduling algorithm should be able to support a variety of RT and NRT applications simultaneously.

1.2 Thesis Contributions

The main contributions of this thesis are:

- The formulation of an HSDPA packet scheduler that considers the channel conditions of users as well as their instantaneous queuing delay in the scheduling assignment. To this end, the proposed scheme known as the Delay Based Scheduler (DBS) can maintain the delay thresholds of RT traffic applications as well as provide QoS prioritization to users that are distinguished based on their service requirements. Furthermore, the DBS increases the scheduling priority of users as their packets approach their expiry term thus giving all users an equal chance of being scheduled at higher delays and thereby enhancing system fairness.

- The development of a mechanism to adjust the tradeoff between throughput maximization and the minimization of queuing delay thus allowing the service provider to balance the two metrics.
- The inclusion of a mathematical proof that illustrates that the DBS can be used to schedule RT as well as NRT traffic simultaneously through the convergence of the scheme to the Max CIR algorithm by the setting of certain parameters.
- The implementation of an HSDPA system model in a network simulating environment and the deployment of the DBS as well as other conventional scheduling algorithms; such as the Max CIR, PF, M-LWDF and ER schemes; in the simulation model.
- The performance evaluation of the proposed scheme and its comparison with other HSDPA schemes in various traffic and simulation environments.

We further demonstrate that the proposed scheme achieves the following six design objectives: maximizing throughput, minimizing queuing delay, fairness, maintaining delay thresholds, QoS prioritization and scheduling RT and NRT traffic classes simultaneously.

1.3 Thesis Organization

The rest of this thesis is organized as follows: Chapter 2 provides a brief overview of the UMTS and HSDPA cellular systems along with their relying technologies and features. In addition, the chapter reviews some of the HSDPA packet scheduling schemes from literature. Chapter 3 outlines the design objectives of a successful HSDPA packet scheduler and presents the formulation of the proposed scheme. Simulation results are provided and discussed in Chapter 4 where the performance of the DBS is compared with to existing HSDPA scheduling schemes. Chapter 5 concludes this document by summarizing the contributions proposed in this thesis as well as suggesting some future research work.

Chapter 2

Background

Before presenting the proposed Delay Based Scheduler (DBS) scheme, it is crucial to provide the background material that will help the reader understand the remainder of this thesis. Therefore, this chapter presents an overview of the Universal Mobile Telecommunication Systems (UMTS) and its downlink extension for packet data services known as High Speed Downlink Packet Access (HSDPA). Firstly, the evolution of UMTS from previous mobile communications is discussed in Section 2.1. Secondly, the UMTS network architecture is described in Section 2.2. An overview of HSDPA is presented in Section 2.3. Various packet scheduling schemes in the literature are reviewed and compared in Section 2.4. Finally, a summary of this chapter is presented in Section 2.5.

2.1 Evolution of Wireless and Mobile Communications

Electromagnetic waves were first discovered as a communications medium at the end of the 19^{th} century [7]. Mobile telephone services were introduced in the late 1940s in the form of a car phone service. However, this service was limited to a single cell environment and hence was constrained by restricted mobility, poor speech quality, low capacity and limited service. Moreover, those early single cell systems were provided with heavy, bulky and expensive equipment which was susceptible to interference. Due to these limitations, less than one million subscribers were registered worldwide by the 1980s [8].

The cellular systems that were introduced in the late 1970s and early 1980s represented a quantum leap in mobile communication due to the introduction of smaller and lighter equipment. These cellular systems only transmitted analog voice information and were known as the first generation (1G) cellular systems. The mobile market showed an annual growth rate which was between 30 to 50% with the introduction of 1G cellular systems [8]. By 1990, around 20 million mobile subscribers were registered worldwide.

The need to improve transmission quality, system capacity and coverage led to the development of second generation (2G) cellular systems. 2G cellular systems saw a further advancement in semiconductor technology as well as microwave devices which brought digital transmission to mobile communications. Moreover, the increasing demand of data services over the cellular network led to the introduction of data transmissions in 2G cellular systems. Security services such as data encryption and fraud prevention, which were standard practices over a wired network, were introduced for cellular networks. Various 2G cellular systems were developed such as: Global System for Mobile communications (GSM), Code Division Multiple Access (CDMA), Digital AMPS (D-AMPS) and Personal Digital Communication (PDC). Presently, multiple 1G and 2G standards are used in mobile communications. The standards vary in terms of capability, mobility, service area and supportable applications. Moreover, GSM is the most successful family of cellular standards with around 250 million subscribers that enjoy international roaming in approximately 140 countries and 400 networks [8].

The European Telecommunications Standard Institute (ETSI) completed Phase 1 of GSM standardization in 1990. GSM has a supportable data rate of up to 9.6 kbps. Due to the increasing popularity of multimedia applications over the Internet; such as text messaging, audio and video streaming; there was a need to improve the supportable data transmission rate over the wireless medium. As a result, GSM standards were enhanced in Phase 2 to incorporate cellular technologies that would improve the supportable data rate. 2.5G and third generation (3G) wireless cellular systems evolved with the release of Phase 2+ of GSM standards which introduced features such as network services with Customized Application for Mobile Enhanced Logic (CAMEL), speech Compression/Decompression (CODEC), Adaptive Multi-Rate (AMR), etc. 2.5G wireless cellular systems include: the General Packet Radio Service (GPRS) and Enhanced Data rate for GSM Evolution (EDGE) [8]. Universal Mobile Telecommunication System (UMTS) is a 3G GSM successor that is backward compatible with GSM. Figure 2.1 compares the data rates of UMTS with other wireless cellular systems.

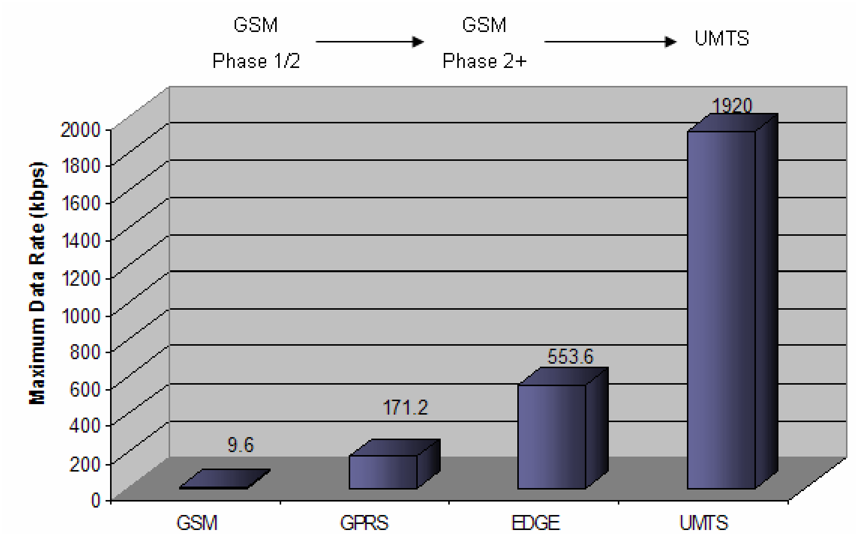


Figure 2.1: Data Rates offered by various cellular systems

2.2 Universal Mobile Telecommunication System

Universal Mobile Telecommunication System (UMTS) is being developed by the Third Generation Partnership Project (3GPP). UMTS provides a data rate of up to 2 Mbps and hence can address the growing demand of data applications such as web browsing, text messaging, etc over the wireless medium. Moreover, UMTS has established a global roaming standard that allows users to access non-local cellular networks that are affiliated with their service providers.

This Section presents an overview of UMTS. Section 2.2.1 describes the UMTS network architecture. The UMTS radio interface protocol is outlined in Section 2.2.2. Finally, the UMTS Quality of Service (QoS) classes are presented in Section 2.2.3.

2.2.1 UMTS Network Architecture

The UMTS network architecture is similar to that employed by various 2G and 2.5G wireless cellular systems such as GSM and GPRS. It consists of three main logical entities: the UMTS Terrestrial Radio Access Network (UTRAN), the Core Network (CN) and the User Equipment (UE) [9]. The UTRAN is further subdivided into of one or more Radio Network Sub-systems (RNS). Each RNS comprises of a Radio Network Controller (RNC) which is connected to one or more base stations (Node Bs). The CN comprises of a Serving GPRS Support Node (SGSN) as well as a Gateway GPRS Support Node (GGSN). It is responsible for data connections to external networks. Figure 2.2 shows a simplified version of the UMTS architecture.

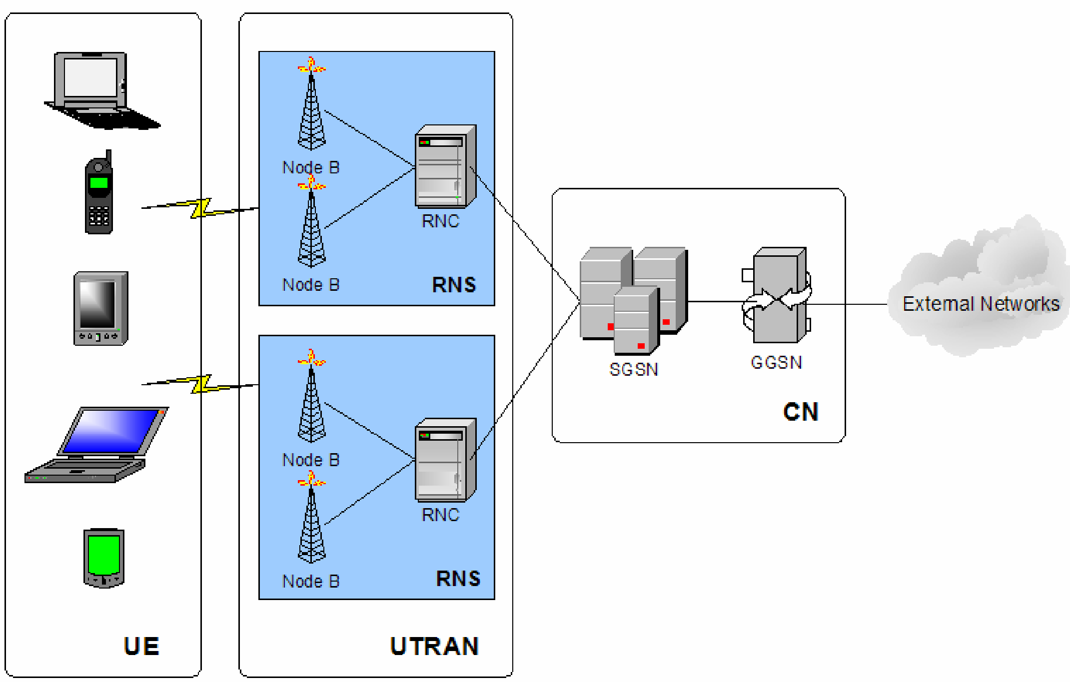


Figure 2.2: UMTS Network Architecture

The UE provides the user with direct access to the network and, therefore, it acts as the interface between the user applications and the network services. It is connected to the cellular network through the Node B via a radio interface and is identified through the UMTS Subscriber Identity Module (USIM). The Node B handles the users that are within its coverage area and may control more than one cell depending on cell sectoring. The main function of the Node B is the conversion of data to and from the radio interface which involves processes such as forward error correction, rate adaptation, spreading, channel coding and Quadrature Phase Shift Keying (QPSK). Furthermore, it provides the channel condition information of each UE to the RNC which is required for the Radio Resource Management (RRM) as well as the packet scheduling operations.

The RNC autonomously controls the radio resources of the UTRAN. It interfaces with the CN as well as manages the procedures between the UE and the UTRAN through the Radio Resource Control (RRC) protocol. The RNC implements all of the RRM procedures which include but are not limited to: power control, handoff management, call admission control and packet scheduling [9]. The power control algorithms determine the optimal amount of power that should be used to transmit signals between the Node B and UE. Handoff management algorithms decide when an ongoing call should be transferred to another Node B as the user moves from the coverage area of one cell to the coverage area of another cell. Call admission control algorithms determine whether a handoff or new call should be accepted based on factors such as the number of available channels, the amount of available power, etc. Finally, the packet scheduling algorithm decides the user that transmits (or receives) data at any given time interval and, hence, manages the radio resources accordingly. Therefore, the packet scheduling algorithm has a huge impact on the performance of the system in terms of throughput, queuing delay, packet loss, etc. The RRC protocol defines the messages and procedures between the UE and the UTRAN. It is used by the RNC to interface between these two network elements. Some of the functions offered by the RRC are broadcasting information, connection management between the UE and UTRAN, ciphering control, paging and notification, etc. Further details of the RRC protocol can be found in [10].

The CN interfaces between the UMTS network and the external networks. The CN operates in two domains which are the Circuit-Switched Domain and the Packet-Switched Domain. The Circuit-Switched Domain provides real-time services such as voice, whereas the Packet Switched Domain connects to external data networks and provides data services such as email and Web browsing.

2.2.2 UMTS Radio Interface Protocol Architecture

The UMTS radio interface protocol architecture consists of three main protocol layers: Radio Link Control (RLC) layer, Medium Access Control (MAC) layer and the Physical layer. Each UMTS network entity consists of one or more of these protocol layers. Moreover, these protocol layers are interfaced through three different channels which are the logical, transport and physical channels. The physical channel is the transmission medium between different network entities and is defined by a specific carrier frequency, scrambling and channelization codes and time duration. The physical layer offers services to the MAC layer by means of the transport channel which is characterized by the method employed for data transmission. Consequently, the MAC layer offers services to the RLC layer via the logical channel which is defined by the type of data transmitted. The physical channel exists between the UE and Node B whereas the transport and logical channels exist between the UE and the RNC [8].

This section outlines the UMTS radio interface protocol architecture. The RLC and MAC protocols that are implemented for UMTS are described in Sections 2.2.2.1 and 2.2.2.2 respectively. The physical layer is discussed in Section 2.2.2.3.

2.2.2.1 Radio Link Control Protocol

The RLC protocol layer is implemented at the RNC as well as the UE. It is responsible for packet segmentation, packet concatenation and padding, in-sequence and out-of-sequence delivery, retransmissions and error control. The senders' RLC protocol layer is responsible for the segmentation of packets or Service Data Units (SDUs) into Protocol Data Units (PDUs). Conversely, the receiving RLC protocol layer is responsible for the concatenation of the PDUs to their respective SDUs. Moreover, segmentation, concatenation and padding are provided by means of header fields added to the data.

Each RLC instance can be configured to operate in one of three modes: Acknowledged Mode (AM), Unacknowledged Mode (UM) or Transparent Mode (TM) [11]. In the AM, an Automatic Repeat Request (ARQ) procedure is used for error control and packet retransmission. However, if the maximum number of retransmissions is reached and the RLC is unable to deliver the data correctly, then the SDU is discarded and the peer entity is informed as the AM is bi-directional and capable of piggybacking[1] data delivery status to the sending user. The AM, strives to maintain a low SDU error ratio at the cost of variable delay by retransmitting every unsuccessfully received RLC PDU. Therefore, the AM is most suitable for non real-time applications that are not delay sensitive, for instance web browsing and File Transfer Protocol (FTP) applications. In the UM, retransmission procedures are non-existent and thus data delivery is not guaranteed. Incorrectly received PDUs are either discarded or marked erroneous. For this reason, an RLC entity in the unacknowledged mode is considered unidirectional and is, therefore, used for real-time applications, such as video streaming and Voice-over-IP (VoIP) services. The TM offers a circuit switched service and does not require any protocol overhead at the higher layers. In the TM mode, SDUs can be transmitted without segmentation and erroneously received PDUs are discarded.

2.2.2.2 Medium Access Control Protocol

As mentioned above, the MAC layer offers various data services to the RLC layer through the logical channels. Each of these data services is mapped to a logical channel. Therefore, a set of logical channel types is defined and characterized by the type of information transferred. Logical channels can be classified into two groups: control channels and traffic channels. The various control channels are:

[1] The technique of temporarily delaying acknowledgements so that they can be added to the next outgoing data frame is known as piggybacking.

- Dedicated Control Channel (DCCH) transmits dedicated control information to and from a UE.
- Common Control Channel (CCCH) transfers control information between the network and a group of UEs.
- Broadcast Control Channel (BCH) for broadcasting system control information.
- Paging Control Channel (PCCH) that transfers paging information.

There are two types of traffic channels, which are:

- Dedicated Traffic Channel (DTCH) used to transfer information dedicated to a particular user in a bi-directional manner.
- Common Traffic Channel (CTCH) used to broadcast information that is relevant to a group of users.

The MAC layer is also responsible for mapping the logical and transport channels. In addition, it selects the appropriate Transport Format (TF) for each transport channel based on the instantaneous source rate. Furthermore, the MAC layer manages the priority handling between various UEs (or between the various data flows of the same UE). A detailed description of the MAC functionalities is described in [12]. In order to support all the functionalities implemented at the MAC layer, it is subdivided into the following three logical entities:

- MAC-b supports the Broadcast Channel (BCH).

- MAC-c/sh is responsible for handling the common and shared channels.
- MAC-d handles the Dedicated Channel (DCH).

The UMTS radio interface protocol architecture along with the various MAC entities is shown in Figure 2.3.

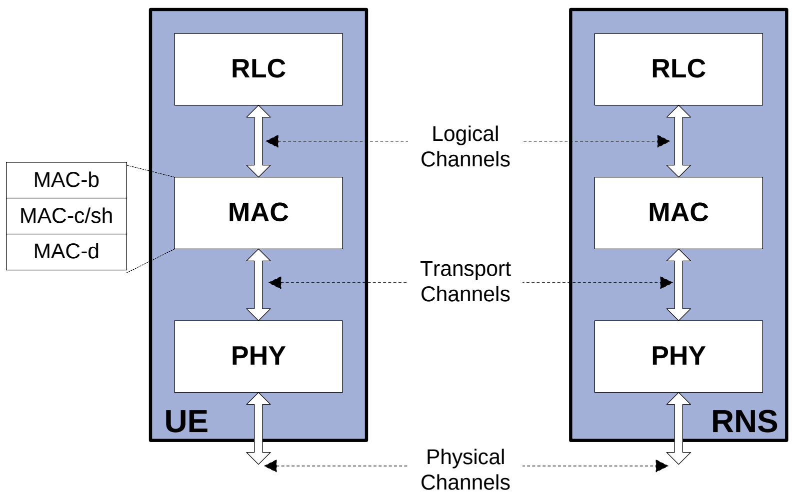


Figure 2.3: UMTS Radio Interface Protocol Architecture

2.2.2.3 Physical Layer

The physical layer is responsible for the transmission of the PDUs. Data is transmitted over the radio interface with the help of transport channels which are mapped to the different physical channels by the physical layer. The various physical channels are described in detail in [13]. There are two different types of transport channels which are the dedicated and common channels. The dedicated channel is a resource available to a single user, identified by a particular code on a certain frequency, whereas the common channel is a resource available to all (or a group of) users. Only one dedicated transport channel exists which is known as the Dedicated Channel (DCH). The six different common transport channels are:

- Broadcast Channel (BCH) broadcasts cell specific information.
- Forward Access Channel (FACH) is a downlink channel that transmits control information to the UEs.
- Random Access Channel (RACH) is an uplink channel that carries control information from the UEs.
- Paging Channel (PCH) carries the paging procedure data in the downlink direction.
- Uplink Common Packet Channel (CPCH) is an uplink channel that carries packet data from the UEs.
- Downlink Shared Channel (DSCH) carries dedicated user data or control information that can be shared by several users.

Data is sent over the radio interface in the form of a radio frame structure which is divided into 15 slots. Each of these radio frames is sent over a time interval of 10 ms.

2.2.3 UMTS QoS Classes

The UMTS Quality of Service (QoS) classes are distinguished based on the delay sensitivity of the various traffic types. The 3GPP has specified four different QoS classes for UMTS in [14], which are:

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

The classes are stated in descending order of delay sensitivity with the conversational class being the most delay sensitive traffic class whereas the background class is the least delay sensitive traffic class. The conversational class is intended for applications such as VoIP and video telephony. The streaming class is more delay-tolerant than the conversational class. However, streaming applications require the end-to-end delay variations to be within a certain limit. These applications buffer the received stream before decoding it in order to maintain a certain value of delay jitter. Audio and video streaming are examples of streaming applications. The conversational and streaming classes are highly delay sensitive and, therefore, are known as Real-Time (RT) traffic classes. The interactive class represents bursty traffic such as email access and Web browsing where the user expects to receive data from a remote server within a certain time limit. The background class is the most delay tolerable traffic class with the least error rate. Server-to-server communication such as an FTP application is an example of this traffic class. The interactive and background classes are delay tolerable and are classified as Non-Real-Time (NRT) traffic classes. A comparison between the above mentioned QoS classes is shown in Table 2.1.

Table 2.1: Comparison of the four UMTS QoS classes

Traffic Class	Delay Sensitivity	Other Characteristics	Application
Conversational	Highly delay sensitive	RT traffic class with less than one tenth of the packets dropped	VoIP
Streaming	Delay sensitive with low delay jitter limit	RT traffic class that maintains a low error rate	Video Streaming
Interactive	Delay tolerable with a low response time	NRT traffic class that preserves the payload content	Web Browsing
Background	Delay insensitive	NRT traffic class which is most error sensitive	FTP

2.3 High Speed Downlink Packet Access

The 3GPP has standardized a 3.5G cellular system known as High Speed Downlink Packet Access (HSDPA) which is an extension to the existing 3G wireless communication system known as UMTS. HSDPA can support a data rate of up to 14.4 Mbps which is much beyond what 2G and 3G cellular systems could offer. Apart from increasing the peak data rate, HSDPA aims at achieving lower delays and improving the spectral efficiency of asymmetrical and bursty packet downlink services. This is accomplished through the implementation of a new shared transport channel known as the High Speed Downlink Shared Channel (HS-DSCH). UMTS has already implemented a downlink shared channel known as the DSCH; however, HSDPA extends this concept in order to significantly improve the overall system capacity. In addition, the HSDPA packet scheduling procedure is deployed at the Node B instead of the RNC (as in the case of UMTS). Since each RNC controls multiple base stations, this deployment distributes the workload among the base stations and consequently increases system efficiency. As will be shown in the forthcoming sections, the scheduling scheme is capable of achieving the desired system QoS. The increased efficiency and high supportable data rate of HSDPA has encouraged the convergence of various multimedia applications to this technology. Moreover, various new features have been added to HSDPA which include: Adaptive Modulation and Coding (AMC) and Hybrid Automatic Repeat Request (HARQ). The 3GPP specifications for the HSDPA cellular system are given in [2].

This section outlines the most important features implemented in HSDPA. Furthermore, it highlights the enhancements made to UMTS in order to implement this wireless communication system. Section 2.3.1 describes the HSDPA network architecture as well as its radio interface. The HSDPA MAC architecture is shown in Section 2.3.2. AMC and HARQ procedures are discussed in Sections 2.3.3 and 2.3.4 respectively. Finally, channel dependent scheduling is explained in Section 2.3.5.

2.3.1 HSDPA Network Architecture and Radio Interface

HSDPA improves the overall system efficiency by making minimal changes to the existing UMTS network architecture. In the UMTS radio interface protocol architecture, the two most important protocol layers that are implemented at the Node B are the Medium Access Control (MAC) layer and the physical layer (as discussed in Section 2.2.2). With the introduction of HSDPA, an additional MAC sub-layer was introduced at Node B which is known as Medium Access Control-high speed (MAC-hs). In UMTS, the packet scheduling procedure is performed at the RNC; however, this procedure is moved to the MAC-hs sub-layer of the Node B in the case of HSDPA. This reduced the time required for the packet scheduling decision, as some of the information required for scheduling is already available at Node B. Moreover, HSDPA converts the scheduling operation from a pipelined to a parallel procedure. This is because each RNC controls multiple Node Bs and is in-charge of the scheduling decision for each of them. With the implementation of the scheduling function at the Node B, the network is converted from a central to a distributed scheduling system. The MAC-hs is further discussed in Section 2.3.2.

UMTS consists of a Dedicated Physical Channel (DPCH) which is subdivided into separate control and data channels in the uplink and downlink direction [13]. It is to be noted that the DPCH is associated with the Downlink Shared Channel (DSCH). New transport and physical channels were added to the UMTS specification in order to support all the features of HSDPA. The new transport channels comprise of:

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

The HS-DSCH is a time shared transport channel that carries dedicated user data that is shared by a group of users. This transport channel maintains some of the characteristics of the release 99 DSCH and is mapped to one or more physical channels [2]. However, unlike the DSCH, the HS-DSCH does not carry user-specific downlink control information. This downlink control information is transmitted by the HS-SCCH thus allowing for the structure of the downlink DPCH to remain unchanged with the implementation of the HS-DSCH. The new physical channels that are added to support HSDPA are:

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

The HS-PDSCH is a physical channel that is mapped to its transport level counterparts i.e. the HS-DSCH and the HS-SCCH. Hence, the HS-PDSCH transmits user data and control information in the downlink direction. The HS-DPCCH is an uplink signaling channel that transmits information relevant for the HARQ process, such as acknowledgements, as well as the Channel Quality Indicator (CQI) that reflects the instantaneous channel condition of a UE. The above mentioned channels and their functionalities are summarized in Figure 2.4.

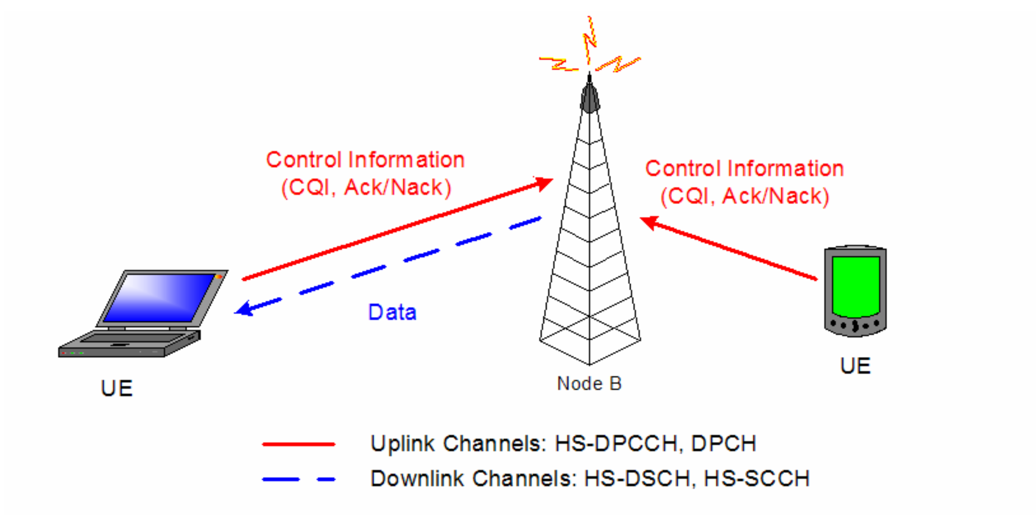


Figure 2.4: Transport and Physical channels of HSDPA

At every Transmission Time Interval (TTI) of 2 ms, the UE measures the strength of the signal transmitted by the Node B. This instantaneous channel condition is then used by the UE to evaluate a recommended Transport Block Size (TBS), modulation and channelization codes. The UE then sends this information to the Node B through the HS-DPCCH in an uplink sub-frame. Each of these sub-frames comprises of three slots where one slot is allocated for the HARQ acknowledgements and the remaining two slots are allocated for the CQI (shown in Figure 2.5). The UE reports the highest value of CQI that will result in a Block Error Rate (BLER) value which is less than or equal to 10%. As the CQI value is recommended by the UE, the Node B is not obliged to act on this information for the scheduling decision.

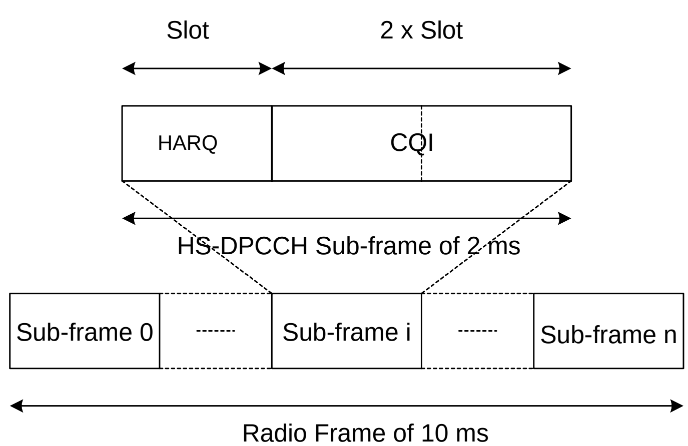


Figure 2.5: Frame Structure of the uplink HS-DPCCH

2.3.2 HSDPA Medium Access Control Layer Architecture

As mentioned in Section 2.3.1, an additional MAC sub-layer known as the MAC-hs is added at the Node B. The MAC-hs is responsible for the scheduling decision that was previously taken at the RNC. In order to accommodate the new MAC sub-layer at the

Node B, corresponding MAC-hs functionality was implemented at the MAC of the UE. A detailed view of the elements found in the UE side MAC-hs are shown in Figure 2.6. The HARQ entity is accountable for managing the HARQ procedure which is discussed in Section 2.3.4. A single HARQ process exists for every UE per TTI that handles all the tasks required for RLC level retransmissions. The configuration of the HARQ protocol is provided by the RRC. Successfully received data blocks are queued according to their transmission sequence number in the reordering queue distribution entity. Every UE may consist of more than one data flows by multi-tasking network applications. These data flows are distinguished based on their priority. The reordering entity further organizes the received data blocks based on their priority class. The de-assembly entity then generates the appropriate MAC-d PDUs from the consecutive data blocks available at the reordering queues.

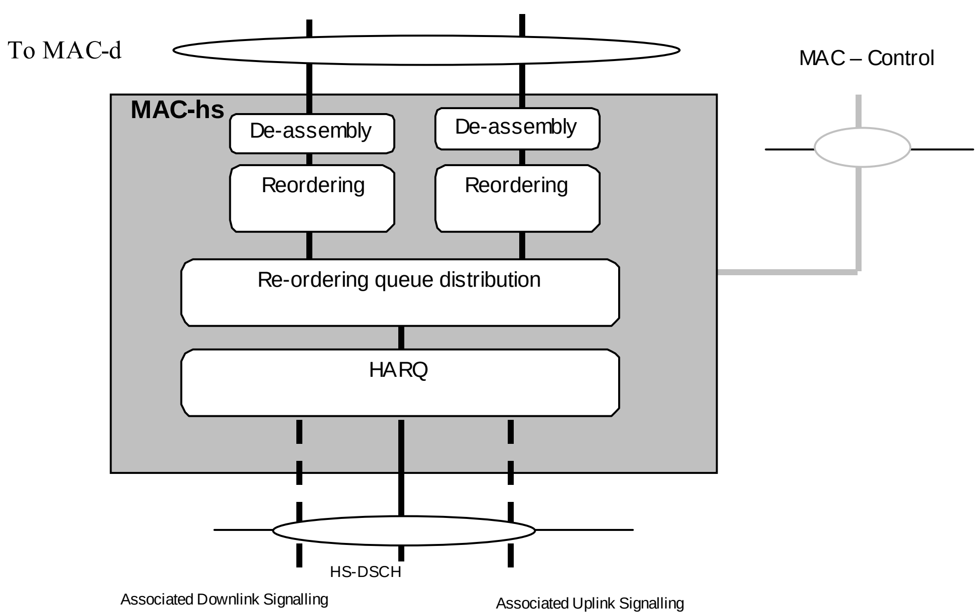


Figure 2.6: UE side MAC-hs architecture [15]

The MAC-hs at the UTRAN side consists of four different functional entities [15]:

- Flow Control
- Scheduling and Priority Handling
- Hybrid Automatic Repeat Request
- Transport Format and Resource Combination (TFRC) selection

The flow control entity between the RNC and Node B ensures that the MAC-hs buffer always contains enough data packets to maximize system throughput while avoiding packet loss due to buffer overflow. Moreover, the flow control mechanism guarantees that the MAC-hs buffer length is kept as low as possible in order to decrease required memory space, round-trip delay and packet loss at handoff. Flow control is provided independently by priority class for each MAC-d flow [15]. The 3GPP specifications in [16] have stated a flow control mechanism for the HS-DSCH known as the credit-based system. The priority handling entity is responsible for dividing the data flows according to their priority class. For instance, if two priority classes exist for each of two MAC-d flows, then a total of four MAC-d priority queues exist. The scheduling entity is responsible for making a scheduling decision based on a number of factors that will be discussed in Section 2.4. The scheduler can combine several MAC-d PDUs, depending on the value of the TBS, along with a MAC-hs header to form a transport block. The HARQ entity is similar to the one implemented at the UE side MAC-hs. The TFRC selection is based on the modulation and number of channelisation codes that are supportable by the UE. These values are indicated by the CQI, as mentioned at the end of Section 2.3.1. The architecture of the MAC-hs at the UTRAN side is shown in Figure 2.7.

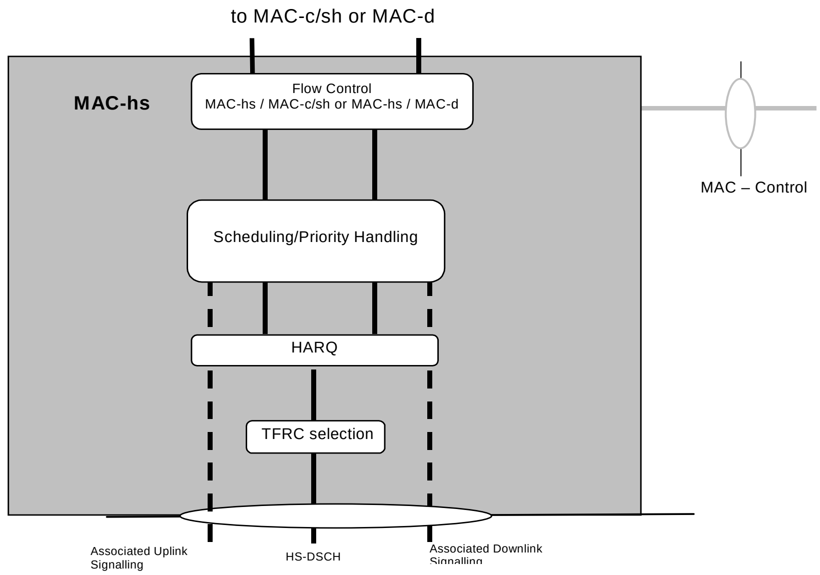


Figure 2.7: UTRAN side MAC-hs architecture [15]

2.3.3 Adaptive Modulation and Coding

In a cellular system, the channel condition experienced by a UE continuously varies due to: path loss, shadowing, multi-path fading and cell interference[1]; which are collectively known as the propagation loss model (see Chapter 4 for details). Earlier releases of UMTS used a power control technique to compensate for the fluctuations in the channel condition. This power control technique ensured a constant data rate by adjusting the transmission power and hence maintained a similar service quality. Therefore, users with poor channel conditions enjoyed data rates that were similar to those users with good channel conditions at the cost of greater transmission power. However, from a system point of view, it is more efficient to allocate the radio resources to a user experiencing good channel conditions because it requires less transmission power. Due to this reason, a

[1] Inter-cell interference occurs due to noise generated by surrounding cells and is most prominent at the cell edge; Intra-cell interference is caused by other user equipment within the same cell.

more efficient best-effort service was introduced that maintained a constant transmission power while varying the data rates. HSDPA employs link adaptation along with a predefined combination of modulation and channel coding to determine the supportable data rate of each UE per TTI. This technique is known as Adaptive Modulation and Coding (AMC). Each UE reports its channel condition by means of the CQI through the uplink physical channel known as HS-DPCCH. The CQI is then used by the Node B to determine the size of the transport block as well as the modulation and coding schemes that the user can support. The transport block size indicates the number of bits that can be sent to a user (given their channel condition) at a particular time interval. Therefore, users with good channel conditions enjoy potentially higher data rates with the help of higher modulation and coding rates. Contrarily, users with poor channel conditions have lower data rates. HSDPA employs a higher order modulation scheme known as Quadrature Amplitude Modulation (16QAM) which allows more bits per modulation symbol and hence higher data rates. In addition to 16QAM, HSDPA uses a scheme that dynamically adjusts to the channel coding rate known as Quadrature Phase Shift Keying (QPSK) which was included in the release 99 of UMTS [17].

2.3.4 Hybrid Automatic Repeat Request

The Automatic Repeat Request (ARQ) is an error detection mechanism used in previous releases of UMTS. In the ARQ scheme, data blocks that are incorrectly received are discarded and retransmitted. The transmitter uses a stop-and-wait procedure which transmits a data block and waits for a response from the receiver before sending a new data block or retransmitting an incorrectly received data block. However, in HSDPA the

Hybrid ARQ (HARQ) scheme is used where incorrectly received data blocks are not discarded but stored and soft-combined with retransmissions of the same information bits. These combined information blocks are then decoded and further retransmissions occur if the decoding process is unsuccessful.

The 3GPP specifications have defined two HARQ processes for HSDPA: Incremental Redundancy and Chase Combining [17]. In the former scheme, successive retransmissions of an incorrectly received data block are sent with additional redundancy that is increased with each consecutive retransmission. The retransmissions consist of redundant information in order to increases the chances of successful delivery. Since each transmitted block is not the same as the previous transmission, it is demodulated and stored at the receiver and consequently soft-combined to reproduce the original data block [7]. In the chase combining strategy, an erroneously received data packet is stored and soft-combined with later retransmissions that are an exact copy of the original transmission [18].

2.3.5 Channel Dependent Scheduling

HSDPA is based on the use of a time shared resource, viz. the HS-DSCH. Therefore, the assignment of this shared resource to a particular user at a given time interval is a crucial decision. HSDPA uses the channel condition of a UE in the scheduling decision apart from other factors such as queuing delay, priority assignments, etc. As the packet scheduler can exploit the short-term variations in the radio conditions of a user, it can maximize the overall system throughput by selecting users with favorable channel conditions. This is because good channel conditions permit higher supportable data rates through the selection of higher order modulation and coding schemes. As mentioned earlier, the HSDPA scheduling decision is made at the MAC-hs sub-layer of the Node B. This reduced the time required for the packet scheduling decision, as there is no delay involved in conveying the channel condition information (or CQI) to the Node B. Moreover, in UMTS the scheduling decision was made at the RNC which is responsible for controlling multiple base stations and hence was in-charge of the scheduling procedure at each of them. With the implementation of the scheduling function at the Node B, the network was converted from a semi-central scheduling system; where all the scheduling decisions were made at the RNC; to a distributed scheduling system resulting in a higher system capacity and efficiency. Various packet scheduling schemes are discussed in Section 2.4.

2.4 HSDPA Scheduling Schemes

Several scheduling schemes for HSDPA have been proposed and studied in literature. These scheduling algorithms consider the channel condition of a user while making the scheduling decision. The scheduler can maximize system throughput by the selection of a user with good channel conditions. Users with favorable radio channel conditions can support a high Transport Block Size (TBS) allowing them to enjoy high data rate downloads. However, it should be noted that favoring users with good channel conditions may lead to the starvation of those with poor channel conditions, as it prevents those users from being served. Therefore, users who pay for the same service can be alienated due to unfavorable channel conditions. A good scheduling scheme should not only try to maximize overall throughput but also manage system resources so that all users are treated in a fair manner.

The various QoS classes (discussed in Section 2.2.3) can be classified into two groups viz. Real-Time (RT) and Non-Real-Time (NRT) classes based on their delay requirements. RT traffic applications, such as VoIP and video streaming, have stringent QoS requirements as they are highly delay sensitive and thus scheduling schemes for these applications aim at maintaining their delay requirements and consequently minimizing the overall queuing delay of the system. Contrarily, NRT traffic applications are delay tolerable and hence NRT scheduling algorithms try to balance the tradeoff between system throughput ad fairness. Moreover, the HSDPA wireless cellular system is capable of measuring the channel conditions of users which can be used for the scheduling prioritization in order to maximize system throughput. Before the scheduling algorithms are discussed, the following notations should be defined:

- t represents the Transmission Time Interval (TTI).
- $SF_i(t)$ is the priority assigned to user i at time t where $i \in \{1,2,\ldots,n\}$ and n is the total number of users in the system. At any given time interval, the user with the highest $SF_i(t)$ value is scheduled.
- $R_i(t)$ is the instantaneous supportable data rate of user i at time t. It is evaluated from the size of the transport block which in turn is determined from the value of the channel quality indicator.

- $S_i(t)$ is the average throughput of user *i* at time *t*. It is the ratio of the sum of the bits transmitted to the user through the HS-DSCH up to time *t* over the total time spent by the UE in the system.
- $Q_i(t)$ is the queuing delay experienced by the head packet of flow *i* at Node B up to transmission time interval *t*.

This Section reviews the HSDPA scheduling schemes. NRT and RT scheduling schemes are discussed in Sections 2.4.1 and 2.4.2, respectively. Finally, a comparison between the scheduling algorithms is given in Table 2.2. It should be noted that all of the below mentioned schemes schedule users at every TTI (2 ms) since they utilize the instantaneous channel conditions of the users for the scheduling criteria.

2.4.1 Non Real-Time Scheduling Schemes in HSDPA

NRT scheduling algorithms are suitable for asymmetric and bursty traffic (i.e. interactive and background traffic). These schemes are based around the instantaneous channel condition of the users and try to maximize system throughput by scheduling users with good channel conditions. This section presents several scheduling algorithms for NRT traffic in HSDPA.

Round Robin

The Round Robin scheme schedules users in a cyclic order. This scheme existed much before the introduction of HSDPA; hence it ignores the channel conditions of users while making the scheduling decision. Therefore, this scheduling scheme is unable to maximize system throughput. However, it is a simple algorithm with a low complexity and thus is easy to implement. Additionally, the Round Robin scheme fairly distributes the system resources among all users in the cell, regardless of their channel conditions.

Maximum Carrier-to-Interference Ratio

The Maximum Carrier-to-Interference Ratio (Max CIR) scheme serves the user with the highest instantaneous supportable data rate at every TTI [3]. Obviously, this scheme maximizes system throughput by selecting users with favorable channel conditions. However, the main drawback of this algorithm is that it does not fairly distribute system resources, which may lead to the starvation of users with poor channel conditions. This scheme was one of the first scheduling algorithms proposed for HSDPA, but the unfairness issue has led to the proposal of several fair scheduling schemes that try to maximize system throughput.

Proportional Fairness

The Proportional Fairness (PF) scheduling scheme serves the user with the largest relative channel quality [4]:

$$SF_i(t) = \frac{R_i(t)}{S_i(t)}$$

This algorithm tends to serve users under favorable instantaneous channel conditions relative to their average ones, thus taking advantage of the temporal variations of the fast fading channel. In other words, this algorithm serves users with lower average throughputs who have favorable channel conditions. Therefore, this scheme provides a better degree of fairness when compared to the Max CIR scheme, while it considers the instantaneous channel conditions of users. However, recent studies in [19] have shown that the PF scheme gives privilege to users with high variance in their channel conditions, and hence fails in allocating the system resources to all users in a fair manner. This is due to the fact that in HSDPA only a finite set of discrete rate values exist.

Data Rate Control Exponent

As mentioned above, the PF scheme encounters an unfairness problem due to the allocation of the shared channel to users with high variance in their channel conditions. To solve this unfairness problem, a modified version of the PF algorithm was proposed in [19], which is known as the Data Rate Control (DRC) Exponent scheme. This scheme attempts to find a solution to the unfairness problem by adding a constant exponent term c to $R_i(t)$:

$$SF_i(t) = \frac{R_i(t)^c}{S_i(t)}$$

The DRC Exponent scheme emphasizes the instantaneous channel conditions of users, as opposed to their relative channel quality. Therefore, if the exponent c is greater than 1, then this scheme schedules users with better channel conditions. However, two main problems arise due to fixing a constant value of c for all users. Firstly, fixing the control parameter in time does not adapt to the time-varying radio conditions of each UE. Secondly, as the control parameter takes a unique value for all users, it is not possible to ensure fairness among all users at the same time [20].

Adaptive Proportional Fairness

To solve the problems encountered with the DRC Exponent rule, a scheme known as the Adaptive Proportional Fairness (APF) algorithm was proposed in [20]. This scheme enhances the DRC scheme by adding an independent exponent term c_i for each user, in order to avoid the dependency between different users. The user selection criterion for this algorithm is given by:

$$SF_i(t) = \frac{R_i(t)^{c_i}}{S_i(t)}$$

Moreover, to account for the fast variations in channel conditions, the APF algorithm adds a monitoring module which updates the values of c_i. The scheduling takes place at every TTI whereas the control parameters are updated at a larger time scale. The updating module verifies whether the difference between the proportional data rate allocated to a user and the average value over all users, is within acceptable limits which are defined by a fixed interval. If the condition is not satisfied, then the control parameters of all users are updated. Simulation results in [20] show that the APF algorithm outperforms the PF algorithm in terms of fairness with no significant loss in total data rate.

Fast Fair Throughput

In [21], an efficient modification of the PF algorithm is proposed that aims at providing a fair throughput distribution among all the users of the HS-DSCH, while taking advantage of the short-term variations of the radio channel. The algorithm schedules users at every TTI, according to the following rule:

$$SF_i(t) = \frac{R_i(t)}{S_i(t)} \cdot \left[\frac{\max_j \left\{ \overline{R_j(t)} \right\}}{\overline{R_i(t)}} \right]$$

where $\overline{R_i(t)}$ denotes the average supportable data rate of user i up to time t and $\max_j \left\{\overline{R_j(t)}\right\}$ is the maximum average data rate among all users of the system at time t. Note that the addition of the $\overline{R_i(t)}$ term in the denominator tends to increase the priority of users with less favorable channel conditions. Therefore, this scheme tries to distribute the cell throughput to all users in an even and fair manner.

Weighted Fair Scheduling

Ki et al. argue that any system is bound to provide lower throughput to some users in order to maximize system capacity in [22]. They describe the fairness of a scheduling scheme as its ability to provide a certain minimal throughput to users with unfavorable channel conditions. According to them, the fairness index of user i can be described as:

$$F_i = \frac{S_i}{S_{avg}} = \frac{S_i}{\frac{1}{n}\sum_{i=1}^{n} S_i}$$

where F_i signifies the fairness metric of user i and S_{avg} denotes the average throughput of all users. This implies that the fairness index of a user can be defined as the ratio of the users' average throughput to the average throughput of all users in the cell. They propose the Weighted Fair Scheduling (WFS) scheme by augmenting the above mentioned fairness index to the PF scheme. The WFS algorithm assigns priorities to users according to the following formula:

$$SF_i(t) = \frac{R_i(t)}{S_i(t)} \cdot \left[\frac{S_{avg}(t)}{S_i(t)}\right]$$

In the WFS scheme, the scheduling priority of a user is inversely proportional to their fairness index, thus giving a higher weight to users who have relatively low fairness. Therefore, users with poor channel conditions as well low fairness values are given a higher priority in order to provide certain minimal throughput to these users.

Throughput Guarantee Scheduling

In [22], it is stated that the network should be able to provide a minimum throughput to all users in the cell, in order to be able to guarantee the QoS requirements. The throughput outage probability of a user is defined as the probability that the users' average throughput cannot satisfy the minimum throughput requirement:

$$\Pr_i\left(S_i < S_{i,req}\right) \leq \rho_i$$

where $S_{i,req}$ is the required minimum throughput of user i and ρ_i is their throughput outage probability. Ki et al. propose a scheme in [22] that modifies the PF scheduler in order to meet the minimum throughput requirements of a user and hence guarantee their QoS. This scheme is known as the Throughput Guarantee Scheduling (TGS) scheme where the priority of a user is generated as follows:

$$SF_i(t) = \begin{cases} \dfrac{R_i(t)}{S_i(t)} \cdot \left(C_i\right)^{S_i / S_{i,req}} & if \quad S_i < S_{i,req} \\ \dfrac{R_i(t)}{S_i(t)} & if \quad S_i \geq S_{i,req} \end{cases} \qquad where \quad C_i = \frac{\max_j \left\{\overline{R_j(t)}\right\}}{\overline{R_i(t)}}$$

where $\max_j \left\{\overline{R_j(t)}\right\}$ is the maximum average data rate among all users of the system and C_i is a weight added to increase a users' scheduling probability if their throughput at a particular time interval is less than that specified by their QoS requirements. Furthermore, the exponent term $S_i / S_{i.req}$ is the normalized throughput of user *i* whose value tends to 1 as the users' throughput tends to their required minimum value. Therefore, this exponent term gives a higher priority to users who can readily achieve their target throughput. Conclusively, it can be said that the TGS scheme aims at maximizing the number of satisfied users in the network environment.

2.4.2 Real-Time Scheduling Schemes in HSDPA

With the advent of the Internet, the use of RT traffic applications; such as video telephony; have become day-to-day events. However, theses applications impose strict delay constraints in order to meet their QoS requirements. For this reason, RT scheduling schemes need to consider the queuing delay of flows at the Node B buffer for the scheduling decision. Moreover, HSDPA measures the instantaneous channel conditions of users which can be used to maximize system throughput. Hence, RT scheduling schemes in HSDPA should utilize the queuing delay as well as the channel conditions of users in order to support delay sensitive applications. This section discusses some well known RT scheduling algorithms for HSDPA.

Modified Largest Weighted Delay First

The Modified Largest Weighted Delay First (M-LWDF) algorithm was proposed by Andrews et al. in [5]. This algorithm attempts to keep the probability of the queuing delay exceeding the due time below a certain ratio all the while trying to utilize the wireless channel efficiently. The M-LWDF computes the priority of user *i* at every TTI as follows:

$$SF_i(t) = -\log(\rho_i) \cdot \frac{R_i(t)}{S_i(t)} \cdot \frac{Q_i(t)}{T_i}$$

where T_i expresses the discard timer parameter for user *i* and ρ_i denotes a QoS parameter that allows to differentiate between users with different QoS requirements (such as end-to-end delay). Moreover, ρ_i can be considered as the probability of a user to exceed their delay requirements:

$$\Pr(Q_i > T_i) \leq \rho_i$$

A user with a higher value of ρ_i has a higher probability of exceeding its delay requirement. Therefore, it can be said that users with a lower value of ρ_i have a higher priority than users with a higher value of ρ_i. Furthermore, it has been proven analytically in [23] that M-LWDF is a throughput optimal algorithm implying that it is able to guarantee the QoS requirements of various traffic classes as well as optimize system throughput. Nevertheless, it is concluded in [24] that M-LWDF is an unfair scheduling scheme where the users with poor average radio propagation conditions suffer from higher delays than the remaining users in the cell and are not able to fulfill the QoS criterion during high load situations.

Exponential Rule

The Exponential Rule (ER) scheduler [6] is a modified version of the PF scheme which has been customized for scheduling RT applications. The ER algorithm prioritizes users based on the following formula:

$$SF_i(t) = a_i \cdot \frac{R_i(t)}{S_i(t)} \cdot \exp\left(\frac{a_i Q_i(t) - \overline{aQ(t)}}{1 + \sqrt{\overline{aQ(t)}}} \right) \qquad where \quad \overline{aQ(t)} = \frac{1}{n} \sum_{i=1}^{n} a_i Q_i(t)$$

where a_i is the priority value used to characterize the desired QoS and n is the total number of users. In order to better understand the ER scheme, let us consider the exponent term. According to the exponent, if the difference between a users' prioritized delay and the average prioritized delay of all users is greater than $\sqrt{\overline{aQ(t)}}$, then the exponent term will become dominant and exceed the impact of channel variations. On the other hand, for small differences in the prioritized delay, the exponent term will approach 1. Hence, this algorithm tries to balance the prioritized delay of all users when the difference between them becomes significant.

Queue-Based Exponential Rule Scheduler

In [25], Wang et al. propose a modified version of the ER scheduler known as the Queue-Based Exponential Rule (QBER) scheduler. The QBER considers the users' queue length in the scheduling decision apart from the delay experienced by their head-of-line packet. The QBER prioritizes users based on the following formula:

$$SF_i(t) = a_i \cdot \frac{R_i(t)}{S_i(t)} \cdot \exp\left(\frac{a_i Q_i(t) - \overline{aQ(t)}}{1 + \sqrt{\overline{aQ(t)}}} \right) \cdot \exp\left(\frac{q_i(t) - \overline{q(t)}}{1 + \overline{q(t)}} \right)$$

$$where \quad \overline{aQ(t)} = \frac{1}{n} \sum_{i=1}^{n} a_i Q_i(t)$$

$$and \quad \overline{q(t)} = \frac{1}{n} \sum_{i=1}^{n} q_i(t)$$

where $q_i(t)$ is the queue length of user i at the beginning of the t^{th} time interval. The second exponent term in the above equation is used to balance the service queue length among multiple users as it gives a higher priority to users that have a high variance in queue length when compared to the average value. The authors argue that the magnitude of the second exponent term does not exceed that of the first exponent term as the denominator of the second exponent term does not comprise of a square root term. Simulation results in [25] show that this scheme achieves a higher fairness index when compared to the ER scheme at the cost of slightly degraded performance in terms of throughput and queuing delay.

Modified Proportional Fairness

The Modified Proportional Fairness (MPF) [26] scheme is an adaptation of the PF algorithm for RT traffic. The MPF assigns a scheduling priority to users in the following manner:

$$SF_i(t) = \begin{cases} \dfrac{R_i(t)}{S_i(t)} & when \quad Q_i < \tau \\ \dfrac{R_i(t)}{S_i(t)} \cdot \left(\dfrac{\max_j \left\{\overline{R_j(t)}\right\}}{\overline{R_i(t)}} \right) & when \quad Q_i(t) \geq \tau \end{cases}$$

where $\max_j \left\{\overline{R_j(t)}\right\}$ is the maximum average supportable data rate among all users at a time interval t, $\overline{R_i(t)}$ is the average supportable data rate of user i and τ is a predefined threshold delay. The MPF algorithm prioritizes users according to the PF algorithm when their queuing delay is below a certain threshold value otherwise the priority is computed according to the FFT scheme. This scheme gives a higher priority to users whose queuing delay is close to their deadline value in order to prevent packet dropping. Moreover, the scheme assigns the HS-DSCH in a fair manner as it increases the priority of those users that have low average throughputs.

Scheduling Algorithm	Scheduling Factors	Supported Applications	Scheduling criteria
Round Robin	None	NRT	Cyclic order
Max Carrier to Interference Ratio [3]	Channel Quality Indicator	NRT	Highest channel quality
Proportional Fairness [4]	Supportable data rate and average user throughput	NRT	Highest relative channel quality
Data Rate Control [19]	Supportable data rate, average user throughput and constant control parameter	NRT	Highest relative channel condition with a constant preference given to the instantaneous channel condition
Adaptive Proportional Fairness [20]	Supportable data rate, average user throughput and dynamic control parameter	NRT	Highest relative channel condition with a dynamic preference given to the instantaneous channel condition
Fast Fair Throughput [21]	Supportable data rate, average user throughput and average throughput of all users in the cell	NRT	Highest relative channel condition with preference given to users with relatively poor channel conditions
Weighted Fair Scheduling [22]	Supportable data rate, average user throughput and average throughput of all users in the cell	NRT	Highest product of relative channel condition and the inverse of a users' fairness index
Throughput Guarantee Scheduling [22]	Supportable data rate, average throughput, average throughput of all users in the cell and minimum required throughput	NRT	Highest relative channel condition if users' throughput is greater than minimum requirement otherwise priority is added in order to guarantee QoS requirements
Modified Largest Weighted Delay First [5]	Supportable data rate, average user throughput, queuing delay of head packet, maximum delay threshold and QoS parameter	RT	Highest relative queuing delay and channel condition with QoS prioritization
Exponential Rule [6]	Supportable data rate, average user throughput, queuing delay of head packet and QoS parameter	RT	Highest relative channel condition or prioritized delay depending on the average prioritized delay of all users
Queue-Based Exponential Rule [25]	Supportable data rate, average user throughput, queuing delay of head packet, QoS parameter and queue length	RT	Highest relative channel condition, prioritized delay or queue length depending on the average prioritized delay or average queue length of all users
Modified Proportional Fairness [26]	Instantaneous supportable data rate, average user throughput, Average supportable data rate, maximum supportable data rate among all users in the cell	RT	Prioritize users based on the PF scheme if the queuing delay is below a certain threshold value otherwise priority is calculated according to the FFT scheme

Table 2.2: Comparison between the various scheduling schemes for HSDPA

2.5 Summary

Wireless communication systems have been evolving since the end of the 19th century when electromagnetic waves were first discovered as a communication medium. The development of communication systems went hand-in-hand with the growth of the semiconductor industry. The second generation of wireless communication systems introduced the most widely used and successful wireless cellular system to date: GSM. Universal mobility led to the development of a global roaming standard, and hence UMTS was introduced. With the growth of multimedia applications on the Internet, there was a need to improve data rates as well as system capacity, which resulted in the development of HSDPA. HSDPA can support data rates of up to 14.4 Mbps as opposed to 2 Mbps that was supportable by UMTS. Four QoS classes have been defined which are distinguished based on the delay sensitivity of the various traffic types. Through the implementation of a new shared transport channel known as the HS-DSCH, HSDPA aims at achieving lower delays and improving the overall system capacity. Furthermore, various new features have been added to HSDPA which include: Adaptive Modulation and Coding (AMC), Hybrid Automatic Repeat Request (HARQ) and channel dependent scheduling. HSDPA has shifted the packet scheduler from the RNC (in the case of UMTS) to the Node B in a MAC sub-layer known as the MAC-hs. With this implementation, the network was converted from a central to a distributed scheduling system which improved the overall system throughput. HSDPA is based on the use of a time shared resource and the assignment of this resource to a particular user at a given time interval is a non-trivial decision; therefore packet scheduling is an important implementation. HSDPA uses the short-term variations in channel conditions for the scheduling decision where the packet scheduler can maximize the overall system throughput by selecting users with favorable radio channel conditions. However, this raises the issue of fairness, as users with bad channel conditions might not get access to radio channels. Hence, a good scheduling scheme should not only try to maximize system throughput but also distribute the system resources in a fair manner. Moreover, RT applications have strict delay constraints and therefore a scheduling scheme for these applications should be able to meet their QoS requirements. In this chapter, UMTS was discussed and the changes that were made for the development of HSDPA were highlighted. New features that were required for the implementation of HSDPA were also discussed. Finally, various NRT and RT scheduling schemes were presented and compared.

Chapter 3

Delay Based Scheduler

In Chapter 2, the HSDPA network architecture was outlined as well as the features required for the conversion of UMTS to HSDPA were presented. Moreover, it was realized that the packet scheduling decision for the HSDPA system is a crucial one. The HSDPA packet scheduler utilizes the short-term radio channel conditions of users in order to prioritize them for scheduling. The shared channel resource is assigned to users with good channel conditions with the aim of maximizing overall system throughput and capacity. However, it was remarked that scheduling users with relatively good channel conditions leads to unfairness as it results in the starvation of those users with poor channel conditions. Therefore, there is a trade-off between system throughput and fairness which is identified as the major system constraint. Moreover, RT traffic applications are highly delay sensitive and impose strict delay restrictions in order to meet their QoS requirements. Hence, the design objective of a RT packet scheduling scheme is to maintain delay requirements for the various traffic classes as well as provide a good balance between system throughput and fairness. In this chapter, a RT traffic scheduler that we call the Delay Based Scheduler (DBS) is proposed. The DBS utilizes the channel conditions of users as well as the delay constraints of their RT traffic applications to determine their scheduling priority. Furthermore, the DBS algorithm converges to the Max CIR scheduling scheme with the adjustment of certain variables, thus allowing the prioritization of NRT traffic classes based on the instantaneous channel conditions of the users. Therefore, the DBS scheme can be adapted to utilize the system resources depending on the traffic environment of the cell as it can support RT as well as NRT services.

The remaining of this chapter is organized as follows: Section 3.1 outlines the problem definition. The scheduler model is discussed in Section 3.2. The DBS algorithm is presented in Section 3.3. Section 3.4 outlines the effect of the priority constant on the scheduling algorithm. Section 3.5 presents special cases where the DBS converges to the Max CIR scheme. Finally, the chapter is summarized in Section 3.6

3.1 Problem Definition

As discussed in Chapter 2, HSDPA achieves peak data rates of up to 14.4 Mbps. Due to its high supportable data rates and system efficiency various packet based RT as well as NRT services have emerged on this cellular network. The packet scheduler strives at maximizing system throughput by assigning users with good channel conditions to the shared channel. Therefore, the Max CIR algorithm was one of the first scheduling schemes proposed for HSDPA. However, it was realized that prioritizing users based on their channel conditions would result in the unfair distribution of system resources. Furthermore, RT applications require the transfer of data packets within a maximum tolerable delay. Based on this discussion and the factors that were outlined in Section 1.1, the scheduling problem definition may be defined as follows:

Given the channel conditions of a set of users sharing a common channel, then the scheduling problem is the assignment of these users on the shared channel so that the following objectives are met:

- *Throughput Maximization: The number of bits that can be transmitted to a user at a particular time interval is dependent on their channel condition. The better the channel condition, the more number of information bits can be transferred. Hence scheduling users with good channel conditions can maximize system throughput.*
- *Minimization of Queuing Delay: If the scheduling algorithm considers the queuing delay of a flow in order to prioritize users for assignment to the shared channel, then the average queuing delay experienced by the system can be minimized. Moreover, the minimization of queuing delay by a scheduling scheme is advantageous for RT traffic applications as they are highly delay sensitive.*
- *Fairness: Fairness is an integral part of the scheduling decision as users with the same service on a wireless cellular network expect similar data rates regardless of their channel conditions. Moreover, users with good channel conditions constitute a small percentage of the total cellular users and catering only these*

users, in order to maximize system throughput, will result in users leaving the service provider due to unsatisfactory service which will consequently lower the total revenue generated.

- *Maintaining Delay Thresholds: RT traffic applications are highly delay sensitive and require that their delay thresholds are maintained. This can be done by dropping packets when they exceed their delay threshold.*
- *QoS Prioritization: Various services are available simultaneously on a cellular network. These services are distinguished based on the requirements of users, for instance if a user transfers data packets in the range of 50 to 100 Mb on the cellular network then they would need a service that caters to users with high transfer requirements whereas occasional users would choose a service with lower transfer thresholds. Offered services can be prioritized based on the QoS requirements of various users. Therefore, a scheduling scheme should consider the prioritization of users based on their QoS requirements for the scheduling assignment.*
- *Scheduling RT and NRT traffic classes simultaneously: A scheduling scheme should be able to support RT as well as NRT applications at the same time.*

The scheduling algorithm that is proposed in this thesis tries to achieve the above mentioned design objectives. This will be discussed in detail in the forthcoming sections.

3.2 Scheduler Model

The scheduler model is defined as follows. It is assumed that the system consists of *n* users and the Node B is responsible for assigning the shared channel resource (i.e. the HS-DSCH) to one user at every Transmission Time Interval (TTI) of 2 ms. UEs can multi-task[1] several data transfer applications simultaneously. Therefore, each UE can consist of several data flows. However, it is assumed in this thesis; without the loss of generality; that each UE comprises of a single data flow at any time interval such that the RNC and Node B have one queue per user. Moreover, a Call Admission Control (CAC) mechanism is required to determine the level of acceptable traffic load in a cell such that the minimum service requirements of users are guaranteed. The CAC mechanism is an auxiliary research topic and the implementation of this mechanism is outside the scope of this thesis.

The flow of data packets within the UTRAN is as follows. The RNC receives packets and segments them into fixed Protocol Data Units (PDUs). The flow of these PDUs from the RNC to the Node B is governed by the flow control mechanism which was discussed in Chapter 2. One or several PDUs stored at the Node B are combined to form a transport block (or MAC-hs PDU) whose size depends on the channel quality of a user. The Node B then assigns the HS-DSCH resource to a single user at every TTI based on a packet scheduling scheme. Furthermore, a retransmission buffer exists at the RNC which stores PDUs in the order that they were transmitted to the Node B. PDUs stored in the

[1] Multi-tasking is the method by which multiple tasks or processes share a common processing resource in such a manner that this resource is frequently switched from one task to another so that the illusion of parallelism is achieved.

retransmission buffer are sent if the original PDUs are received incorrectly by the UE. The HARQ procedure is responsible for these retransmissions. Figure 3.1 illustrates the system model.

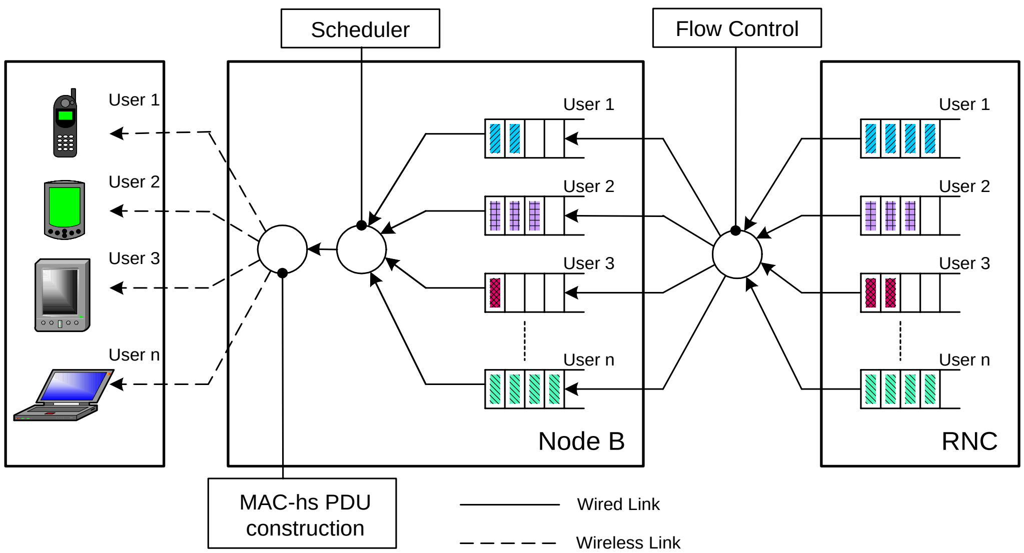


Figure 3.1: The System Model

3.3 Scheduling Algorithm

In this section, the Delay Based Scheduler (DBS) is proposed and its formulation is discussed so that all of the above mentioned design objectives are accounted for. DBS is presented in Section 3.3.1. Section 3.3.2 discusses how the proposed scheduling scheme satisfies the design objectives.

3.3.1 Delay Based Scheduler

The DBS is a packet based scheduling algorithm implemented for HSDPA. It can support RT traffic applications as it tries to maintain the delay thresholds of various traffic classes by minimizing their packet queuing delay. DBS is located at the MAC-hs sub-layer of the Node B in conformance with the 3GPP specifications for HSDPA.

3.3.1.1 DBS Notations

Before presenting the proposed algorithm, it is imperative that the following notations used for its formulation are defined:

- *i* represents the user index. Since it is assumed that the system consists of a maximum of *n* users at any time interval then $i \in \{1,2,...,n\}$.
- *t* represents the transmission time interval or TTI.
- *c* defines the traffic class where $c \in \{1,2,...,m\}$ and at any time interval, there are a total of *m* traffic classes.
- $SF_i(t)$ is the value of the scheduling function and hence represents the priority assigned to user *i* at time *t*. The user with the highest $SF_i(t)$ value is scheduled at every TTI.
- $CQI_i(t)$ defines the channel condition or CQI of user *i* at time *t*.
- $\max\_CQI(t)$ is the maximum value of CQI experienced among all users at time *t*.

- $Norm\_CQI_i(t)$ is the normalized channel quality of a user *i* at time interval *t*. Its calculation is given in Section 3.3.1.2.
- $Q_i(t)$ is the queuing delay experienced at Node B by the head-of-line packet of the data flow of user *i* up to transmission time interval *t*.
- $\max\_delay_c$ is the maximum tolerable delay associated with each traffic class *c*.
- $th1\_delay_c$ and $th2\_delay_c$ are two intermediate threshold delays that are assigned for each traffic class *c*. These two intermediate delays depend on $\max\_delay_c$ and their derivation is given in the next Section.
- E_i is the priority value assigned to each user *i* based on their QoS requirements.

3.3.1.2 DBS Formulation

The HSDPA cellular system promises a peak data rate of up to 14.4 Mbps. This has been made possible due to the implementation of the packet scheduler at a MAC sub-layer of the Node B known as the MAC-hs. As remarked previously, the implementation of the packet scheduler at Node B converted the central scheduling system to a distributed one thus increasing overall system throughput and capacity. Moreover, HSDPA utilizes the instantaneous channel conditions of users for the scheduling decision. These channel conditions along with the TBS and the number of supportable channelisation codes are transmitted to the Node B by every UE in the system via the CQI (see Section 2.3.1 for details). As the channel condition information of every user is available at the Node B, it can make the fastest scheduling decision and reduce overall system delay. In addition, users with relatively good channel conditions can download a higher number of bits per second when compared to users with poor channel quality. Therefore, the system throughput can be further increased by prioritizing users with good channel conditions. DBS uses a normalized value of the channel condition of a user for the scheduling decision, which is calculated as follows:

$$Norm\_CQI_i(t) = \frac{CQI_i(t)}{\max\_CQI(t)} \qquad (3.1)$$

where $CQI_i(t)$ is a measure of the channel condition of user *i* at time *t* and $\max\_CQI(t)$ is the maximum value of CQI experienced among all users at the t^{th} time interval. As can be seen in Equation (3.1), $Norm\_CQI_i(t)$ is the ratio of a users' channel condition to the best channel quality experienced among all users of the cell. Therefore, $Norm\_CQI_i(t)$ is a shot-term relative measure of a users' channel condition. It should be noted that at any given time interval the normalized channel condition of user *i* is governed by the following constraint: $0 < Norm\_CQI_i(t) \leq 1$. By utilizing the channel condition information in the scheduling decision, the DBS attempts to maximize system throughput.

In Chapter 2, it was discussed that RT traffic applications have strict delay constraints. Unlike NRT services, RT applications can tolerate a certain error rate in order to meet these delay thresholds. For instance, in a VoIP application users can tolerate a small amount of packet loss as they can infer a conversation if some words of a sentence are unclear. However, if the response time exceeds a certain tolerable amount, then the call is rendered useless. Delay threshold values are dependent on the type of application, for example conversational traffic applications are more delay sensitive than streaming applications. DBS considers the delay constraints of RT applications and hence defines a maximum threshold delay for each traffic class *c* which is know as $\max\_delay_c$. In order to maintain the threshold delay constraints, head packets with queuing delay, $Q_i(t)$, greater than the maximum tolerable delay are dropped. Therefore, the priority of a user, $SF_i(t)$, can be calculated as follows:

$$SF_i(t) = \begin{cases} Norm\_CQI_i(t) & If \quad Q_i(t) < \max\_delay_c & (3.2) \\ Packet \quad Dropped & If \quad Q_i(t) \geq \max\_delay_c & (3.3) \end{cases}$$

At every transmission time interval, the user with the highest $SF_i(t)$ value is scheduled. According to Equations (3.2) and (3.3), if the value of $Q_i(t)$ is below the maximum tolerable delay of the traffic class of user *i* then the user is prioritized based on their relative channel condition. On the other hand, if the queuing delay exceeds the maximum threshold delay then the packet is dropped and the priority of a user is calculated for the next packet in the queue. Traffic prioritization based on the above mentioned equations is illustrated in Figure 3.2, for a maximum threshold delay value of 2 s.

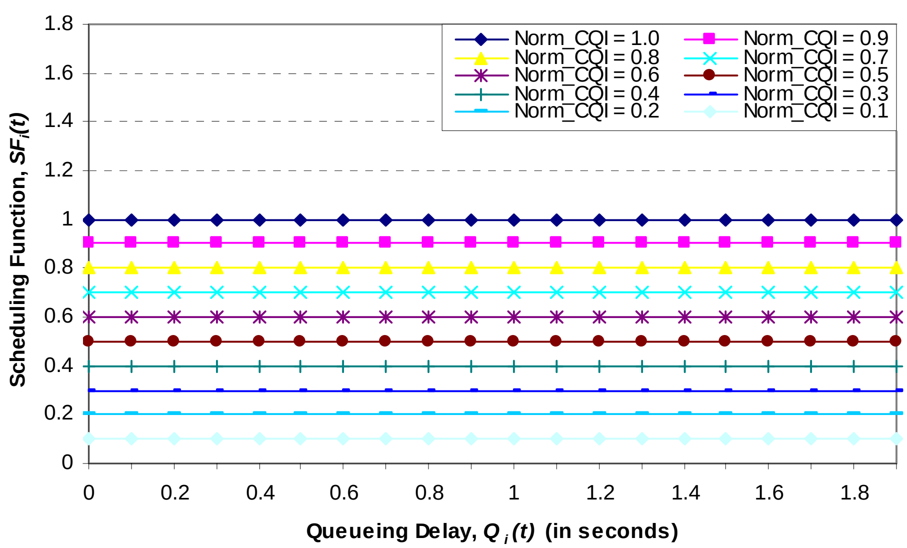


Figure 3.2: The effect of channel condition on traffic prioritization without delay considerations for *max_delay*$_c$ = 2 s

The scheduling function depicted in Figure 3.2 has been graphed for various values of $Norm\_CQI_i(t)$ representing the priority of users with different channel conditions. If it is assumed that the users have a relatively constant channel condition in the short time duration of 2 s then it can be depicted (from the figure) that their scheduling priority is constant as their queuing delay increases. In other words, Equations (3.1) and (3.2) try to maintain the threshold delay of a traffic class by dropping packets when they exceed their maximum delay requirements and prioritize users based on their relative channel conditions without considering the instantaneous queuing delay experienced by the packet. Although the scheduling function described in the above mentioned equations achieves the objective of maintaining the delay constraints of RT traffic, it has two problems: Firstly, it fails to minimize delay as it does not prioritize users based on the instantaneous queuing delay experienced by their data flows. Secondly, this scheduling function may lead to an excessive amount of packet loss as users with poor channel conditions might not be served before their delay term expires.

In order to overcome the shortcomings of the above mentioned equations, a delay factor was added to the scheduling function so that the aggregate queuing delay of the system is minimized Before presenting the modified scheduling function, two more delay thresholds are introduced viz. $th1\_delay_c$ and $th2\_delay_c$. These intermediate delay thresholds can be derived in the following manner:

$$th1\_delay_c = \frac{\max\_delay_c}{\gamma_c} \tag{3.4}$$

$$th2\_delay_c = \frac{1}{2}\left(\max\_delay_c + th1\_delay_c\right) \tag{3.5}$$

where γ_c is a constant defined for every traffic class such that $\gamma_c \geq 1$, $th1\_delay_c$ is a fraction of the maximum threshold delay and $th2\_delay_c$ is the median between the maximum and first threshold delay values. Thus, the scheduling function defined by Equations (3.2) and (3.3) can be redefined as follows:

$$SF_i(t) = \begin{cases} Norm\_CQI_i(t) & If \quad Q_i(t) < th1\_delay_c \\ Norm\_CQI_i(t) + \left(\dfrac{Q_i(t) - th1\_delay_c}{\max\_delay_c - th1\_delay_c} \right) & If \quad th1\_delay_c \leq Q_i(t) < th2\_delay_c \\ Norm\_CQI_i(t) + \left(\dfrac{Q_i(t) - th1\_delay_c}{\max\_delay_c - th2\_delay_c} \right) & If \quad th2\_delay_c \leq Q_i(t) < \max\_delay_c \\ Packet \;\; Dropped & If \quad Q_i(t) \geq \max\_delay_c \end{cases} \tag{3.6}$$

According to the function shown in Equation (3.6), a user is prioritized based on their relative channel condition as well as the queuing delay experienced by their head packet. This function assigns different scheduling values to users depending upon their queuing delay. As discussed above, the maximum threshold delay of a class is divided into three parts, based on these divisions three cases have been defined: $Q_i(t) < th1\_delay_c$, $th1\_delay_c \leq Q_i(t) < th2\_delay_c$, $th2\_delay_c \leq Q_i(t) < \max\_delay_c$. If the queuing delay experienced by the head-of-line packet of a data flow is less than $th1\_delay_c$ then the priority of that user depends on their relative channel condition only. However, if the queuing delay experienced by a data flow is in between $th1\_delay_c$ and $th2\_delay_c$ then the scheduling function is a sum of the relative channel condition of that user and a linear function which is expressed by the users' queuing delay and the threshold delays defined for its traffic class: $(Q_i(t) - th1\_delay_c)/(\max\_delay_c - th1\_delay_c)$. Note that this linear function has a constant slope which is expressed by:

$1/\left(\max\_delay_c - th1\_delay_c\right)$ and the value of the function increases as the difference between $Q_i(t)$ and $th1\_delay_c$ increases. Consecutively, if the queuing delay experienced by a user is between $th2\_delay_c$ and $\max\_delay_c$ then the slope of the linear function increases such that it is represented as: $1/\left(\max\_delay_c - th2\_delay_c\right)$, thus user *i* is prioritized based on its normalized channel condition, $Norm\_CQI_i(t)$, and the linear function: $\left(Q_i(t) - th1\_delay_c\right)/\left(\max\_delay_c - th2\_delay_c\right)$. The purpose of increasing the slope of the above mentioned linear function was to increase the priority of a user as its queuing delay approaches the maximum threshold delay. This discussion can be described as follows:

$$
\begin{aligned}
&If \quad th1\_delay_c \le Q_i(t) < th2\_delay_c \\
&then \quad SF_i(t) = Norm\_CQI_i(t) + \left(\frac{Q_i(t) - th1\_delay_c}{\max\_delay_c - th1\_delay_c}\right) \\
&\qquad\qquad = Norm\_CQI_i(t) + \underbrace{\left(\frac{1}{\max\_delay_c - th1\_delay_c}\right)\cdot\left(Q_i(t) - th1\_delay_c\right)}_{Linear \quad Expression \quad (3.7a)}
\end{aligned}
$$

$$
\begin{aligned}
&Else \quad If \quad th2\_delay_c \le Q_i(t) < \max\_delay_c \\
&then \quad SF_i(t) = Norm\_CQI_i(t) + \left(\frac{Q_i(t) - th1\_delay_c}{\max\_delay_c - th2\_delay_c}\right) \\
&\qquad\qquad = Norm\_CQI_i(t) + \underbrace{\left(\frac{1}{\max\_delay_c - th2\_delay_c}\right)\cdot\left(Q_i(t) - th1\_delay_c\right)}_{Linear \quad Expression \quad (3.7b)}
\end{aligned}
$$

The slope of expression (3.7a) can be derived as

$$
\Rightarrow \frac{1}{\max\_delay_c - th1\_delay_c}
$$

$$
\Rightarrow \frac{1}{\max\_delay_c - \dfrac{\max\_delay_c}{\gamma_c}} \qquad\qquad \left\{\because \quad th1\_delay_c = \frac{\max\_delay_c}{\gamma_c} \quad by \quad (3.4)\right\}
$$

$$\Rightarrow \frac{\gamma_c}{\gamma_c \cdot \max\_delay_c - \max\_delay_c}$$

$$\Rightarrow \frac{\gamma_c}{\max\_delay_c(\gamma_c - 1)} \qquad (3.8)$$

Similarly, the slope of expression (3.7b) is derived as

$$\Rightarrow \frac{1}{\max\_delay_c - th2\_delay_c}$$

$$\Rightarrow \frac{1}{\max\_delay_c - \frac{1}{2}\cdot(\max\_delay_c + th1\_delay_c)}$$

$$\left\{\because \quad th2\_delay_c = \frac{1}{2}\cdot(\max\_delay_c + th1\_delay_c) \quad by \quad (3.5)\right\}$$

$$\Rightarrow \frac{1}{\max\_delay_c - \frac{1}{2}\cdot\left(\max\_delay_c + \frac{\max\_delay_c}{\gamma_c}\right)} \qquad \{by \quad (3.4)\}$$

$$\Rightarrow \frac{1}{\max\_delay_c - \frac{\max\_delay_c}{2} - \frac{\max\_delay_c}{2\cdot\gamma_c}}$$

$$\Rightarrow \frac{2\cdot\gamma_c}{2\cdot\gamma_c\cdot\max\_delay_c - \gamma_c\cdot\max\_delay_c - \max\_delay_c}$$

$$\Rightarrow \frac{2\cdot\gamma_c}{\max\_delay_c\cdot(2\cdot\gamma_c - \gamma_c - 1)}$$

$$\Rightarrow \frac{2\cdot\gamma_c}{\max\_delay_c\cdot(\gamma_c - 1)} \qquad (3.9)$$

Therefore, the slope of expression (3.7b) is greater than the slope of expression (3.7a)

$$\because \frac{2\cdot\gamma_c}{\max\_delay_c\cdot(\gamma_c - 1)} > \frac{\gamma_c}{\max\_delay_c\cdot(\gamma_c - 1)} \qquad \{by \quad (3.8) \quad and \quad (3.9)\}$$

Hence, this shows that the assignment of prioritization of a user (as shown in Equation (3.6)) changes as their queuing delay transitions from one intermediate delay to the next until it approaches the maximum threshold defined for their traffic class. This is done by increasing the slope of the linear function that is used to utilize the queuing delay information of a user in the calculation of their scheduling function.

Most RT scheduling schemes try to minimize the average queuing delay of the system while considering the users' channel conditions in the scheduling assignment, in an attempt to balance the tradeoff between throughput maximization and the minimization of queuing delay. However, these schemes cannot adjust the desired level of system throughput or queuing delay. The introduction of the intermediate threshold delays in Equation (3.6) has enabled the adjustment of the desired system performance through the attunement of $th1\_delay_c$ or specifically γ_c. According to Equation (3.4), smaller values of γ_c will allow $th1\_delay_c$ to be closer to $\max\_delay_c$. In other words, if $\gamma_c \to 1$ then $th1\_delay_c \to \max\_delay_c$. This will allow the first delay interval, specified by: $Q_i(t) < th1\_delay_c$ to be longer which in turn increases the period during which the scheduling function depends only on the channel condition information of a user. As mentioned previously, users with good channel conditions can receive a higher number of bits per second and, thus, scheduling these users will improve system throughput. On the other hand, larger values of γ_c will make the first threshold delay smaller i.e. if $\gamma_c \to \infty$ then $th1\_delay_c \to 0$ (from Equation (3.4)). Smaller values of $th1\_delay_c$ and consequently $th2\_delay_c$ (according to Equation (3.5)) will result in the second and third delay intervals specified by: $th1\_delay_c \le Q_i(t) < th2\_delay_c$ and

$th1\_delay_c \le Q_i(t) < th2\_delay_c$ respectively, to be longer. This increases the time duration during which the scheduling function depends on the users' queuing delay as well as their channel condition. However, during the second and third delay intervals, the dependence of the scheduling function on the flows' queuing delay increases and its dependence on the users' channel condition information decreases as the packet approaches its deadline. Hence, longer second and third delay intervals allow the scheduling function to give a higher priority to users with higher queuing delay, thus minimizing the average queuing delay of the system. Therefore, γ_c can be attuned to either maximize system throughput or minimize the average queuing delay of the system. Traffic prioritization based on the function described in Equation (3.6) is illustrated in Figure 3.3 where $\max\_delay_c = 2$ s, $\gamma_c = 2$, $th1\_delay_c = 1$ s and $th2\_delay_c = 1.5$ s.

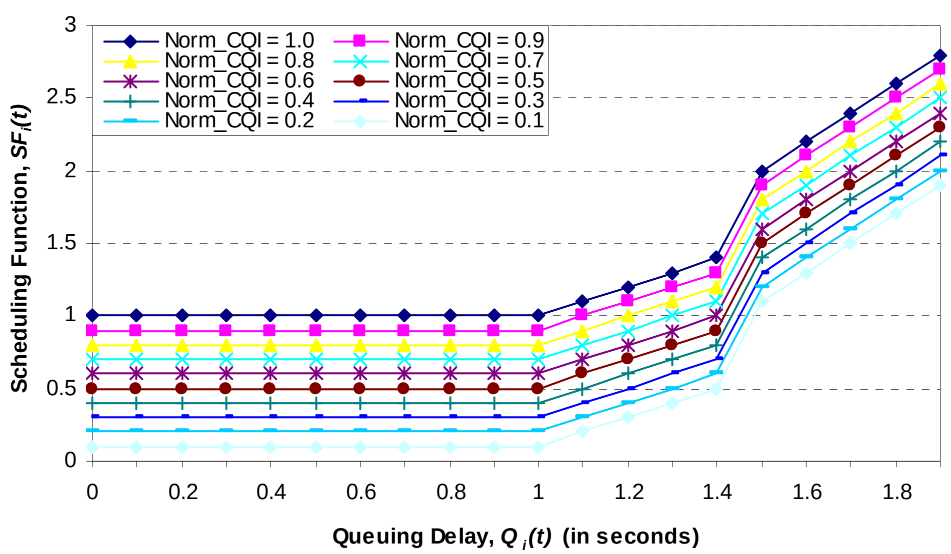


Figure 3.3: Effect of Queuing Delay on the Scheduling Function for various channel conditions

Similar to Figure 3.2, the scheduling function shown in Figure 3.3 has been plotted for various values of $Norm\_CQI_i(t)$. It can be depicted from the figure, that as the duration of a packets stay at the Node B increases, its priority also increases. Therefore, the scheduling function described in Equation (3.6) prioritizes users based on their relative channel condition as well as their instantaneous queuing delay at the Node B buffer. Moreover, this function is capable of maintaining the delay thresholds of RT traffic applications by dropping packets once they exceed their maximum threshold delay. However, under the assumption that a users' $Norm\_CQI_i(t)$ value fluctuates in such a way that its average value stays relatively constant for a short time duration, then the above scheme distributes resources in an unfair manner as packets of users with poor channel conditions might not be served before they are dropped. This may also eventually lead to an excessive amount of packet loss.

Fairness is introduced in the scheme described in Equation (3.6) in order to prevent starvation of users with relatively poor channel conditions and consequently reduces the amount of packet loss. This fairness is implemented in terms of the queuing delay experienced by a flow. The modified and fair version of the scheme in Equation (3.6) is:

$$SF_i(t)=\begin{cases} Norm\_CQI_i(t) & If \quad Q_i(t)<th1\_delay_c \\ \left(Norm\_CQI_i(t)\right)^{\alpha}+\left(\dfrac{Q_i(t)-th1\_delay_c}{\max\_delay_c-th1\_delay_c}\right)^{\beta} & If \quad th1\_delay_c\le Q_i(t)<th2\_delay_c \\ \left(Norm\_CQI_i(t)\right)^{\alpha}+\left(\dfrac{Q_i(t)-th1\_delay_c}{\max\_delay_c-th2\_delay_c}\right)^{\beta} & If \quad th2\_delay_c\le Q_i(t)<\max\_delay_c \\ Packet\ \ Dropped & If \quad Q_i(t)\ge\max\_delay_c \end{cases} \tag{3.10}$$

$$Where \quad \beta=\frac{Q_i(t)}{\max\_delay_c} \tag{3.11a}$$

$$\alpha=1-\beta \tag{3.11b}$$

According to the modified scheme defined by Equation (3.10), users are prioritized in such a way that the dependence of the scheduling decision on their channel condition decreases as their queuing delay increases. This is because the value of β increases as a packets' stay at the Node B increases, thereby increasing the influence of queuing delay on the scheduling decision. Consequently, as β increases the value of α decreases which in-turn decreases the dependence of the channel condition on the scheduling function. This point is further clarified in Figure 3.4.

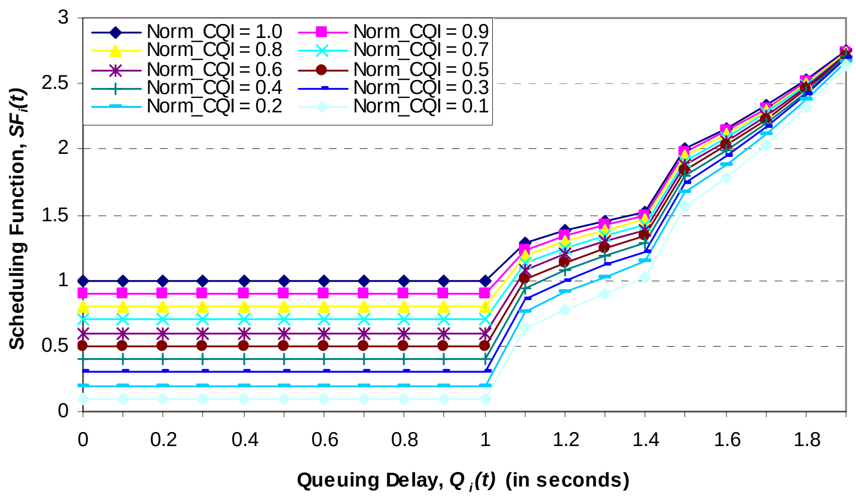


Figure 3.4: Effect of fairness on the scheduling function for various channel conditions

As in Figures 3.2 and 3.3, in Figure 3.4 it is assumed that a users' normalized channel condition varies in a way that its average value is relatively constant for a short time interval. The figure has been illustrated with the following values: $\max\_delay_c$ = 2 s, γ_c = 2, $th1\_delay_c$ = 1 s and $th2\_delay_c$ = 1.5 s. It can be depicted from the figure that (with the scheme obtained in Equation (3.10)) users with poor channel conditions can have a similar priority when compared to users with relatively good channel conditions as opposed to the scheme illustrated in Equation (3.6). This is because the scheme described in Equation (3.10) increases the dependence of the queuing delay experienced by a flow on the scheduling function and hence reduces the variance in the scheduling probability as a flows' queuing delay approaches its maximum value. Therefore, this scheme not only considers the channel condition of a user for prioritization but it increases their scheduling probability as their queuing delay increases. Hence, users with poor channel conditions have a greater chance of being scheduled as their $Q_i(t)$ value increases which addresses the fairness issue that previously discussed. Nonetheless, it should be noted that at any time interval, if the queuing delay experienced by two users is the same but their channel qualities are different then the user with the better channel condition has a higher chance of being scheduled. This property assures that the scheme maximizes system throughput while considering the delay requirements of various traffic classes. This scheme reduces the starvation of users with poor channel conditions and thus solves the problem of excessive packet loss that was seen in the previous scheme.

In the problem definition discussed in Section 3.1, QoS prioritization was identified as a design objective. In order to accommodate this objective, the scheme in Equation (3.10) was modified as follows:

$$SF_i(t)=\begin{cases} \left[Norm\_CQI_i(t)\right]\cdot E_i & \mathit{If} \quad Q_i(t)<th1\_delay_c \\ \left[\left(Norm\_CQI_i(t)\right)^{\alpha}+\left(\dfrac{Q_i(t)-th1\_delay_c}{\max\_delay_c-th1\_delay_c}\right)^{\beta}\right]\cdot E_i & \mathit{If} \quad th1\_delay_c\le Q_i(t)<th2\_delay_c \\ \left[\left(Norm\_CQI_i(t)\right)^{\alpha}+\left(\dfrac{Q_i(t)-th1\_delay_c}{\max\_delay_c-th2\_delay_c}\right)^{\beta}\right]\cdot E_i & \mathit{If} \quad th2\_delay_c\le Q_i(t)<\max\_delay_c \\ \mathit{Packet\ \ Dropped} & \mathit{If} \quad Q_i(t)\ge \max\_delay_c \end{cases} \tag{3.12}$$

where E_i is the priority assigned to every user i based on their QoS requirements. As discussed previously, various services are available simultaneously on a cellular network. These services can be distinguished based on the requirements of users. The scheme presented in Equation (3.12) prioritizes offered services based on the QoS requirements of various users. The effect of QoS prioritization on the scheduling function is shown in Figure 3.5 where two classes of service are distinguished based on their delay requirements. Class 1 has a lower delay tolerance than class 2 and thus has a higher priority. The parameters for class 1 are: $\max\_delay_1 = 2$ s, $\gamma_1 = 2$, $th1\_delay_1 = 1$ s, $th2\_delay_1 = 1.5$ s and $E_1 = 1$. The parameters for class 2 are defined as: $\max\_delay_2 = 4$ s, $\gamma_2 = 2$, $th1\_delay_2 = 2$ s, $th2\_delay_2 = 3$ s and $E_2 = 0.5$.

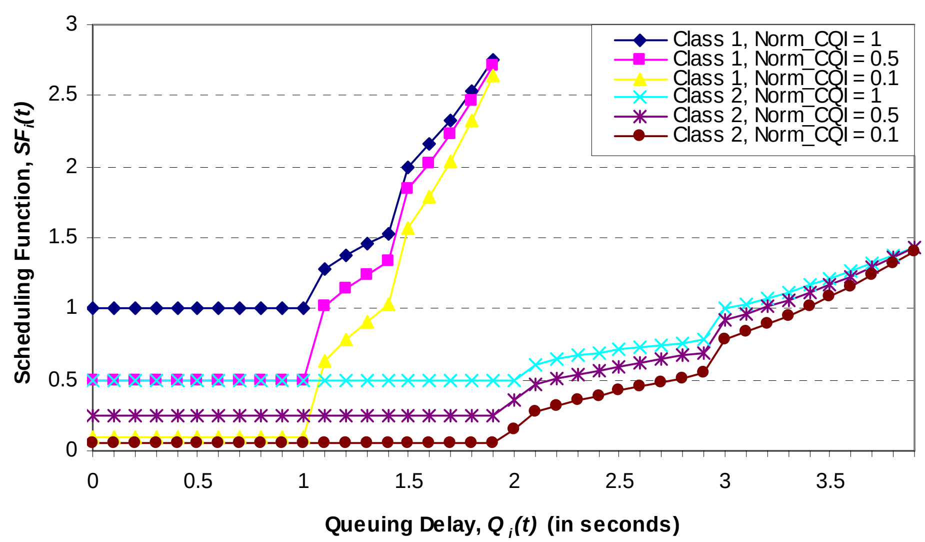


Figure 3.5: Effect of QoS Prioritization on the Scheduling Function

Figure 3.5 has been illustrated for users belonging to two distinct traffic classes with various values of $Norm\_CQI_i(t)$. It can be seen from the figure that users belonging to the traffic class with higher QoS requirements have a higher $SF_i(t)$ value over time.

Hence the scheduling scheme described in Equation (3.12) prioritizes users based on their relative channel condition, queuing delay and QoS requirements. Furthermore, this scheme improves system fairness as well maintains the delay thresholds of RT traffic applications. This scheme represents the scheduling function of the DBS.

3.3.1.3 DBS Algorithm

The Delay Based Scheduler (DBS) is a scheduling scheme designed for HSDPA. As discussed in Section 3.3.1.2, DBS aims at achieving maximum system throughput while distributing system resources in a fair manner so as to prevent the starvation of users with poor channel conditions. Moreover, this scheme can minimize the average queuing delay of the system, maintain the delay thresholds of various RT traffic applications as well as provide certain QoS prioritization. The scheduling function of the DBS is described as follows:

$$SF_i(t)=\begin{cases}[Norm\_CQI_i(t)]\cdot E_i & If \quad Q_i(t)<th1\_delay_c\\ \left[(Norm\_CQI_i(t))^{\alpha}+\left(\dfrac{Q_i(t)-th1\_delay_c}{\max\_delay_c-th1\_delay_c}\right)^{\beta}\right]\cdot E_i & If \quad th1\_delay_c\le Q_i(t)<th2\_delay_c\\ \left[(Norm\_CQI_i(t))^{\alpha}+\left(\dfrac{Q_i(t)-th1\_delay_c}{\max\_delay_c-th2\_delay_c}\right)^{\beta}\right]\cdot E_i & If \quad th2\_delay_c\le Q_i(t)<\max\_delay_c\\ Packet \;\; Dropped & If \quad Q_i(t)\ge\max\_delay_c\end{cases}$$

$$Where \;\; \beta=\frac{Q_i(t)}{\max\_delay_c}$$
$$\alpha=1-\beta$$

(3.13)

The notations used in Equation (3.13) are defined in Section 3.3.1.1. Moreover, $SF_i(t)$ represents the priority assigned to a user such that the user with the highest $SF_i(t)$ value is scheduled at every TTI. $Norm\_CQI_i(t)$ represents the normalized channel condition of a user and is calculated as follows:

$$Norm\_CQI_i(t) = \frac{CQI_i(t)}{\max\_CQI(t)}$$

The two intermediate delays: $th1\_delay_c$ and $th2\_delay_c$ are assigned to each traffic class *c* and are derived as follows:

$$th1\_delay_c = \frac{\max\_delay_c}{\gamma_c}$$

$$th2\_delay_c = \frac{1}{2}\left(\max\_delay_c + th1\_delay_c\right)$$

γ_c is a constant defined for every traffic class. Therefore, $th1\_delay_c$ is a fraction of the maximum threshold delay value and $th2\_delay_c$ is the median between the maximum and first threshold delay values. The above mentioned expression that governs the prioritization of a user based on the DBS scheme translates to the following rules:

- When the queuing delay of a data flow is less than $th1\_delay_c$ then the priority of that user is determined by their channel condition and QoS prioritization.
- When the value of $Q_i(t)$ is in between the intermediate threshold delays then the value of $SF_i(t)$ depends on the users' normalized channel quality, their relative queuing delay and their QoS prioritization assignment. It should be noted that the exponent terms α and β increase the dependence of the priority assignment on the queuing delay consequently inhibiting its dependence on the users' channel condition as a packets' stay at the Node B approaches its expiry term. This ensures that the overall system delay is minimized as well as guaranteeing that the HS-DSCH is assigned to users with relatively poor

channel conditions in a fair manner. Furthermore, two users with the same queuing parameters are scheduled depending on their channel condition thus ensuring that the system throughput is maximized.

- When $th2\_delay_c \le Q_i(t) < \max\_delay_c$ then a users' priority assignment is similar to the one explained in the previous point. However, with the exception of an increased scheduling probability depending on the queuing delay.
- When the queuing delay of a users' data flow exceeds its maximum threshold value then their head packet is dropped and the scheduling priority of that user is calculated for the next packet in the queue. This property maintains the delay thresholds of RT traffic applications at the cost of a certain error rate.

So far it has been shown that the DBS achieves the design objectives of a scheduling scheme for the HSDPA system that were outlined in Section 3.1 with the exception of scheduling RT and NRT traffic classes simultaneously, which will be discussed in Section 3.3.1.5. The DBS algorithm collects and stores the channel condition information of all users as well as the queuing delay of their flows' head packet in their corresponding transmission buffer at the Node B. It then determines the maximum channel quality among all users of the cell in order to calculate the relative channel condition of all users. Upon calculation of the scheduling function the users are sorted in descending order of scheduling priority. Finally, PDUs of the user with the highest priority are scheduled if the number of PDUs to be sent by that user exceeds zero. The DBS algorithm is illustrated by means of a flowchart which is shown in Figure 3.6.

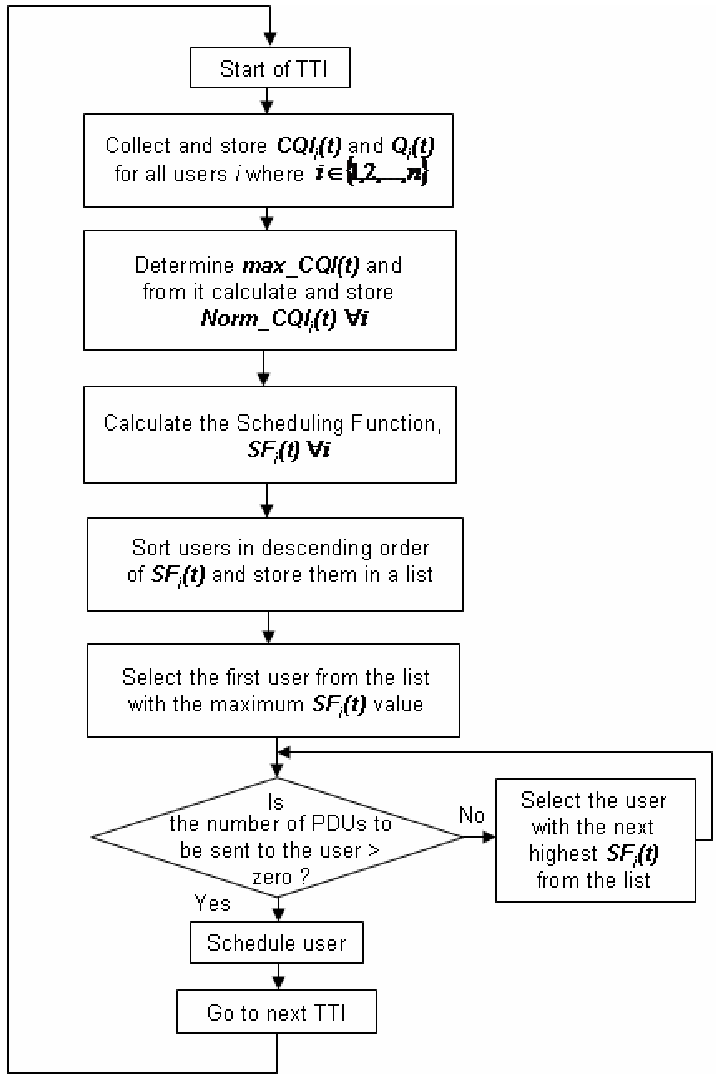


Figure 3.6: Overall Description of the DBS Algorithm

3.3.1.4 DBS Illustrations

In this Section, the scheduling of users based on the DBS scheme is described with the help of numerical illustrations. The DBS algorithm depends on various factors for the scheduling decision which include: relative channel condition, instantaneous queuing delay, QoS prioritization and delay thresholds of various traffic applications. Selection of these parameters can greatly vary system performance. The effects of delay thresholds as well as QoS prioritization on the scheduling decision and consequently on the system performance are discussed in Sections 3.3.1.4.1 and 3.3.1.4.2 respectively.

3.3.1.4.1 The Effect of Delay Thresholds on the Scheduling Decision

Two traffic classes are considered for the purpose of illustration (i.e. $m = 2$). The parameters defined for these classes are shown in Table 3.1. At the t^{th} time interval, three users are connected to the Node B viz. UE_1, UE_2 and UE_3 such that the former two users belong to traffic class 1 whereas UE_3 operates applications that are associated with traffic class 2. As mentioned previously, these users send their channel condition information to the Node B via the CQI at every TTI of 2 ms. The MAC-hs sub-layer of the Node B comprises of the HSDPA packet scheduler which is responsible for making the scheduling decision at every TTI. DBS utilizes the channel condition information to derive the $Norm\_CQI_i(t)$ values for all users which is the ratio of a users' instantaneous channel condition to the maximum channel quality experienced among all users at a given time interval, thus $0 < Norm\_CQI_i(t) \le 1$. Hence, at any time interval there will be one or more users that have a relative channel quality value of 1. Moreover, a single data flow is associated with every user at the Node B in such a way that packets are queued for this user in a first-in-first-out manner. Therefore, the instantaneous queuing delay, $Q_i(t)$, associated with each user, is the queuing delay experienced by the head packet of their flow at the Node B. It is assumed that all users have the same QoS priority such that

$E_i = 1 \; \forall i \in \{1,2,3\}$. Finally, the DBS uses these factors for the calculation of the priority of each user, $SF_i(t)$, according to the function presented in Equation (3.13).

Class	*max_delay$_c$ (in s)*	γ_c	*th1_delay$_c$ (s)*	*th2_delay$_c$ (s)*
1	2	2	1	1.5
2	4	2	2	3

Table 3.1: Definition of Class Parameters

The relative channel condition and queuing delay of each user for three time intervals (*t* to *t+4*) is shown in Table 3.2 and their $SF_i(t)$ value is calculated accordingly. It can be seen from the table that at time interval *t*, UE_1 has the best channel quality but the queuing delay experienced by UE_2 is higher and between its intermediate threshold delays which increases the dependence of the scheduling function on the delay experienced by the flow (as explained in Section 3.3.1.2). Therefore, the priority assigned to UE_2 is greater than the priority assigned to UE_1. This allows the DBS to minimize the average queuing delay of the system. However, if the Max CIR algorithm is used in this scenario then UE_1 would have been chosen for transmission in order to maximize system throughput. Nonetheless, it should be noted that RT applications have defined delay thresholds and minimizing queuing delay will ensure that packets are sent before the expiry of their delay term.

As UE_2 is assigned to the HS-DSCH at the t^{th} interval, it is assumed that the queuing delay of the next head-of-line packet is smaller than the previous one. At time interval *t+2*, a similar case is seen where UE_3 has the best channel conditions but UE_1 was scheduled due to its higher queuing delay value. Moreover, it can be seen that UE_2

experiences a higher queuing delay in this time frame but is not scheduled due to its poor channel quality. Thus, the DBS not only considers the delay experienced by users but also their channel quality in the scheduling assignment and tries to minimize queuing delay as long as the system throughput is not compromised greatly. In the third time interval, it can be seen that UE_1 and UE_3 have the same scheduling parameters (i.e. $Norm\_CQI_i(t)$ and $Q_i(t)$) but UE_1 was selected over the other because it has a higher relative queuing delay. In other words, the maximum tolerable delay of class 1 is lower than that of class 2 which allows these users to be scheduled earlier as their packet expiry deadline is closer. For this reason, it was seen that UE_3 was never selected for transmission although it had preferable channel conditions. It can be argued that this will reduce system throughput. But the channel condition of a user does not fluctuate by a significant amount in the short time durations that are defined for the delay thresholds of RT applications and thus has a sticky[1] property. In other words, if a user has good channel conditions then their aggregate value over a short time will remain the same. This property can be utilized to schedule users that have poor channel conditions so that they can meet their delay requirements and the compromised user (with high channel quality) can be served at a later time without considerably inhibiting system throughput.

It can be inferred from this illustration, that the DBS aims at minimizing overall queuing delay in order to meet the delay requirements of RT traffic applications. Furthermore, the instantaneous channel conditions of users are utilized in order to maximize system throughput even if the delay experienced by them is not the highest among all users.

[1] Sticky property is the ability to retain a certain condition even if the environment is constantly changing.

Conclusively, it can be said that there is a fine balance between throughput maximization and the minimization of queuing delay which can be tampered with the variable: γ_c. This will be shown through simulation results in Chapter 4.

TTI (in ms)	*t*	*t + 2*	*t + 4*
Norm_CQI$_1$(t)	1.0	0.95	1.0
Q$_1$(t)	1000	1002	1002
SF$_1$(t)	1.0	1.019	1.044
Norm_CQI$_2$(t)	0.5	0.45	0.3
Q$_2$(t)	1496	1108	1110
SF$_2$(t)	1.4316	0.9918	0.629
Norm_CQI$_3$(t)	0.95	1.0	1.0
Q$_3$(t)	998	1000	1002
SF$_3$(t)	0.95	1.0	1.0

Table 3.2: Calculation of the Scheduling Function at varying time intervals

3.3.1.4.2 The Effect of QoS Prioritization on the Scheduling Decision

The two traffic classes that were defined in the previous section will be used here along with the class assignment of the three users. The QoS prioritization is changed so that $E_1, E_3 = 1$ and $E_2 = 0.5$. Table 3.2 was modified to reflect these changes which are shown in Table 3.3. The QoS prioritization changed the scheduling assignment from *UE$_2$* to *UE$_1$* in the first interval. This is because the priority assignment of *UE$_2$* halved its scheduling function value although it was the most preferred transmission user based on the cumulative effect of its queuing delay and channel condition. Furthermore, the selection of *UE$_1$* in the t^{th} time interval resulted in lowering the queuing delay of this user in the next TTI which consequently reduced its $SF_i(t)$ value. This in-turn led to the selection of UE_3 at time interval $t+2$. Finally, UE_1 was scheduled in the third time interval. It should be noted that UE_2 is not selected for transmission in this scenario although it has the highest queuing delay. This is because the user has been given a much lower QoS priority compared to other users in the system. In other words, higher priority users are given preference over channel condition as well as queuing delay in the scheduling assignment of the DBS. Therefore, it can be concluded from this illustration that the QoS priorities assigned to users should be selected carefully as they can control the flow of traffic.

TTI (in ms)	***t***	***t + 2***	***t + 4***
$Norm\_CQI_1(t)$	1.0	0.95	1.0
$Q_1(t)$	1000	1000	1002
$SF_1(t)$	1.0	0.9747	1.044
$Norm\_CQI_2(t)$	0.5	0.45	0.3
$Q_2(t)$	1496	1498	1500
$SF_2(t)$	0.7158	0.7058	0.87
$Norm\_CQI_3(t)$	0.95	1.0	1.0
$Q_3(t)$	998	1000	992
$SF_3(t)$	0.95	1.0	1.0

Table 3.3: Calculation of the Scheduling Function with QoS prioritization

3.3.1.5 Special Cases

In this section, it is proved that the DBS converges to the Max CIR algorithm and a prioritized version of the scheme under special cases.

Lemma 1: DBS can converge to a prioritized version of the Max CIR algorithm

Proof:

If the value of the class parameter γ_c is set to 1 for all traffic classes $c \in \{1,2,...,m\}$, then the intermediate threshold delays, defined in Equations (3.4) and (3.5), can be modified as follows:

$$th1\_delay_c = \frac{\max\_delay_c}{\gamma_c} \qquad \{by \quad (3.4)\}$$

$$\Rightarrow \quad th1\_delay_c = \frac{\max\_delay_c}{1} \qquad \{\because \gamma_c = 1\}$$

$$\Rightarrow \quad th1\_delay_c = \max\_delay_c \tag{3.14}$$

Similarly,

$$th2\_delay_c = \frac{1}{2}\left(\max\_delay_c + th1\_delay_c\right) \qquad \{by \quad (3.5)\}$$

$$\Rightarrow \quad th2\_delay_c = \frac{1}{2}\left(2 \cdot \max\_delay_c\right) \qquad \{\because \quad th1\_delay_c = \max\_delay_c \quad by \quad (3.14)\}$$

$$\Rightarrow \quad th2\_delay_c = \max\_delay_c \tag{3.15}$$

Therefore, if $\gamma_c = 1$, then the intermediate threshold delays converge to the maximum threshold delay as shown by Equations (3.14) and (3.15). Consecutively, this change in the intermediate threshold delay values modifies the DBS scheduling function shown in Equation (3.13) in the following manner:

$$SF_i(t) = \begin{cases} [Norm\_CQI_i(t)] \cdot E_i & If \quad Q_i(t) < \max\_delay_c \\ Packet \ \ Dropped & If \quad Q_i(t) \geq \max\_delay_c \end{cases} \tag{3.16}$$

$$\left\{ \begin{array}{l} \because th1\_delay_c = \max\_delay_c \quad by \quad (3.14) \\ \quad th2\_delay_c = \max\_delay_c \quad by \quad (3.15) \end{array} \right\}$$

where E_i is the priority assigned to a user *i* that defines its QoS requirements. Equation (3.16) schedules users based on their relative channel condition and QoS definitions. This scheme drops packets if the delay experienced by them exceeds a certain maximum value which is defined by their traffic applications. However, the Max CIR scheme does not drop packets as it supports NRT applications that do not have delay constraints. Therefore, in order to prevent packet dropping, $\max\_delay_c$ can be set to a very high value which will modify Equation (3.16) as follows:

$$\begin{aligned}&If \quad \max\_delay_c \rightarrow \infty \\ &then \quad SF_i(t) = \begin{cases} [Norm\_CQI_i(t)] \cdot E_i & If \quad Q_i(t) < \max\_delay_c \rightarrow \infty \\ Packet\ \ Dropped & If \quad Q_i(t) \geq \max\_delay_c \rightarrow \infty \end{cases}\end{aligned}$$

Since it is theoretically impossible to reach a queuing delay value that is greater than infinity then the above expression can be defined as:

$$SF_i(t) = [Norm\_CQI_i(t)] \cdot E_i \qquad If \quad Q_i(t) < \infty \tag{3.17}$$

Equation (3.17) represents a prioritized version of the Max CIR scheme that was derived from the DBS if $\gamma_c = 1$ and $\max\_delay_c \rightarrow \infty$. Note that $Norm\_CQI_i(t)$ is a relative measure of a users' channel quality and essentially represents their channel condition.

Lemma 2: DBS can converge to the Max CIR algorithm

Proof:

The proof for this Lemma is adopted from that of Lemma 1 with the exception of the QoS priority assignment. The value of E_i in Equation (3.17) is dependent on the priority assigned to a user *i*. If this value is kept constant (e.g. $E_i = 1$) for all users then Equation (3.17) is modified to:

$$SF_i(t) = Norm\_CQI_i(t) \qquad\qquad If \quad Q_i(t) < \infty \qquad (3.18)$$
$$\{\because \quad E_i = C \quad \forall i \in \{1,2,...,n\}\}$$

As all users will be assigned the same priority, the scheme defined by Equation (3.18) expresses the Max CIR scheme. Therefore, the DBS converges to the Max CIR algorithm if $\gamma_c = 1$, $\max\_delay_c \to \infty$ and E_i is constant $\forall i \in \{1,2,...,n\}$.

In this section, it was shown that the DBS can converge to the Max CIR scheduling scheme on setting certain parameters. As was discussed in Section 2.4.1, the Max CIR scheme is suitable for NRT applications that are delay tolerable but are highly error sensitive. On the other hand, the DBS algorithm minimizes queuing delay as well as drops packets that exceed their maximum threshold delay term in order to maintain certain delay constraints. Hence, it is suitable for scheduling RT traffic applications. The convergence of the DBS scheme to the Max CIR algorithm will allow the prioritization of RT and NRT traffic classes with the implementation of a single scheduling scheme. This can be done by setting the per-class parameters of NRT data flows so that the DBS mimics the scheduling assignment of the Max CIR scheme whereas RT data flows can have their class parameters set according to the requirements of their applications. For instance, at a given time interval if two types of traffic classes are found in the system viz. conversational and interactive classes, then the conversational traffic class can have a γ_c value greater than one and an appropriate $\max\_delay_c$ value whereas the interactive class will have $\gamma_c = 1$ and $\max\_delay_c \to \infty$. Since the NRT classes do not necessarily have any delay constraints, the scheduling assignment of packets belonging to these classes does not have to consider their queuing delay. Nonetheless, if packets are scheduled in this manner then the scheduling of NRT users will become a best-effort service and these users will be served only if the RT data flows are being scheduled based on their users' CQI values. For this reason, an appropriate maximum threshold delay is defined for NRT applications in the simulations shown in Chapter 4.

3.3.2 Discussions

In Section 3.1, the scheduling problem was defined as the successful implementation of an HSDPA scheduling scheme which is identified by six design objectives. This section investigates and discusses how the DBS accomplishes those objectives.

- ***Throughput Maximization*:** The transmission bit rate of a user at a particular time interval is dependent on their channel condition. Users with better channel conditions are capable of transmitting a higher number of bits per second. The DBS utilizes the instantaneous radio channel information of users for the calculation of their scheduling priority and schedules users that have relatively good channel conditions, hence maximizing system throughput. However, as discussed in Section 3.3.1.4.2, since the DBS utilizes the queuing delay information of a data flow apart from their respective users' channel condition, a fine balance exists between throughput maximization and the minimization of queuing delay. Nevertheless, the DBS can accommodate both these properties with the attunement of γ_c. This will be shown with the help of simulations in Chapter 4.

- ***Minimization of Queuing Delay:*** The DBS considers the queuing delay experienced by a data flow at the Node B for the scheduling decision. Moreover, as discussed in Section 3.3.1.2, the dependence of the scheduling function on the queuing delay increases as a packet approaches its maximum tolerable delay. It will be shown in Chapter 4 through simulation results that the DBS has the lowest average queuing delay when compared to other well known RT scheduling schemes.
- ***Fairness:*** Fairness is an integral part of any scheduling scheme as users with the same service expect similar performance regardless of their channel conditions. However, in order to maximize system throughput HSDPA scheduling schemes monitor the instantaneous channel environment and schedule users with good channel conditions. Therefore, there is a tradeoff between system throughput and fairness. Nonetheless, the DBS increases the scheduling priority of a user as their packets approach their delay term. This is done by increasing the dependence of the scheduling function on the queuing delay factor and consecutively decreasing its dependence on the users' channel condition (shown in Section 3.3.1.2). As all users get an equal chance of being scheduled at higher queuing delays the system fairness is enhanced.
- ***Maintaining Delay Thresholds:*** RT traffic applications are highly delay sensitive and require that their delay thresholds are maintained at the cost of a certain error rate. To accommodate this feature, the DBS drops packets of data flows when their queuing delay exceeds the maximum tolerable amount defined by their traffic class. Furthermore, as the DBS tries to minimize queuing delay, majority

of the packets are transmitted before their expiry term which in turn reduces the overall fraction of packets dropped.

- ***QoS Prioritization:*** A variety of services are available on a cellular network simultaneously which are distinguished based on the QoS requirements of users. The DBS defines a QoS parameter for each user and utilizes these prioritization parameters for the scheduling decision (illustrations were shown in Section 3.3.1.4.2). This will allow service provider to cater the offered services based on the individual user requirements.
- ***Scheduling RT and NRT traffic classes simultaneously:*** The DBS is essentially a RT scheduling scheme as it preserves the delay constraints of applications belonging to this traffic type. Moreover, NRT traffic applications are delay tolerable but expect a very low error rate. In other words, RT applications can tolerate a certain amount of error but expect packets to be transmitted within a certain time interval whereas NRT applications are highly error sensitive but do not have a strict response time. For this reason, the Max CIR scheme can best schedule NRT applications. In Section 3.3.1.5, it was shown that under certain conditions (or with the setting of certain parameters), the DBS converges to the Max CIR scheme which has brought up the interesting possibility of scheduling RT and NRT traffic classes simultaneously, with the implementation of a single scheduling scheme. RT scheduling schemes can be used to schedule NRT traffic by setting certain delay thresholds for them. But this varies the NRT traffic characterization as these applications do not have delay thresholds. The DBS, on the other hand, can be used to support RT as well as NRT applications without

changing the traffic characterization of these applications by converging to the Max CIR scheme for certain traffic classes.

The implementation of these design objectives in the DBS will be further affirmed with the help of simulations in the next chapter.

3.4 Summary

HSDPA can support data rates of up to 14.4 Mbps through increased spectral efficiency and the utilization of instantaneous channel conditions in the scheduling decision which has led to the emergence of various packet based RT as well as NRT services on to this cellular network. Some of the original HSDPA schedulers only considered the channel conditions of users for the allocation of the HS-DSCH in order to improve system throughput. However, such scheduling techniques have short-comings such as unfair channel allocation, failure to meet delay requirements, etc. Hence, in this chapter, it was remarked that the design objectives of a successful HSDPA packet scheduler can be identified as: throughput maximization, minimization of queuing delay, fairness, maintaining delay thresholds, QoS prioritization and scheduling RT and NRT traffic classes simultaneously. Furthermore, a Delay Based Scheduler (DBS) was proposed that tries to implement the above mentioned design objectives. The proposed scheme not only considers the short-term channel conditions of users but also the instantaneous queuing delay of their data flows' head packet in the calculation of the scheduling priority. Moreover, the DBS increases the dependence of the scheduling function on the users' queuing delay and conversely decreases its dependence on their channel quality as their queuing delay approaches its maximum value which is defined by their traffic class. This ensures that the aggregate queuing delay at the packet scheduler is minimized in such a way that it does not considerably compromise system throughput. Additionally, it improves fairness as users with poor channel conditions have a higher probability of being scheduled as their queuing delay increases. In order to maintain the delay constraints of RT applications, certain delay thresholds are defined on a per class basis, so that packets whose queuing delay exceeds this delay limit are dropped. The DBS defines a QoS parameter for each user based on their service requirements and utilizes them for the scheduling assignment. The effect of such QoS prioritization on the system performance was also illustrated in this chapter. Finally, it was shown that the DBS can converge to the Max CIR algorithm under certain conditions and hence can inter-operate between RT and NRT applications without changing the characterization of these traffic classes. It has been shown in this chapter that the DBS achieves the design objectives that were previously outlined. Nevertheless, these proofs will be confirmed through performance evaluation which is the subject of the next chapter.

Chapter 4

Performance Evaluation

In this chapter, the performance of the DBS algorithm is evaluated and compared with those of other well-known HSDPA scheduling algorithms. Moreover, the simulations that are presented in the forthcoming sections demonstrate that the design objectives of a successful scheduling scheme (which were defined in Section 3.1) have been realized by the DBS. Section 4.1 describes the simulation environment along with its traffic and channel models. The performance metrics that will be used to evaluate the various scheduling schemes are defined in Section 4.2. The performance of the DBS under various traffic and mobility conditions is analyzed and compared to other algorithms in Section 4.3. Finally, Section 4.4 presents a summary of the observations.

4.1 Simulation Methodology

A dynamic simulation environment is designed and implemented using a simulation tool known as the Network Simulator version 2 (NS2). NS2 is an open source simulation tool developed by a research team at the University of California, Berkley. Although NS2 is capable of simulating various network environments, it cannot support HSDPA by itself. Therefore, an extension to NS2 was used for the simulation which is known as the Enhanced UMTS Radio Access Network Extension for NS2 (EURANE). EURANE can support UMTS as well as HSDPA and it was used along with NS2 for the performance evaluation of the DBS. Further details about EURANE and NS2 can be found in [27, 28] and [29] respectively.

This section outlines the simulation methodology that was used for the performance evaluation of the DBS. The simulation model is defined in Section 4.1.1. The traffic model is discussed in Section 4.1.2. Finally, the propagation model used in EURANE is described in Section 4.1.3.

4.1.1 Simulation Model

A single cell environment was simulated involving a single Node B which is located at the center of the cell with the UEs evenly distributed throughout the cell (shown in Figure 4.1). Data packets are transmitted to the UEs through the HS-PDSCH (see Section 2.3.1 for details) whereas control information is sent by each user through the HS-DPCCH. At every TTI, each UE calculates the strength of the signal transmitted by the Node B in order to quantitatively measure its channel condition. This instantaneous channel condition is then used to evaluate a recommended size of the transport block as well as modulation and channelisation codes that are supportable by the UE. This information is then sent to the Node B through the HS-DPCCH in the form of a CQI. EURANE defines the CQI in such a manner that it is reported with a Block Error Rate (BLER) value which is less than or equal to 10%.

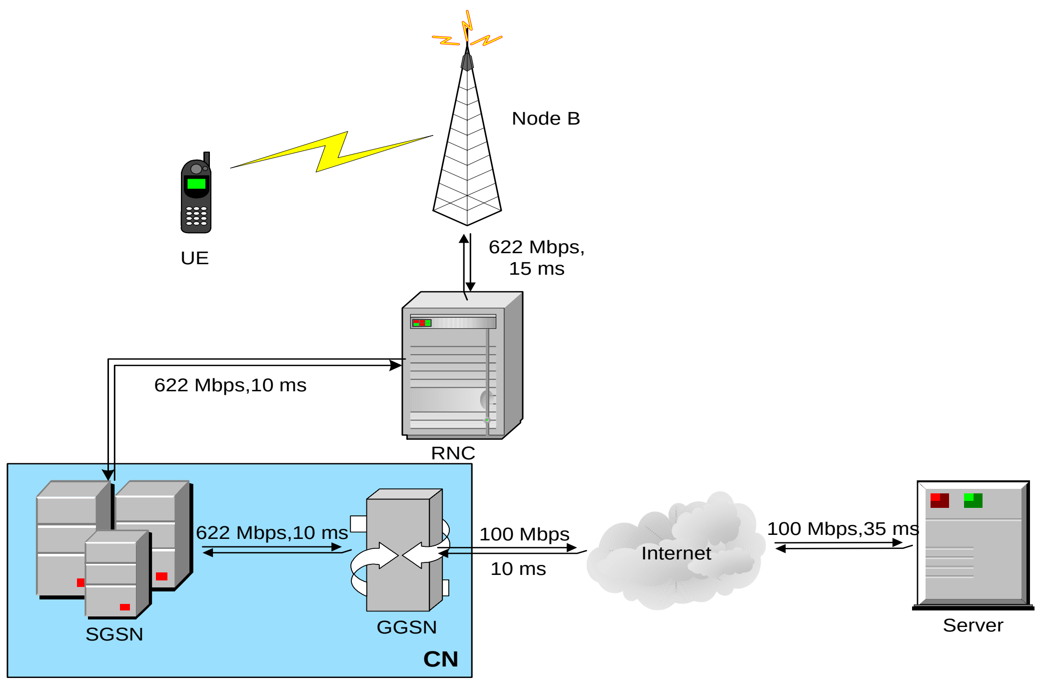


Figure 4.1: The Simulation Model

The Node B is connected to the RNC via the Iub which is a wired duplex link of 622 Mbps bandwidth and 15 ms delay. Consecutively, the RNC is connected to the SGSN with a duplex link of 622 Mbps bandwidth and 10 ms delay. The two CN nodes: SGSN and GGSN are connected by a duplex link of 10 ms delay and 622 Mbps bandwidth. The CN is then connected to the external IP networks via a 100 Mbps and 10 ms delay duplex link. Finally, the traffic sources are connected to the Internet with a 100 Mbps and 35 ms delay duplex link. Moreover, users are uniformly distributed in the cell and the environment is simulated for 200 s with a TTI value of 2 ms.

The 3GPP has recommended two simulation environments for the evaluation of scheduling schemes in HSDPA i.e. Pedestrian A and Vehicular A environments [28]. The Pedestrian A environment simulates UEs that have a speed of 3 km/hr in cell of 500 m diameter thus this environment is used to emulate pedestrians that are using the cellular network. Similarly, the Vehicular A environment simulates vehicle bound UEs traveling at a speed of 60 km/hr in a cell of 1000 m diameter. Furthermore, a third environment is used for the evaluation process under which the channel conditions of users are fixed over time. This is known as the fixed channel environment. The purpose of using this environment for the evaluation process is to study the degree of fairness of the evaluated scheduling algorithms. In other words, under this environment users will always have a fixed channel condition with some users with good channel conditions and others with average or poor channel conditions. The various scheduling schemes under the evaluation process use the channel conditions of users in order to prioritize them for scheduling. A scheduling algorithm that distributes system resources in a fair manner assigns priorities to users such that all UEs have a similar performance regardless of their signal strength. As this environment sets the channel conditions of users to a fixed value the performance of UEs with good channel conditions can be compared to those with poor channel conditions by comparing their fairness index. Such tests are not possible under varying channel environments such as Pedestrian A and Vehicular A.

4.1.2 Traffic Model

In the forthcoming sections, the performance of the DBS is evaluated and compared with other scheduling schemes for a RT as well as NRT traffic environment with each test case involving two traffic classes. For this reason, the simulation environment uses all four traffic classes that are specified by the 3GPP viz. conversational, streaming, interactive and background traffic.

The conversational and streaming classes will be used for the evaluation of the various HSDPA scheduling schemes in a RT traffic environment. Conversational traffic is modeled using an exponential traffic application that generates traffic according to an exponential on-off distribution. Packets are sent at a fixed rate during on periods, and no packets are sent during off periods. Both on and off periods are taken from an exponential distribution, with a constant packet size and a traffic rate of 64 kbps. Streaming traffic is often modeled using a Constant Bit Rate (CBR) type traffic application [7, 24] that transmits a fixed amount of bits per second. In the forthcoming simulations, a CBR traffic application with a data rate of 128 kbps is used to model streaming traffic.

The NRT traffic environment involves the interactive and background classes. Interactive traffic is represented using a CBR type traffic application with a data rate of 64 kbps. Background traffic is modeled using an FTP type traffic application with a packet size of 320 bytes. FTP applications are used over networks that support the Transmission Control Protocol (TCP). These applications transmit new data packets based on acknowledgments received for previously transmitted data packets. Therefore, an FTP application does not have a fixed data rate as their transmission rate depends on the round-trip-time of acknowledgements sent by the receiver. Nonetheless, the number of transmitted packets for FTP application is higher when the load on the network is lower and vice versa. The traffic model parameters for RT as well as NRT applications are given in Table A.1 in Appendix A.

In the RT traffic environment, the offered traffic load is equally contributed by each of its traffic classes. For instance, if the total traffic load is equivalent to 512 kbps then 256 kbps are contributed by each class implying that there would be four users in the cell belonging to the class with a data rate of 64 kbps and two users with a data rate of 128 kbps. Moreover, the simulation results for the NRT traffic environment are plotted against the number of users in the system since the transmission rate of FTP applications is not fixed and hence the offered traffic load for this environment cannot be determined at any given time interval.

4.1.3 Propagation Model

At every TTI, the radio signal broadcasted by the Node B (to all the users in the cell) has a constant power. However, this radio signal is attenuated on its way from the Node B because of the users' environment and mobility. This is known as the propagation model which describes the change in the radio signal received by the user and the resulting variation in their channel condition with time. EURANE has divided the propagation model into five parts: inter-cell interference, intra-cell interference, path loss, shadowing and multi-path fading. The inter-cell interference is caused by the activities of the surrounding cells whereas the intra-cell interference is caused by other UEs in the same cell. Path loss is the attenuation of the radio signal as it propagates through space and is highly influenced by the distance between the Node B and the user. Its value is proportional to the distance between these two network elements and hence increases as distance increases and vice versa. Shadow fading is caused by obstacles in the propagation path between the Node B and the UE. These obstacles block the direct transmission of the radio signal but due to the properties of reflection and diffraction of electromagnetic waves, this signal is still transmitted to the user. Moreover, the diffraction of radio signals because of obstacles in its path results in the replication of this signal such that the received signal is the sum of these replications. This phenomenon is known as multi-path fading or fast fading where the radio signal may have reached the receiving antenna by two or more paths. Furthermore, path loss, shadowing and multi-path fading are collectively known as the channel model. The five components of the propagation model are considered independent of each other and are expressed in dB [28]. All the relevant propagation model parameters are show in Table A.2.

In EURANE, both the inter-cell as well as the intra-cell interference are modeled as a constant power as real life fluctuations in interference have little impact on the end result of the simulator when compared to the variations introduced by the channel model. Their values correspond to 30 dBm and -70 dBm respectively. Path loss is calculated using the following formula:

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

where L_{init} is the distance loss and its value at 1 km is equal to 1.374e2, d is the distance between the Node B and the UE in km and β is the path loss exponent which is equal to 3.52. It should be noted that the distance and path loss components depend on the environment and their values mentioned above are for the most practical situations. Shadow fading is modeled through a lognormal distribution using a correlated slow fading model which is constructed in the following manner:

$$S(x+\Delta x) = a \cdot S(x) + b \cdot \tilde{\sigma} \cdot N$$
$$a = \exp(-\Delta x / D)$$
$$b = \sqrt{1-a^2}$$

where Δx is the variation in distance between two subsequent time samples, N is the standard normal distribution and has a random value, $\tilde{\sigma}$ is the standard deviation which has a typical value of 8 dB in suburban areas and, finally, D is the correlation distance which depends on the speed and environment of the UE. In EURANE, multi-path fading corresponds to the 3GPP channel models for the Pedestrian A and the Vehicular A environments [28]. At the UE, the received signal from the Node B is used to determine the signal strength by the calculation of the Signal-to-Noise Ratio (SNR) with the following formula:

$$\begin{aligned} SNR &= P_{Tx} - L_{Total} - 10\log_{10}(10^{\frac{I_{Intra}-L_{Total}}{10}} + 10^{\frac{I_{Inter}}{10}}) \\ &= P_{Tx} - 10\log_{10}(10^{\frac{I_{Intra}}{10}} + 10^{\frac{I_{Inter}+L_{Total}}{10}}) \end{aligned}$$

where P_{Tx} is the Node B transmission power in dBm, L_{Total} is determined by the channel model i.e. it is the sum of the path loss, shadow fading and multi-path fading components. I_{Intra} and I_{Inter} are the intra and inter cell interference respectively in dBm.

The SNR is then mapped to the CQI value through a linear function that approximates a BLER value of 0.1:

$$CQI = \begin{cases} 0 & SNR \le -16 \\ \left\lfloor \dfrac{SNR}{1.02} + 16.62 \right\rfloor & -16 < SNR < 14 \\ 30 & SNR \ge 14 \end{cases} \tag{3.19}$$

Once the CQI is calculated, it can be used to determine the size of the transport block (TBS) as well as the modulation and channelisation codes that are supportable by the UE. The 3GPP has defined different UE categories which map different TBS values to the same CQI values. A full combination of these categories along with their TBS and CQI values are stated in [30]. In this simulation, it is assumed that all UEs belong to categories 1 to 6. Moreover, the TBS represents the number of bits that can be transmitted to a user in a particular TTI. This information can then be used to calculate the users' supportable data rate at the given time interval. For example, if the TBS of a user is 100 bits in a time interval of 2 ms then this user has an instantaneous supportable data rate of 50 kbps. The scheduling schemes that will be compared with the DBS in the forthcoming sections use this instantaneous data rate for the calculation of a users' priority.

4.2 Performance Metrics

The performance of the DBS is evaluated in terms of the following performance metrics:

- ***Throughput:*** The total number of bits transmitted to all the users in the cell, through the HS-DSCH, during the simulation period.

- ***Queuing Delay:*** The average queuing delay experienced at the Node B by the head-of-line packet of all data flows.
- ***Fraction Dropped:*** Packets are dropped at the Node B if their queuing delay exceeds the maximum tolerable amount stated by their traffic class. The fraction of packets dropped is defined as the average ratio of the number of PDUs dropped over the total number of PDUs that arrived at the Node B.
- ***Jain Fairness Index (JFI):*** A fairness index that that uses the variance in user throughput to calculate fairness among them [31]. It is calculated in the following manner:

$$JFI = \frac{\left(\sum_{i=1}^{n} S_i\right)^2}{n \cdot \sum_{i=1}^{n} (S_i)^2}, S_i \geq 0 \qquad \forall i$$

  where S_i is the average throughput for user i and $i \in \{1,2,...,n\}$. It should be noted that the JFI is equal to 1 if all users have the same average throughput. Lower JFI values indicate that users have high variances in their average throughputs and hence they practice unfairness employing this scheme and vice versa.

4.3 Simulation Results

In this section, the performance of the proposed DBS scheme is evaluated and compared with the following HSDPA scheduling schemes: Max CIR, PF, M-LWDF and ER (discussed in Section 2.4). In the simulations that follow, it is assumed that users are given QoS prioritization based on the maximum tolerable delay requirements of their traffic class. Therefore, in each environment users belonging to traffic classes with lower delay tolerance are given a higher QoS priority whereas users with higher delay tolerance are given a lower QoS priority. In other words, QoS prioritization is done based on a per-class basis rather than a per-user basis such that $E_i = E_c \forall i \in c$. As each simulation environment comprises of only two traffic classes, only two values of E_c are defined. Furthermore, the simulations are performed in the Pedestrian A, Vehicular A and the fixed channel environments with the results illustrated in Sections 4.3.1, 4.3.2 and 4.3.3 respectively.

4.3.1 Pedestrian A Environment

We have demonstrated in Chapter 3 that the DBS achieves the design objectives of a successful HSDPA scheduling scheme that were outlined in Section 3.1. However, it was mentioned previously that the attunement of γ_c can balance the tradeoff between throughput maximization and the minimization of queuing delay. This will be shown in Section 4.3.1.1 along with the effect of varying constants (which were defined for the DBS) on the performance of the scheme. We study the performance of the DBS for the RT traffic environment on a per-class basis in Section 4.3.1.2 and we compare it with other existing HSDPA scheduling schemes in Section 4.3.1.3. We then evaluate the scheduling schemes for the NRT and mixed traffic environments in Sections 4.3.1.4 and 4.3.1.5 respectively.

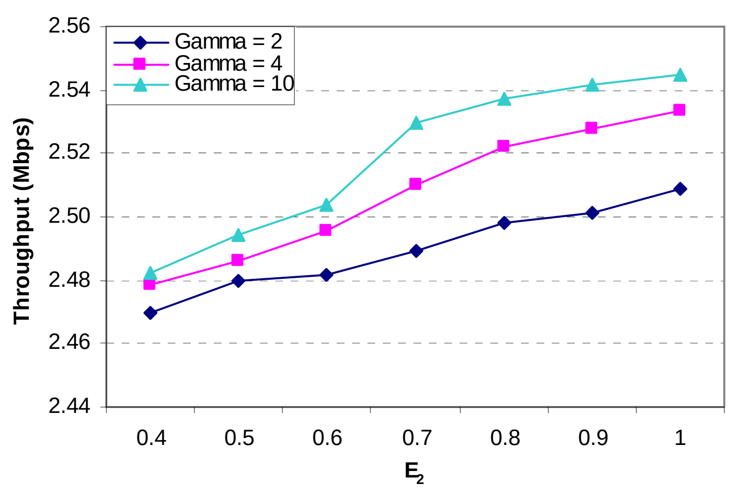
4.3.1.1 Performance Evaluation of DBS for Varying Constants

The three main control parameters defined for the DBS algorithm are: E_i, γ and $\max\_delay_c$. However, it was mentioned above that users are prioritized based on the delay requirements of their traffic class such that $E_i = E_c \forall i \in c$. As there are only two traffic classes in every simulation environment (i.e. $c \in \{1,2\}$), only two values of E_c are defined. It should be noted that classes are categorized in descending order of QoS prioritization, thus $E_1 > E_2$. In this section RT traffic is evaluated which includes the conversational and streaming traffic classes. Since conversational traffic is more delay sensitive than streaming traffic, it is given a higher priority of E_1 whereas streaming traffic is given a lower priority of E_2. Moreover, E_1 is given a constant value of 1 and the value of E_2 is varied so that the effect of QoS prioritization can be seen on the performance of the DBS. Furthermore, these results are simulated in the Pedestrian A environment.

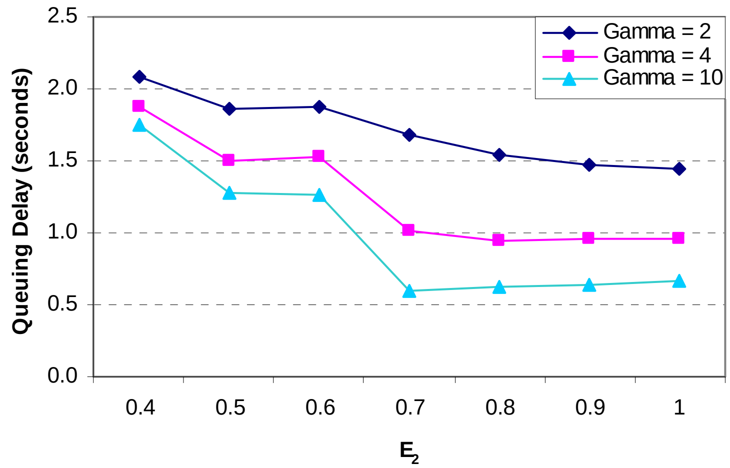
Figure 4.2 illustrates that the aggregate system performance improves as the value of E_2 increases. This is because increasing E_2 enhances the priority of the streaming class which consequently reduces the number of packets that are dropped by the data flows of this class due to exceeding queuing delays and in-turn increases its throughput. In Figure 4.2(a), it can be seen that the overall cell throughput increases from 2.48 Mbps to 2.542 Mbps when the value of E_2 varies from 0.4 to 1.0 for a γ value of 10. This indicates a 2.44% increase in cell throughput. Figures 4.2(b) and 4.2(c) represent the decrease in queuing delay and fraction dropped respectively, for varying values of E_2 and γ. It is shown in these figures that as E_2 varies from 0.4 to 1.0, the average queuing delay decreases from 1.62 s to 0.66 s whereas fraction dropped decreases from 2.2% to 0.49% for $\gamma = 10$. Therefore, the QoS prioritization of traffic classes can significantly vary the performance of the DBS algorithm, hence proving that this design objective is achieved by the algorithm. It should be noted that these results were simulated for a traffic load of 2.56 Mbps with $\max\_delay_1 = 2s$ and $\max\_delay_2 = 4s$.

Although the definition of the prioritization parameters is determined by the service provider, they should be defined in a manner that prevents the starvation of users with lower priorities. This is because the starvation of lower priority users might degrade the overall system performance depending upon the traffic load. In these simulations the traffic load is equally contributed by each of the traffic classes (i.e. traffic load is contributed on a 1:1 ratio by each of the traffic classes). However, because the higher priority class is more delay sensitive, this class is given preference in the scheduling assignment such that the QoS parameters are defined in the following manner: $E_1 = 1.0$ whereas $E_2 = 0.7$ (i.e. scheduling assignment is done based on a 1:0.7 ratio). In the rest of this thesis, the above mentioned QoS prioritization is maintained for the remaining simulations.

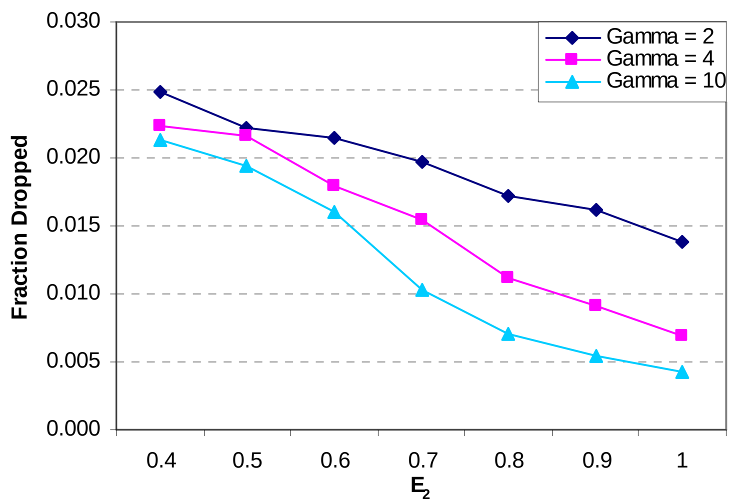
It can be seen in Figure 4.2 that higher values of γ uniformly improve system performance in terms of throughput, queuing delay as well as fraction dropped since a γ value of 10 provided the best overall performance whereas lower γ values offered worse system performance. Therefore, it could be inferred that increasing γ will improve the aggregate system performance (i.e. if $\gamma \leq 10$). However, as can be seen in Figure 4.3, increasing γ beyond a value of 10 does not necessarily improve the overall system performance. This is because as γ increases the values of the intermediate threshold delays that were defined in Section 3.3 decrease. Consequently, as the values of the threshold delays decrease, the dependence of the scheduling assignment on the queuing delay increases which in-turn decreases its dependence on the channel condition of a user. However, in order to maximize system throughput it is imperative that users with good channel conditions are scheduled so that higher number bits are transmitted in any given interval. Therefore, increasing the value of γ minimizes the average queuing delay experienced by all data flows at the cost of lower overall cell throughput. Hence, γ is the parameter that can be attuned to balance the tradeoff between throughput maximization and the minimization of queuing delay.



(a)



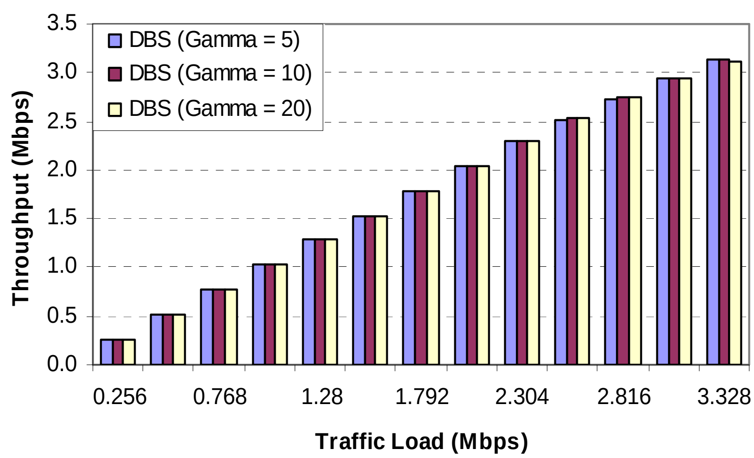
(b)



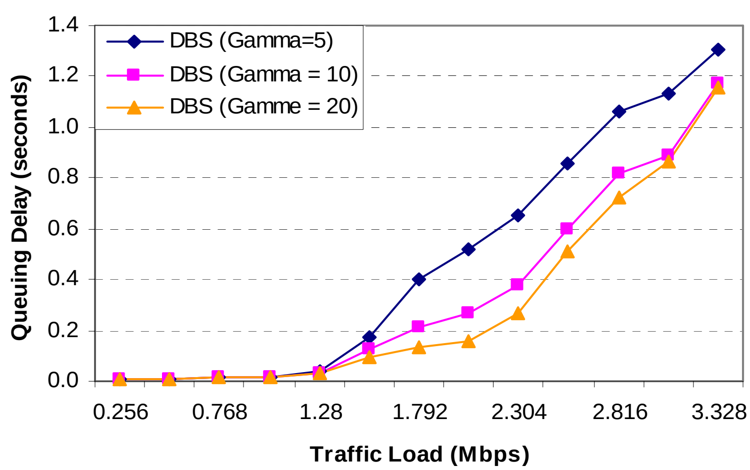
(c)

Figure 4.2: The effect of QoS Prioritization on the Performance of the DBS in the RT traffic environment for a traffic load of 2.56 Mbps
(a) Throughput (b) Queuing Delay (c) Fraction Dropped

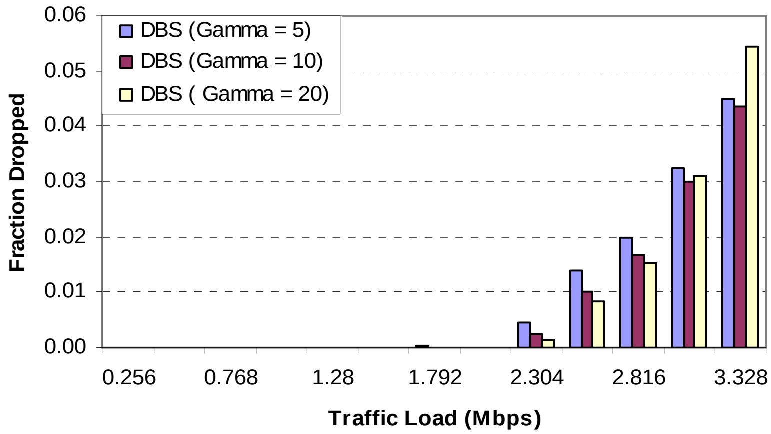
Figure 4.3(a) shows that the cell throughput for γ values of 5, 10 and 20 does not vary up to a load of 2.048 Mbps beyond which the throughput starts decreasing. Figure 4.3(b) shows that a higher value of γ diminishes the average queuing delay of the system with the variance in queuing delay between γ values of 10 and 20 being minimal (i.e. less than 200 ms for any traffic load). As shown in Figure 4.3(c), packets are dropped beyond an arrival rate of 2.048 Mbps due to traffic congestion which is in agreement with Figure 4.3(a). Moreover, it is seen in the figure that when $\gamma = 10$ the overall fraction of packets dropped is the least. Figure 4.3(d) shows that the JFI for the Pedestrian A environment is greater than 0.99 for any traffic load implying that the DBS prioritizes users in a fair manner. Since a γ value of 10 maximizes system throughput while maintaining a low average queuing delay, it represents the best value for this parameter and will be used in the remaining simulations.



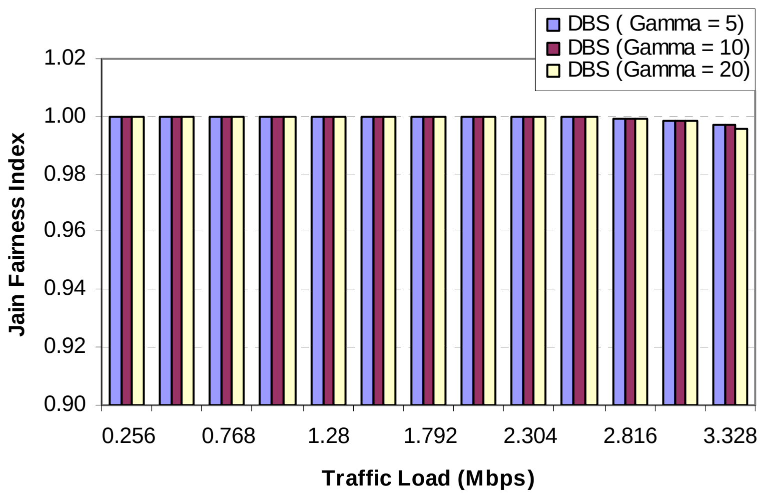
(a)



(b)



(c)



(d)

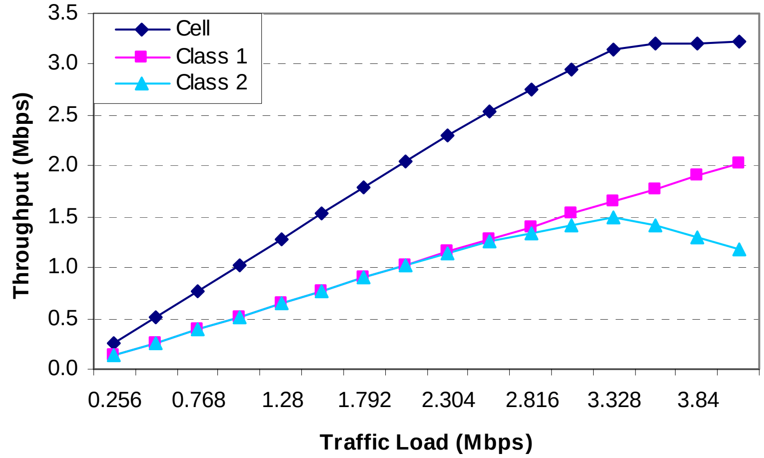
Figure 4.3: The effect of γ on the performance of the DBS in the RT traffic environment for varying traffic loads.
(a) Throughput (b) Queuing Delay (c) Fraction Dropped (d) Jain Fairness Index

It was concluded from simulation results that increasing the maximum threshold delay of a traffic class increases the total system throughput apart from increasing its average queuing delay. These results are shown in Figure B.1 in Appendix B. However, it was seen that doubling the $\max\_delay_c$ value of a traffic class enhanced the cell throughput by less than 3%. Therefore, the maximum threshold delays should be kept minimal in order to minimize queuing delay while avoiding excessive packet dropping which will drastically reduce system throughput. For this reason, the threshold delay values have been set to 2 s and 4 s for the higher and lower priority classes respectively. The values of the DBS parameters that have been defined for the forthcoming simulations are shown in Table A.3.

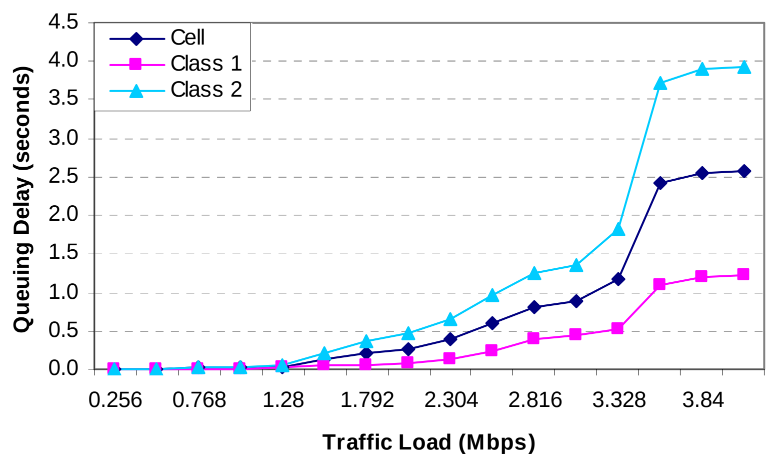
4.3.1.2 Per-Class Performance Evaluation of DBS

The DBS categorizes users based on their QoS requirements. The assignment of the QoS priorities is determined by the service provider. However, as mentioned previously, in these simulations it is assumed that users are prioritized based on the delay requirements of their traffic classes. This section compares the performance of the conversational and streaming traffic classes that were defined for the RT traffic environment.

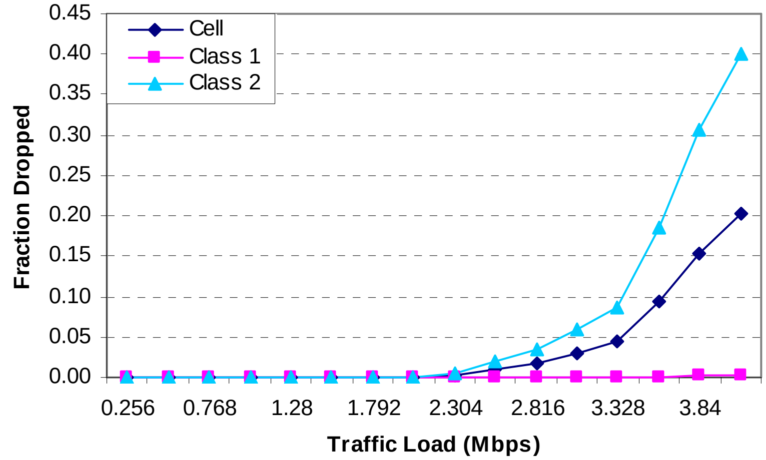
Figure 4.4 illustrates the per-class as well overall cell performance of the system when the DBS is employed for the scheduling decision. As the conversational class is strictly delay sensitive it is given a higher priority ($E_1 = 1$) and is represented in the figure by class 1. Lower priority is given to the streaming class ($E_2 = 0.7$) which is symbolized as class 2. Figure 4.4(a) shows the throughput of the classes as well as the total cell throughput for varying traffic loads. Recall that the traffic load is equally contributed by each of the traffic classes and hence the throughput of both classes is equal up to a traffic load of 2.56 Mbps. However, in these simulations UE categories 1 to 6 are used which have a maximum supportable data rate of 3.584 Mbps, as defined by the 3GPP in [30]. The reduction in throughput after a traffic load of 2.56 Mbps can be accounted by cell congestion since this value of traffic load exceeds 71% of that supportable by the system. For traffic loads greater than or equal to 2.816 Mbps, it is seen that the total cell throughput is the sum of 100% of the load offered by class 1 and 70% of class 2s' offered load since these classes are prioritized on a 1:0.7 ratio respectively. For instance, at a load of 4.096 Mbps, 2.048 Mbps is contributed by each class but only the throughput of class 1 is equal to this value whereas the throughput of class 2 is approximately equal to 70% of this value which is 1.4336 Mbps. Moreover, 3.584 Mbps is a theoretical maximum data rate but in the simulations it is seen that the data rate does not exceed 3.25 Mbps. A maximum threshold delay value of 2 s has been set for the conversational class whereas this value is set to 4 s for the streaming class. Figure 4.4(b) illustrates that the DBS maintains the delay thresholds of the traffic classes as the average queuing delay experienced by class 1 and class 2 is below 2 s and 4 s respectively. Moreover, it is seen that the queuing delay radically increases beyond a traffic load of 3.584 Mbps. This is because the cell can support a maximum load of 3.584 Mbps, as explained above and increasing the traffic load beyond this point causes cell congestion. Furthermore, it can be depicted from the figure that the value of the average queuing delay is less than half of that defined as the maximum tolerable amount up to the point of cell congestion. The fraction of packets dropped, shown in Figure 4.4(c), never exceeds 20% even at high traffic loads. Moreover, the higher priority class has a drop rate of less than 1% which is preferable as the 3GPP has defined this value to be less than 10% for conversational classes. However, this is achieved at the cost of higher drop rates for the lower priority class. Figure 4.4(d) illustrates that the variance in throughput for the Pedestrian A environment is small which, once again, implies that the DBS scheme employs fairness in the assignment of users to the shared channel. In addition, it is seen that the conversational or higher priority class has higher values of JFI when compared to the streaming class.



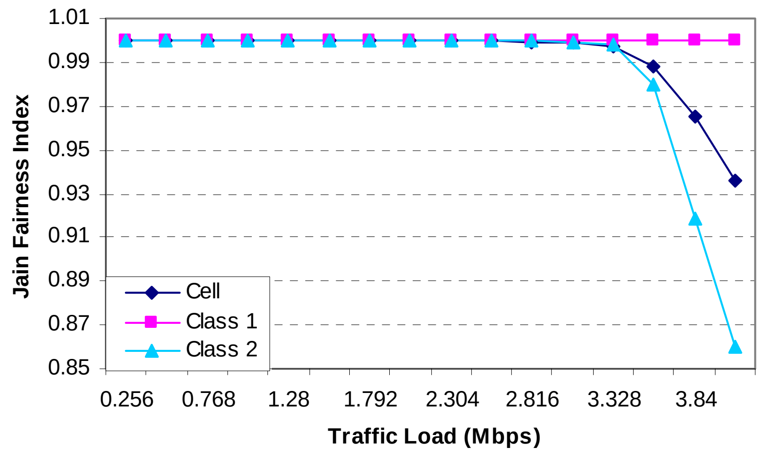
(a)



(b)



(c)



(d)

Figure 4.4: Per-Class Evaluation of the DBS in a RT traffic environment for varying traffic loads (a) Throughput (b) Queuing Delay (c) Fraction Dropped (d) Jain Fairness Index

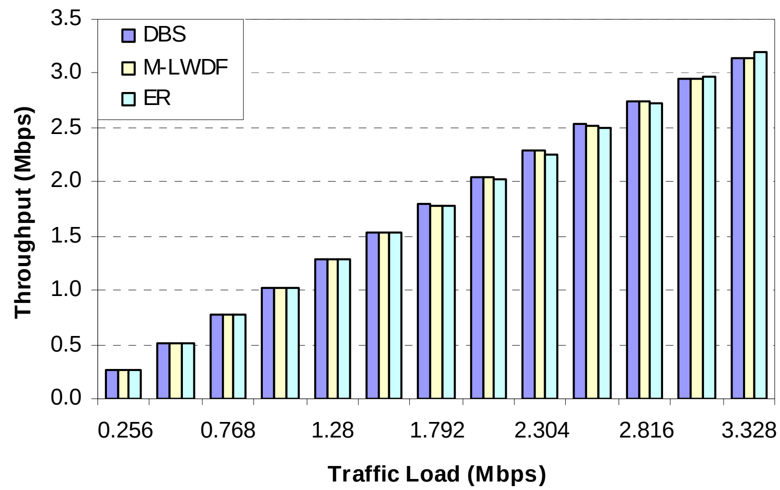
The above discussion along with Figure 4.4 shows that the DBS not only prioritizes users based on their service requirements but also provides better performance to the higher priority class when compared to the lower one. Furthermore, this scheme maintains a high cell performance in terms of throughput, queuing delay, fraction dropped and JFI as it prevents starvation.

4.3.1.3 Performance Evaluation of Different Scheduling Schemes for a Real-Time Traffic Environment

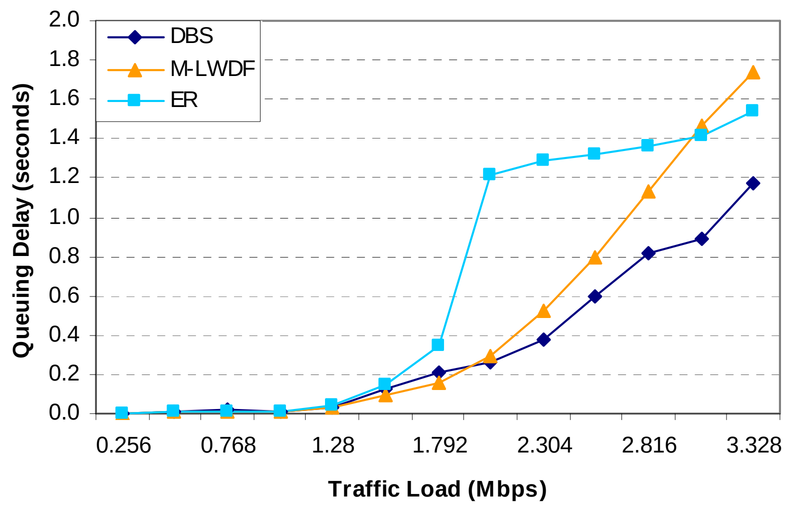
The performance of the proposed scheme is compared with other RT scheduling schemes in this section. In particular, the DBS is compared with the M-LWDF and ER scheduling algorithms which were discussed in Section 2.4. Both schemes utilize the instantaneous supportable data rate of a user at a given time interval and their throughput up to that time for the scheduling assignment. The calculation of a users' instantaneous supportable data rate from their TBS, at a particular time interval, is discussed in Section 4.1.3. Throughput can be calculated as the ratio of the sum of the bits transmitted to a user through the HS-DSCH up to a given time interval over the total time spent by the UE at the system. The delay thresholds and prioritization parameters that have been set for the DBS scheme are also used for the M-LWDF and ER schemes.

The performance of the DBS is compared with the M-LWDF and ER scheduling algorithms and the simulation results are shown in Figure 4.5.The throughput of the three schemes for varying traffic loads is illustrated in Figure 4.5(a). The three schemes maintain the same throughput up to a traffic load of 1.536 Mbps. However, beyond a traffic load of 1.792 Mbps the DBS achieves a higher overall throughput (1.790, 2.045, 2.295, 2.529, 2.744, 2.942, 3.139 Mbps) when compared to the M-LWDF (1.787, 2.039, 2.284, 2.510, 2.738, 2.954, 3.146 Mbps) and ER schemes (1.774, 2.022, 2.258, 2.488, 2.731, 2.969, 3.203 Mbps). The slight drop in the throughput attained by the DBS at traffic loads of 3.072 Mbps and 3.328 Mbps can be accounted by traffic congestion as these load values exceed 85% of that supportable by the system. Moreover, the DBS strives to maintain the performance of higher priority classes at the cost of lower performance for the lower priority classes. Figure 4.5(b) illustrates the average queuing delay of the aforementioned schemes at varying traffic loads. The figure shows that the DBS maintains the lowest value of average queuing delay when compared to other RT schemes. This is because, the DBS scheduling algorithm reduces its dependence on the channel condition as a data flows' queuing delay approaches its deadline and consequently increases its dependence on queuing delay. Since the channel condition information is given less preference at higher queuing delays the throughput of the DBS at higher traffic loads is diminished (as shown in Figure 4.3(a)). It should be noted that this figure is plotted in terms of seconds and a small variance in delay represents an enormous difference in milliseconds and in-turn in TTI (e.g. difference of 0.2 s is equivalent to 200 ms or 100 TTIs). The fraction of packets dropped by the DBS, M-LWDF and ER schemes is shown in Figure 4.5(c) for increasing traffic loads. The fraction of packets dropped for any traffic load is the ratio of the total number of packets that were transmitted by the Node B over the total number of packets that arrived at the Node B. Therefore, the variance in the throughput achieved by the scheduling schemes can be accounted by the fraction of packets that were dropped when these scheduling schemes were utilized. In other words, it was observed in Figure 4.3(a) that the throughput of the three schemes is comparable up to a traffic load of 1.536 Mbps. This is shown in Figure 4.3(c) as a small percentage of packets are dropped up to this load. Moreover, it was remarked that the DBS achieves a higher throughput up to a load value that exceeds 85% of that supportable by the system. Once again, Figure 4.3(c) illustrates that the DBS has the least number of packets dropped up to a traffic load of 3.072 Mbps.

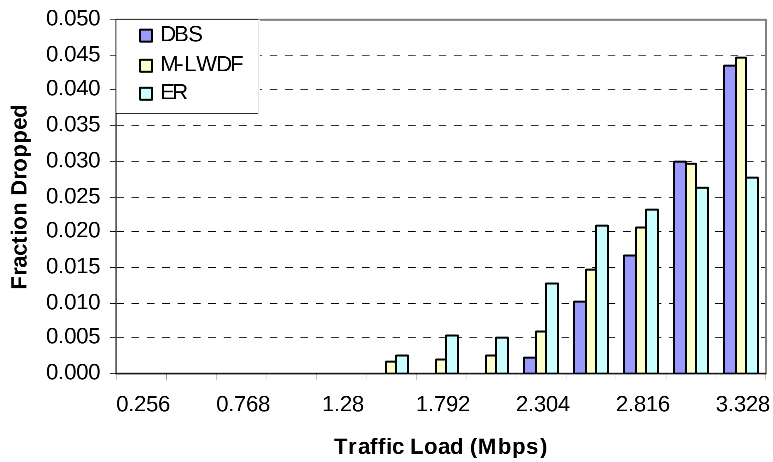
Furthermore, the DBS drops a higher percentage of packets for traffic loads of 3.072 Mbps and 3.328 Mbps but this value never exceeds 4.5% as confirmed by Figure 4.4(c) which was presented in the previous section. All RT scheduling schemes fairly assign users to the HS-DSCH in the Pedestrian A environment as shown in Figure 4.5(d).



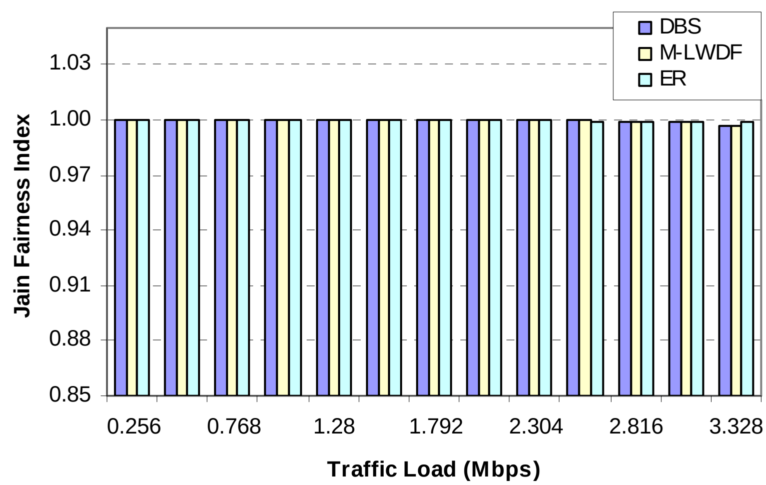
(a)



(b)



(c)



(d)

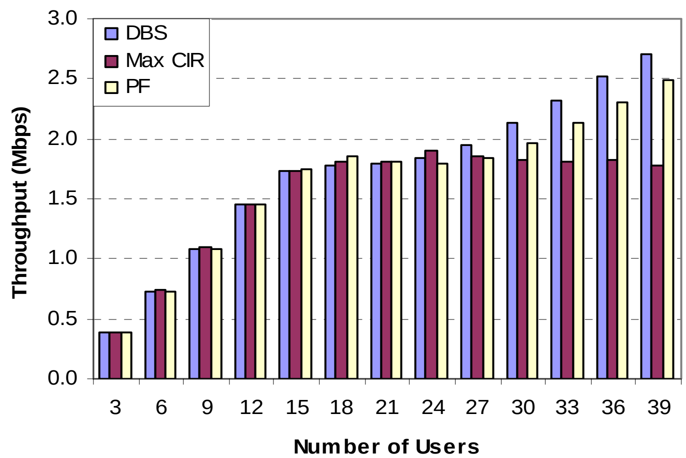
Figure 4.5: Performance of RT scheduling schemes in a Pedestrian A environment with varying traffic loads
(a) Throughput (b) Queuing Delay (c) Fraction Dropped (d) Jain Fairness Index

In this section, it was shown that the DBS achieves the least queuing delay when compared to other RT scheduling algorithms designed for HSDPA. Moreover, this scheme minimizes queuing delay without sufficiently compromising system throughput. In fact, the DBS achieved higher throughputs than the M-LWDF and ER schemes up to a load which is equivalent to 85% of the systems' total supportable data rate.

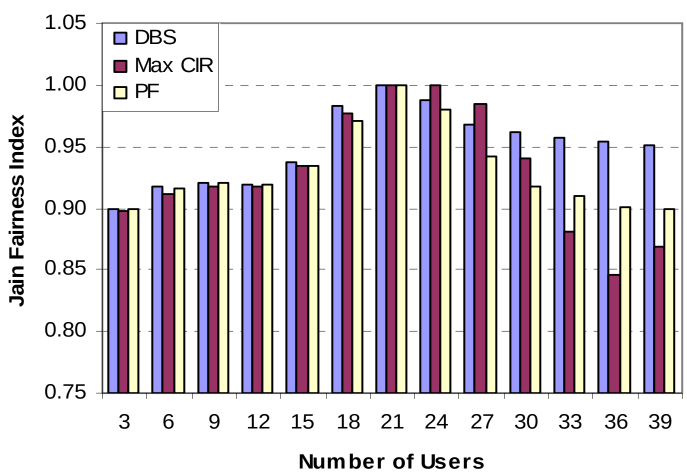
4.3.1.4 Performance Evaluation of Different Scheduling Schemes for a Non-Real-Time Traffic Environment

In this section, the performance of the DBS scheduling algorithm is compared against that of the Max CIR and PF schemes for a NRT traffic environment. The interactive and background traffic classes used for this environment are simulated with CBR and FTP type applications. Threshold delay parameters for the DBS have been set as follows: $\max\_delay_1 = 2$ and $\max\_delay_2 = 4$ where the delay tolerable background class is given a lower priority than the interactive traffic class and the scheduling priority is expressed in Equation (3.13).

Figure 4.6 shows the performance of the DBS, Max CIR and PF schemes for a NRT traffic environment. As can be seen in Figure 4.6(a), the DBS outperforms the Max CIR and PF schemes at high traffic loads in terms of throughput. This is because the DBS utilizes the instantaneous queuing delay of a data flow in the scheduling decision apart from the channel condition information of a user. The Max CIR scheme schedules users with good channel conditions which leads to the starvation of users with poor channel quality as can be seen in Figure 4.6(b). Moreover, the PF scheme schedules users based on their relative channel condition and hence provides better QoS when compared to the Max CIR scheme. However, it has been shown recently that the PF scheme gives preference to users with high variance in their channel conditions, and hence fails in allocating the system resources to all users in a fair manner especially when a greater number of users are connected to the cell. The DBS, however, achieves a higher fairness than both of the above mentioned NRT scheduling schemes as it prioritizes users based on their queuing delay which gives a higher scheduling probability to users with relatively poor channel conditions. Figure 4.6(b) shows lower fairness indices for both schemes at lower traffic loads due to the unpredictable data rates of FTP type applications which leads to a higher variance in throughput. Furthermore, the maximum tolerable queuing delay of NRT traffic classes (for the DBS scheme) can be set to a higher value in order to maintain their traffic characterization.



(a)



(b)

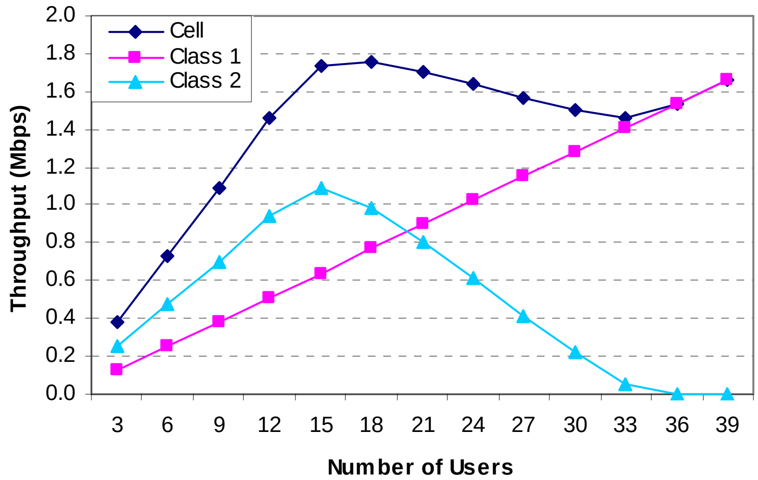
Figure 4.6: The performance evaluation of the DBS and other NRT scheduling schemes in the Pedestrian A environment
(a) Throughput (b) Jain Fairness Index

4.3.1.5 Performance Evaluation of the DBS for a Mixed Traffic Environment

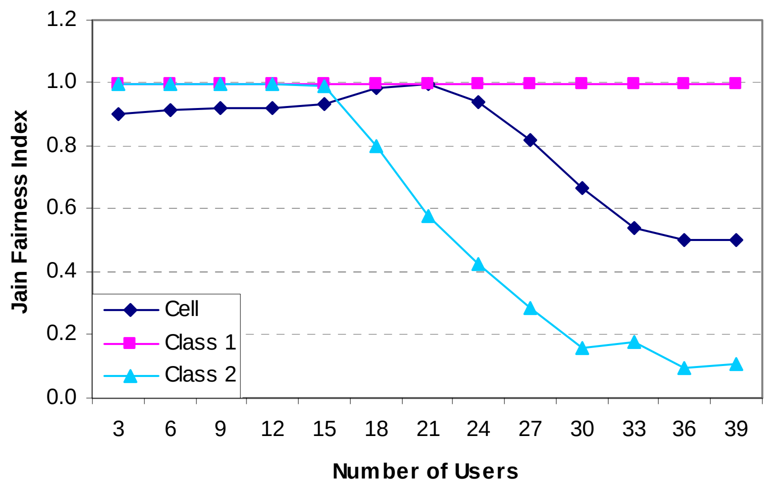
It was proved in Section 3.3.1.5 that the DBS can converge to a prioritized version of the Max CIR scheme in order to prioritize RT and NRT traffic applications without changing the characterization of their respective traffic class. Therefore, in order to show this ability and study its effect on the overall system performance, this section evaluates the DBS for a mixed traffic environment involving two traffic classes: the conversational class which is a RT traffic class with a data rate of 64 Kbps and the background class that is considered a NRT traffic class is emulated with an FTP application. The conversational class is the most delay sensitive traffic class and hence it is given a higher priority ($E_1 = 1$) than the delay tolerable background class ($E_2 = 0.7$). Moreover, the maximum threshold delay for the conversational class has been set to 2 s. Therefore, RT users are scheduled based on their channel condition, instantaneous queuing delay and their maximum delay threshold with the scheduling function shown in Equation (3.13). On the other hand, NRT users are assigned to the shared channel according to the function shown in Equation (3.17) which is based on a users' instantaneous channel condition as well as their QoS prioritization parameter. As discussed in Section 3.3.1.5, the DBS scheduling assignment for a particular class can converge to a prioritized version of the Max CIR scheme when $\gamma_c = 1$ and $\max\_delay_c \to \infty$. Therefore, in the following simulation, $\gamma_2 = 1$ and $\max\_delay_2 = 205$ s which is a relatively high queuing delay considering that the simulation runs for 200 s. These simulation parameters are catalogued in Table A.4.

The performance of the DBS for a mixed traffic environment is shown in Figure 4.7. In this traffic environment, the DBS converges to the Max CIR scheme for the scheduling assignment of NRT traffic while maintaining a delay based prioritization for RT traffic. Figure 4.7(a) illustrates the overall cell throughput as well as the throughputs of the individual traffic classes for an increasing number of users where class 1 represents the RT traffic class and class 2 symbolizes the NRT traffic class. It was remarked in Section 4.1.2 that the data rate of FTP applications at any time interval depends on the overall network load as these applications require the confirmation of previously sent packets for the transmission of new PDUs. For this reason, the throughput of NRT traffic decreases as the number of users in the system increases beyond 15, as shown in the figure. Moreover, since in this kind of DBS scheduling assignment, NRT users are prioritized based on their channel conditions only, their scheduling priority is lower than that of RT users at higher system delay (which is due to higher system loads). In addition, RT traffic is modeled through a CBR application which accounts for its linear increase in throughput as the number of users in the system increases. Figure 4.7(b) shows that if the DBS is converged to the Max CIR algorithm for NRT traffic while maintaining the QoS requirements of RT traffic then NRT users are scheduled in an unfair manner when compared to the RT users. This results in a lower system JFI value.

It was shown through the above discussion that the convergence of the DBS to the Max CIR algorithm for NRT traffic (while utilizing its original scheduling assignment for RT traffic) results in the starvation of this traffic class as RT users are given a priority due to their relative queuing delays.



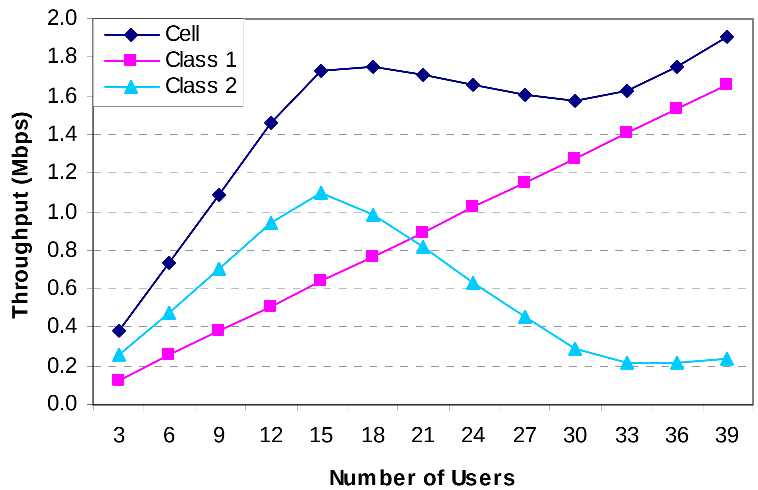
(a)



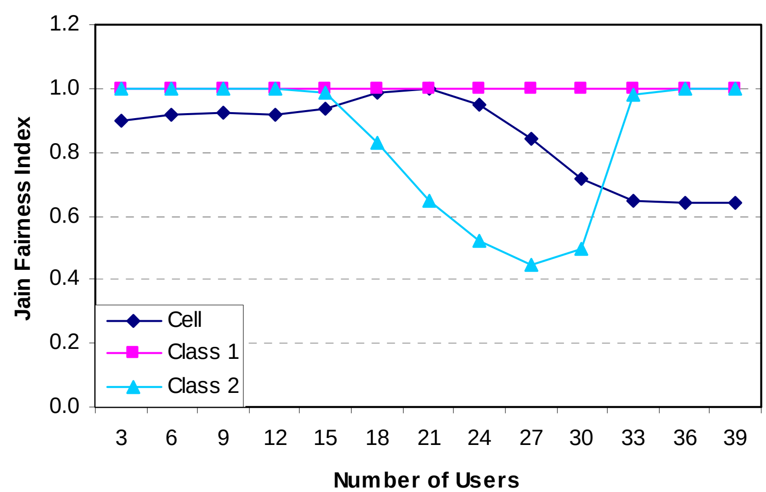
(b)

Figure 4.7: The Performance of the DBS in the Pedestrian A environment for a Mixed Traffic Load when the scheme converges to the Max CIR algorithm for NRT traffic.
(a) Throughput (b) Jain Fairness Index

In Figure 4.8, the performance of the DBS for the mixed traffic environment is shown when high delay constraints are set for NRT traffic applications. A $\max\_delay_2$ value of 25 s was set for the NRT traffic class which resulted in an increased average throughput for this class (shown in Figure 4.8(a)). Setting higher delay constraints for NRT applications does not necessarily change the characterization of this traffic class as these applications have a finite response time. In contrast with Figure 4.7(a) the throughput of class 2 is higher in this scenario for higher traffic loads. This prevents the starvation of NRT traffic as well as increases the overall cell throughput and user fairness as shown in Figure 4.8(b).



(a)



(b)

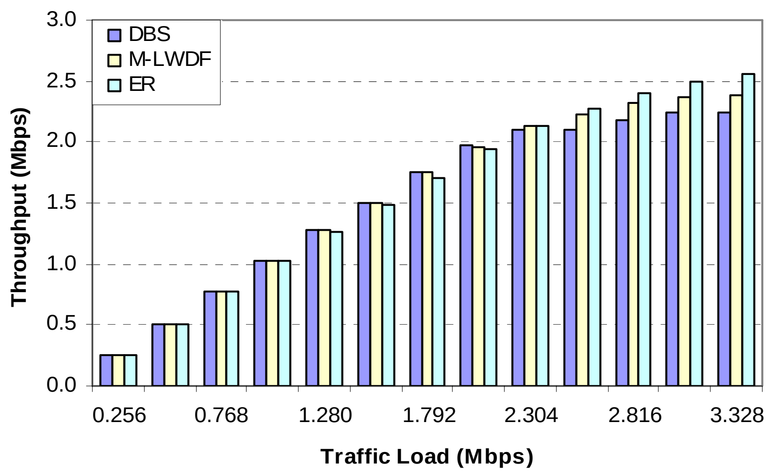
Figure 4.8: The Performance of the DBS in the Pedestrian A environment for a Mixed Traffic Load when delay constraints are set for NRT traffic
(a) Throughput (b) Jain Fairness Index

The convergence of the DBS scheme to the Max CIR scheme for the purpose of scheduling RT and NRT traffic classes simultaneously leads to the starvation of the NRT class especially if this class is given a lower QoS priority. Hence, future work will involve the modification of the scheme so that NRT traffic is given an equivalent scheduling probability in the event of scheme convergence.

4.3.2 Vehicular A Environment

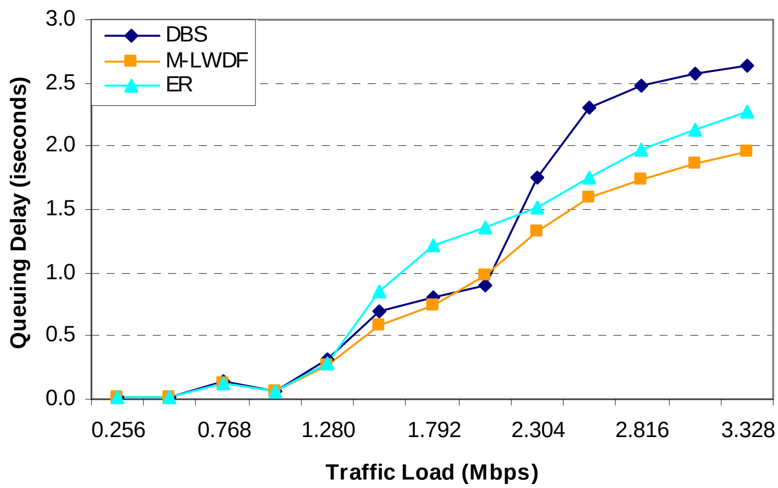
This section evaluates the performance of the DBS against the M-LWDF and ER scheduling schemes in the Vehicular A environment. The scheduling parameters used in this section are the same as those defined for the above mentioned scheduling schemes in the Pedestrian A environment which was discussed in Section 4.3.1.3.

Figure 4.9 illustrates the performance of the DBS, M-LWDF and ER schemes in the Vehicular A environment. As can be seen in Figure 4.9(a), the throughput of the three schemes is approximately the same up to a traffic load of 1.024 Mbps. Between the traffic loads of 1.28 Mbps and 2.048 Mbps, the DBS (1.278, 1.507, 1.752, 1.978 Mbps) outperforms the M-LWDF (1.275, 1.502, 1.746, 1.963 Mbps) and ER (1.27, 1.477, 1.699, 1.946 Mbps) schemes in terms of throughput. However, when the traffic load exceeds 64% of the load supportable by the system, the performance of the DBS starts declining. As discussed in previous sections, this can be accounted by cell congestion. Moreover, in the Vehicular A environment, users have high fluctuations in channel condition due to higher mobility as well as a lower overall channel quality because of the larger cell assigned for this environment (see Appendix A for environment parameters). This results in a lower overall throughput in the Vehicular A environment when compared to the Pedestrian A environment. For instance, a traffic load of 3.328 Mbps resulted in a throughput of 3.139, 3.146 and 3.203 Mbps by the DBS, M-LWDF and ER schemes respectively in the Pedestrian A environment (shown in Figure 4.5(a)) whereas a throughput of 2.245, 2.383 and 2.56 Mbps was achieved by these schemes, for the same traffic load, in the Vehicular A environment. Figure 4.9(b) illustrates that the ER scheme maintains the lowest overall queuing delay when compared to the DBS and M-LWDF scheduling schemes, in the Vehicular A environment. This figure confirms that the performance of the DBS degrades at higher traffic loads in the Vehicular environment. There are two reasons for this problem: firstly, as the traffic load increases, the average queuing delay experienced by the system also increases since the same amount of resources are assigned to a greater number of users. Secondly, at higher queuing delays, the dependence of the channel condition on the scheduling assignment is reduced. Therefore, users with good channel conditions but lower instantaneous queuing delays are given lesser preference than users with poor channel conditions but higher queuing delays which compromises the opportunity of increasing overall system throughput. Furthermore, higher queuing delays also leads to an increased packet drop rate which can be seen in Figure 4.9(c). RT scheduling schemes fairly assign users to the HS-DSCH in the Vehicular A environment as shown in Figure 4.9(d), although the JFI tends to be lower at higher traffic loads due to an increased variance in user throughput when compared to the Pedestrian A environment (Figure 4.5(d)).



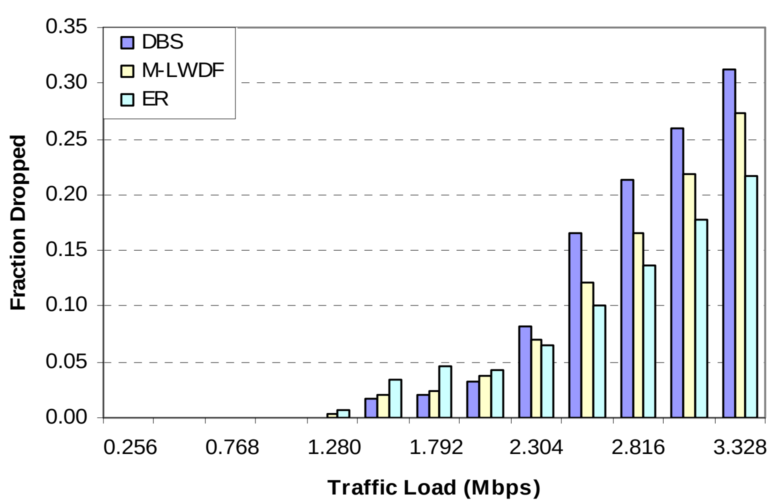
DBS
M-LWDF
ER
Throughput (Mbps)
Traffic Load (Mbps)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.256
0.768
1.280
1.792
2.304
2.816
3.328

(a)



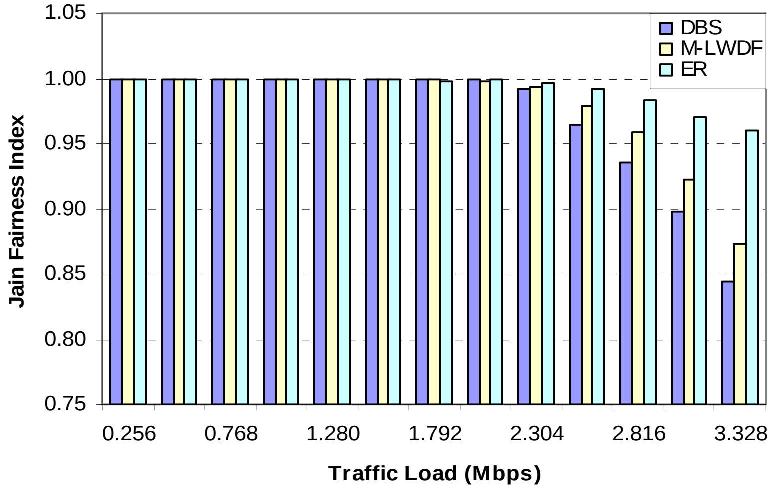
DBS
M-LWDF
ER
Queuing Delay (iseconds)
Traffic Load (Mbps)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.256
0.768
1.280
1.792
2.304
2.816
3.328

(b)



DBS
M-LWDF
ER
Fraction Dropped
Traffic Load (Mbps)
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.256
0.768
1.280
1.792
2.304
2.816
3.328

(c)



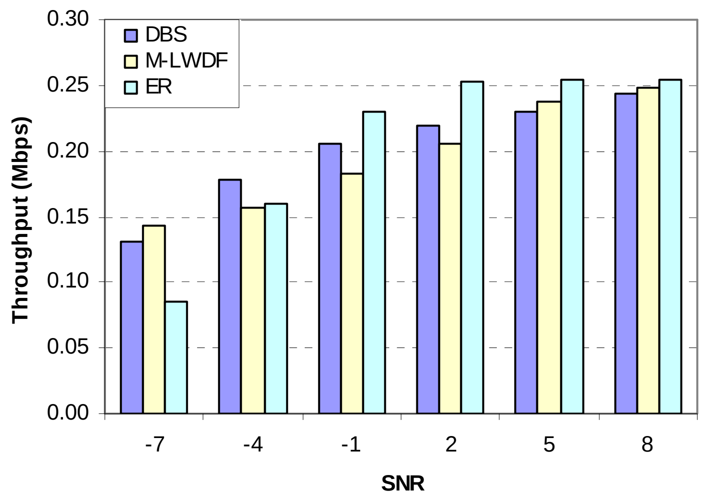
(d)

Figure 4.9: The Performance of RT scheduling schemes in the Vehicular A environment for varying traffic Loads.
(a) Throughput (b) Queuing Delay (c) Fraction Dropped (d) Jain Fairness Index

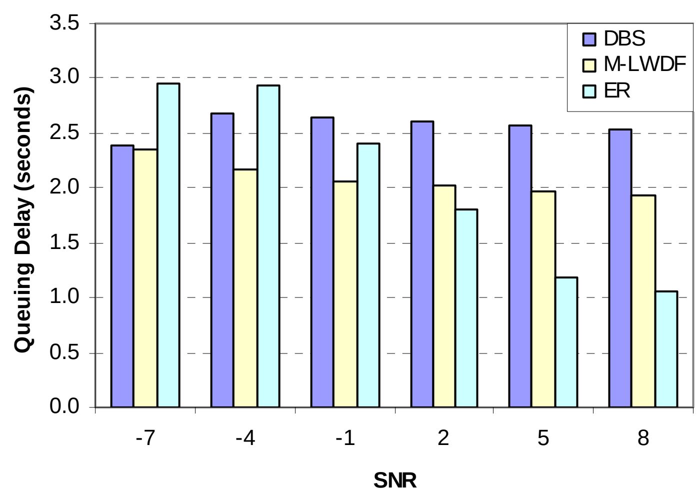
4.3.3 Fixed Channel Environment

RT scheduling schemes are evaluated in this section in a fixed channel environment with six SNR values viz. -7, -4, -1, 2, 5 and 8 dB where the CQI is calculated using Equation (3.19). This simulation is emulated with a traffic load of 1.536 Mbps with 18 users out of which 0.768 Mbps are contributed by 12 conversational class users (with a data rate of 64 Kbps) and the remaining load is supplied by 6 streaming users (with a data rate of 128 Mbps). Each SNR value is assigned to 4 conversational class users and 2 streaming data flows. Results are collected for each SNR value such that the performance metrics for one SNR value is evaluated separately from the others. This demonstrates the performance of users based on their channel condition. It should be noted that for this simulation users are prioritized based on the delay sensitivity of their traffic class.

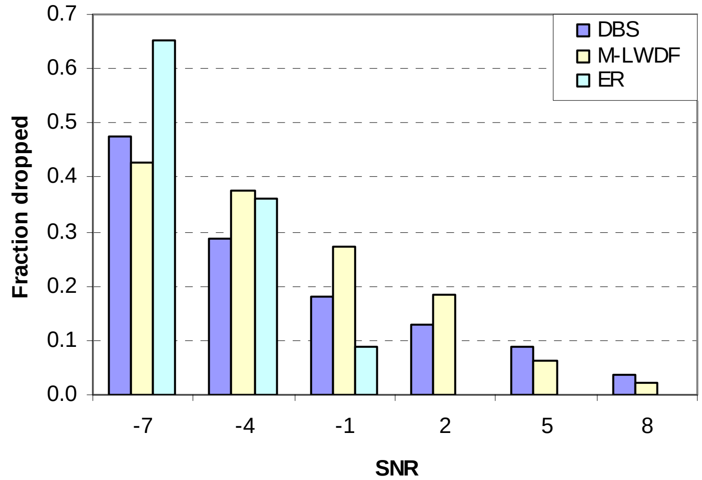
The performance of the DBS, M-LWDF and ER schemes is compared in a fixed channel environment for RT traffic and the results are illustrated in Figure 4.10. Users with an SNR value of 8 dB enjoy a higher average throughput when compared to users with worse channel conditions, as shown in Figure 4.10(a). For example, an SNR value of 8 dB achieves a throughput of 0.243, 0.247 and 0.254 Mbps for the DBS, M-LWDF and ER schemes respectively whereas a lower SNR value of -7 dB attains a throughput of 0.13, 0.142 and 0.085 Mbps for the same schemes. This is because better channel conditions allow greater data rates in a fixed time frame. Moreover, the overall accumulative throughput of the three schemes at this traffic load is similar but the ER scheme achieves a higher throughput for better channel conditions when compared to the DBS and M-LWDF schemes which compromises the throughput of users with poor channel conditions. Figure 4.10(b) shows the average queuing delay of the three schemes which are classified based on SNR and it can be seen that the DBS maintains a queuing delay of approximately 2.6 s for all SNR values whereas the M-LWDF and ER schemes have lower queuing delays for higher SNR values as a greater number of bits per second are transmitted at these SNR values. Figure 4.10(c) shows the fraction of packets dropped at various channel conditions. Since, all three RT schemes utilize the channel condition information of a user along with their data flows queuing delay in the scheduling assignment, it can be seen in the figure that users with poor channel conditions have a higher drop rate with the ER scheme having the highest drop rate at lower SNR values (which goes in concordance with Figure 4.10(a)). From the above discussion it can be concluded that the DBS schedules users in a fair manner when compared to the M-LWDF and ER schemes as it provides similar average throughput and delay to users regardless of their channel condition which is confirmed in Figure 4.10(d). The figure shows that the ER scheme has the highest variance in throughput and thus provides a lower fairness index of 0.83 when compared to the M-LWDF that has a JFI value of 0.89 and the DBS with the highest JFI of 0.9 (for a traffic load of 1.536 Mbps).



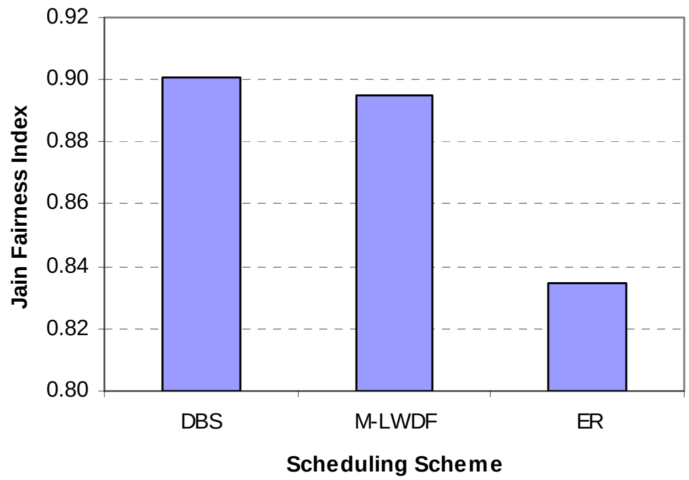
(a)



(b)



(c)



(d)

Figure 4.10: Performance of RT scheduling schemes in the fixed channel environment (a) Throughput (b) Queuing Delay (c) Fraction dropped (d) Jain Fairness Index

It should be noted that the results obtained through all of the above simulations have a 95% confidence level with 5% confidence intervals. The confidence interval has been calculated using the *t-distribution* (see reference [32]).

4.4 Summary

In this chapter, we evaluate the performance of the proposed DBS scheme and show, through simulations, that it achieves the design objectives of a successful HSDPA scheduling scheme that were defined in Chapter 3. Moreover, the scheme can balance the tradeoff between throughput maximization and the minimization of queuing delay by the adjustment of a parameter known as γ_c. We also found that increasing the delay thresholds of RT traffic classes improves overall system throughput at the cost of higher average queuing delay. The ability of the DBS to drop packets when they reach their delay threshold maintains the delay requirements of RT traffic classes. Furthermore, we found that if the variation in the QoS parameters is high then, depending upon on the traffic load, the overall system performance may be compromised as the effect of QoS prioritization overrides that of queuing delay. RT traffic applications are delay sensitive and thus scheduling schemes designed for these applications define certain delay requirements for them. But NRT traffic applications are delay tolerable and the scheduling of their PDUs through the definition of delay requirements changes their traffic characterization. The DBS can maintain the traffic characterization of NRT applications while scheduling RT traffic based on their queuing delay through the convergence of the scheme to the Max CIR algorithm for NRT traffic classes. Therefore, the DBS is capable of scheduling RT as well as NRT traffic classes simultaneously. However, this caused the starvation of NRT traffic classes as the dependence of the DBS scheduling function increases as packets approach their delay deadline whereas the Max CIR scheme is not affected by delay. For this reason, if a large threshold delay is assigned to the NRT class then their performance is slightly increased which improves overall system performance. Finally, we show that the DBS is a fair scheduling algorithm through the measurement of the variance in user throughput known as the JFI.

Chapter 5

Conclusions and Future Work

The Delay Based Scheduler (DBS) for High Speed Downlink Packet Access (HSDPA) is proposed in this thesis. Conventional HSDPA schedulers such as Max CIR and PF utilize the channel condition information of a user for the scheduling decision. Scheduling users with good channel conditions increases overall system throughput as users with better radio conditions are capable of receiving a higher number of bits per second in a given time interval. However, this results in the starvation of users with poor relative channel conditions which in-turn leads to an unfair assignment of the shared channel. Moreover, these schemes are not suitable for Real Time (RT) applications as they do not account for the instantaneous queuing delay of the respective data flows and consequently cannot maintain the Quality of Service (QoS) requirements of these delay sensitive applications. Furthermore, an HSDPA scheduling algorithm should be able to schedule packets belonging to RT and Non-Real-Time (NRT) applications simultaneously without changing the traffic characterization of these classes. For this reason, six design objectives were defined for a successful HSDPA packet scheduler: throughput maximization, minimization of queuing delay, fairness, maintaining delay thresholds, QoS prioritization and scheduling RT and NRT traffic classes simultaneously.

The DBS implements the above mentioned design objectives by utilizing the short-term channel condition information of a user as well as the instantaneous queuing delay of their data flows' head packet in the calculation of the scheduling priority. Maximum delay thresholds have been defined on a per class basis so that packets whose queuing delay exceeds their delay limit are dropped. This ensures that the delay constraints of RT applications are maintained. In order to minimize the average queuing delay of the system, the DBS has been designed such that as the users' queuing delay approaches its deadline, the dependence of the scheduling function on their queuing delay increases which conversely decreases its dependence on their channel quality. Therefore, it can be inferred that there exists a tradeoff between throughput maximization and the minimization of queuing delay. Nevertheless, the DBS can accommodate both these properties with the attunement of a parameter known as γ_c. Additionally, the DBS improves fairness as users with poor channel conditions have a higher probability of being scheduled as the queuing delay experienced by them increases. Furthermore, users choose from a range of services provided by the cellular network based on their requirements. The service provider may then prioritize users depending upon their QoS requirements. The DBS is capable of providing QoS prioritization as it defines a QoS parameter for each user and utilizes these parameters for the scheduling assignment. Finally, we mathematically show that the proposed scheme can converge to the Max CIR algorithm under certain conditions and hence can inter-operate between RT and NRT applications without changing the characterization of their respective traffic classes.

Simulation results revealed that the DBS achieves a lower queuing delay compared to the M-LWDF and ER schemes in the Pedestrian A environment while maintaining similar system throughput and fairness. Moreover, we found that γ_c is capable of reducing the average queuing delay of the system at the cost of lower throughput and thus it can be adjusted in order to achieve the desired system performance. We also found that the DBS out-performed the Max CIR and PF schemes in terms of throughput and fairness when certain delay thresholds are defined for NRT traffic classes. This is expected since the proposed scheme considers the instantaneous queuing delay of data flows in the scheduling assignment unlike NRT schemes. However, NRT traffic is delay tolerable and hence assigning delay thresholds for these traffic classes changes their traffic characterization. The DBS can address this problem by converging to the Max CIR scheme when RT and NRT traffic classes coexist in the cell. A mixed traffic environment was emulated and we found that the convergence of the DBS to the Max CIR scheme for scheduling NRT users led to the starvation of these users as a delay based higher priority is assigned to RT users at higher traffic loads. For this reason, if a large threshold delay is defined for the NRT traffic classes then their throughput can be increased which consequently improves overall system performance. It should be noted that setting a high delay threshold for NRT traffic does not necessarily change its traffic characterization as these applications have a certain finite response time. We also show that the effect of QoS prioritization overrides the collective effects of queuing delay and channel condition in the DBS algorithm and thus a high variation in the QoS parameters may compromise system performance depending upon on the traffic load. Furthermore, the performance of the system degraded in the Vehicular A and fixed channel environments for all scheduling schemes. This is because of the fluctuation in the signal strength due to user mobility in the Vehicular A environment and the large difference in the channel condition experienced by users in the fixed channel environment.

There are some limitations with the present implementation of the DBS scheme in that there is a need to support multiple data connections per user, as well as handoff management and Call Admission Control (CAC) procedures. In the future, the DBS would be adapted so that each user can support multiple network applications simultaneously. This will be made possible with the help of multiple queues at the Node B for every user where each queue will be assigned to a single traffic application. Moreover, a handoff management algorithm will be implemented to decide when an ongoing call can be transferred to another base station as the user moves from the coverage area of one cell to that of another cell. Furthermore, a CAC mechanism will be deployed in order to determine the level of acceptable traffic load in the cell such that certain minimum service requirements of users are guaranteed. The CAC mechanism will be able to control the performance of ongoing calls by preventing the acceptance of new or handoff users to the cell. In addition, the CAC scheme can be used to determine the optimal value of γ_c for a particular traffic load and thus this parameter can be dynamically adjusted as the system load changes. The QoS parameter, E_i, of a user can also be changed dynamically in order to increase the priority of that user such that the overall fairness of the system is enhanced. The DBS can presently support RT and NRT traffic applications simultaneously by the convergence of the scheme to the Max CIR algorithm. However, simulation results show that this leads to an unfair assignment of system resources as users running NRT applications were given a lower scheduling priority at higher queuing delays as the Max CIR scheme is delay independent. This problem will be addressed in the future by the appropriate adjustment of the DBS scheduling function.

References

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

[2] 3GPP TS 25.308, "High Speed Downlink Packet Access (HSDPA); Overall Description", Release 5, March 2003.

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

[4] 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.

[5] M. Andrews et al., "Providing QoS over a Shared Wireless Link", *IEEE Communications Magazine*, volume 39, pp. 150-154, February 2001.

[6] S. Shakkottai and A. Stolyar, "Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR", *Proceedings of the 17th International Teletraffic Congress (ITC-17),* Salvador de Bahia, Brazil, September 2001.

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

[8] Textronics, "UMTS Protocols and Protocol testing", Available: http://www.tek.com/Measurement/App_Notes/2F_14251/eng/, June 2007.

[9] H. Kaaranen, A. Ahtiainen, L. Laitinen, S Naghian and V. Niemi, "*UMTS Networks, Architecture, Mobility, and Services*", John Wiley & Sons, 1st edition, 2001.

[10] 3GPP TS 25.331, "Radio Resource Control (RRC); Protocol Specification", version 3.11.0, Release 1999, June 2002.

[11] 3GPP TS 25.322, "Radio Link Control (RLC); Protocol Specification", version 3.11.0, Release 1999, June 2002.

[12] 3GPP TS 25.321, "Medium Access Control (MAC); Protocol Specification", version 7.2.0, Release 7, September 2006.

[13] 3GPP TS 25.302, "Services provided by the Physical Layer", version 5.1.0, Release 5, June 2002.

[14] 3GPP TS 23.107, "QoS Concept and Architecture", version 4.4.0, Release 4, March 2002.

[15] 3GPP TS 25.321, "Medium Access Control (MAC); Protocol Specification", version 7.2.0, Release 7, September 2006.

[16] 3GPP TS 25.877, "High Speed Downlink Packet Access: Iub/Iur protocol aspects", version 5.1.0, Release 5, June 2002.

[17] T. Kolding, K. Pedersen, J. Wigard, F. Frederiksen and P. Mogensen, "High Speed Downlink Packet Access: WCDMA Evolution", *IEEE Vehicular Technology Society News,* volume 50, pp. 4-10, February 2003.

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

[19] H. Kim, K. Kim, Y. Han and J. Lee, "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'02),* volume 5, pp. 2244–2248, September 2002.

[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*, volume 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*, volume 3, pp. 652-656, 2002.

[22] Y. M. Ki and D. K. Kim, "Packet Scheduling Algorithms for Throughput Fairness and coverage Enhancement in TDD-OFDMA Downlink Network", *Proceedings of the Institute of Electronics, Information and Communication Engineers (IEICE) Transactions on Communications*, volume E-88B, no. 11, pp. 4402-4405, November 2005.

[23] M. Andrews et al., "CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions", *Bell Labs Technical Memo*, April 2000.

[24] P. Ameigeiras, J. Wigard and P. Mogensen, "Performance of the M-LWDF Scheduling Algorithm for Streaming Services in HSDPA", *Proceedings of the IEEE VTC 2004*, volume 7, pp. 999-1003, 2004.

[25] L. C. Wang and M. C. Chen, "Comparisons of link-adaptation-based scheduling algorithms for the WCDMA system with high-speed downlink packet access", *Canadian Journal of Electrical and Computer Engineering*, volume 29, no. 1/2, pp. 109-116, 2004.

[26] K. Khawam, "The Modified Proportional Fair Scheduler", *Proceedings of the 17th IEEE International Symposium on PIMRC'06*, pp. 1-5, September 2006.

[27] Enhanced UMTS Radio Access Network Extensions for NS2, Available: http://www.ti- wmc.nl/eurane/, June 2007.

[28] Deliverable D3. 2v2, "End-to-end Network Model for Enhanced UMTS", Available: http://www.ti-wmc.nl/eurane/, June 2007.

[29] Network Simulator 2, Available: http://www.isi.edu/nsnam/ns/, June 2007.

[30] 3GPP TS25.214, "Physical Layer Procedures", version 5.5.0, Release 5, June 2003.

[31] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems", *DEC Research Report TR-301*, September 1984.

[32] M. Abramowitz and I. A. Stegun, "*Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*", Dover, 1964.

Appendix A

Simulation Parameters

In this Appendix, we present the values of the simulation parameters which are used for the performance evaluation of the DBS and other conventional HSDPA algorithms.

A.1 Traffic Model Parameters

Four traffic classes are used for the simulations in Chapter 4 which comprise of the conversational, streaming, interactive and background classes. The traffic model parameters that were discussed for these application classes in Section 4.1.2 are shown in Table A.1.

Conversational Traffic Parameters	
Distribution	Exponential
Traffic Rate	64 kbps
Burst Time	900 ms
Idle Time	100 ms
Packet Size	320 bytes
Streaming Traffic Parameters	
Traffic Encoding	CBR
Traffic Rate	128 kbps
Packet Size	320 bytes
Interactive Traffic Parameters	
Traffic Type	CBR
Traffic Rate	64 kbps
Packet Size	320 bytes
Background Traffic Parameters	
Traffic Application	FTP
Packet Size	320 bytes

Table A.1: The Traffic Model Parameters

A.2 Propagation Model Parameters

Table A.2 illustrates the propagation loss model parameters. It should be noted that these parameters are relevant for the Pedestrian A and Vehicular A environments as the channel conditions of users are fixed in the fixed channel environment.

Parameter	Value
Node B Transmission Power	38 dBm
Node B Antenna Gain	17 dBi
Intra Cell Interference	30 dBm
Inter Cell Interference	-70 dBm
Distance Loss at 1 km	1.374e2
Path Loss Exponent	3.52
Mobile Speed for the Pedestrian A environment	3 Km/hr
Cell Diameter for the Pedestrian A environment	500 m
Mobile Speed for the Vehicular A environment	60 Km/hr
Cell Diameter for the Vehicular A environment	1000 m

Table A.2: The Propagation Loss Model Parameters

A.3 DBS Parameters

All the relevant DBS parameter settings are shown in Table A.3.

Parameter	Value
$\max\_delay_1$	2 s
$\max\_delay_2$	4 s
E_i for Class 1 users	1.0
E_i for Class 2 users	0.7
γ_c for Class 1 and Class 2	10

Table A.3: The DBS Parameters

A.4 Mixed Traffic Environment Parameters

The mixed traffic environment comprises of conversational and background applications. The DBS was converged to the Max CIR algorithm for NRT traffic classes while the scheme maintained a delay based prioritization for RT traffic. The DBS parameters involved in this traffic scenario are shown in Table A.4.

Parameter	Value
$\max\_delay_1$	2 s
$\max\_delay_2$	205 s
E_i for Class 1 users	1.0
E_i for Class 2 users	0.7
γ_1	10
γ_2	1

Table A.4: The DBS Parameters for a Mixed Traffic Environment

A.5 System Parameters

The system and other parameters are shown in Table A.5.

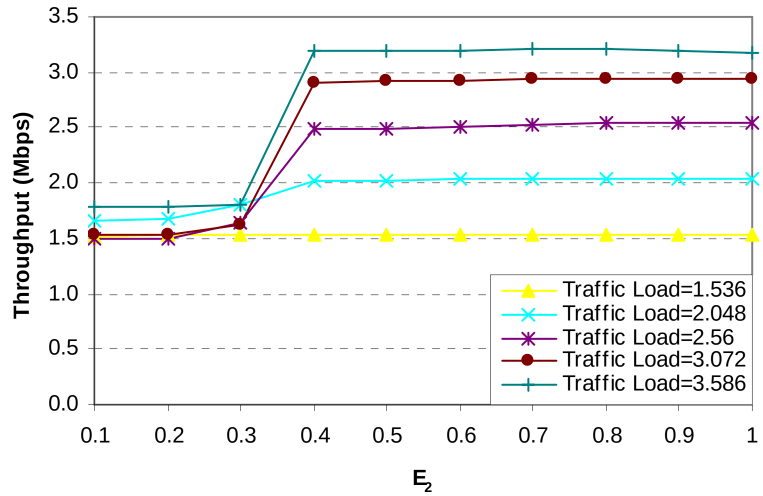
Parameter	Value
Simulation Time	200 s
TTI	2 ms
Distribution of Users in the Cell	Uniform
Minimum CQI	0
Maximum CQI	30
Number of Cells	1

Table A.5 System Parameters

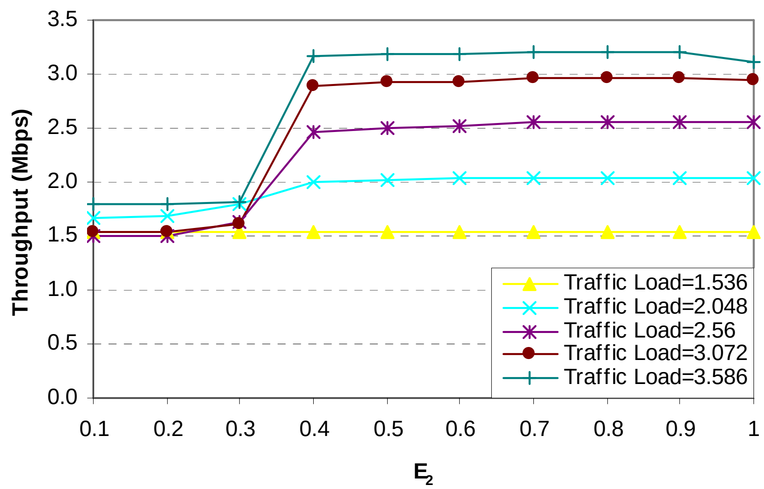
Appendix B

Auxiliary Simulation Results

The effect of varying DBS parameters was illustrated through simulations in Section 4.3.1.1. It was remarked that increasing the threshold delay of a traffic class reduces the fraction of packets dropped which consequently improves system throughput at the cost of higher queuing delays. These results are shown in Figures B.1, B.2 and B.3. Figure B.1(a) and Figure B.1(b) illustrate the system throughput for a threshold delay value of 4 s and 8 s respectively. It can be seen from the figures that increasing the threshold delay of a traffic class increases the overall throughput of the system by a small amount. However, this improvement in throughput is less than 3% of the aggregate value. Similarly, Figure B.2(a) and Figure B.2(b) show the effect of doubling the threshold delay on the average queuing delay of the system. Clearly, it can be seen that this queuing delay is almost doubled as the threshold values are increased.



(a)

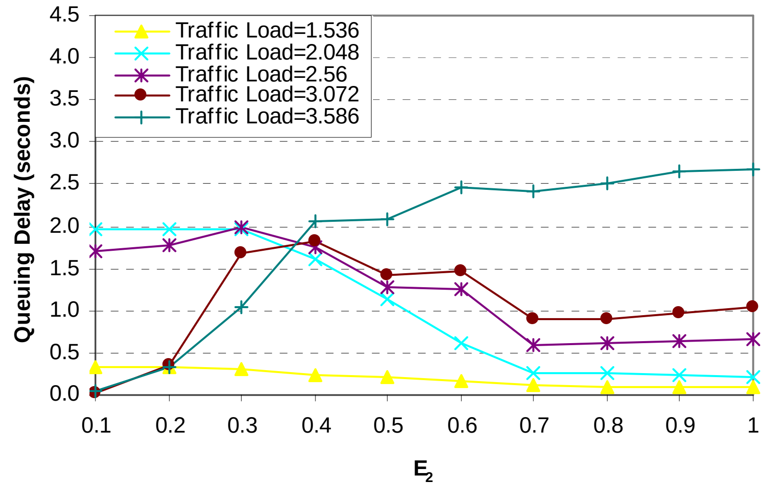


(b)

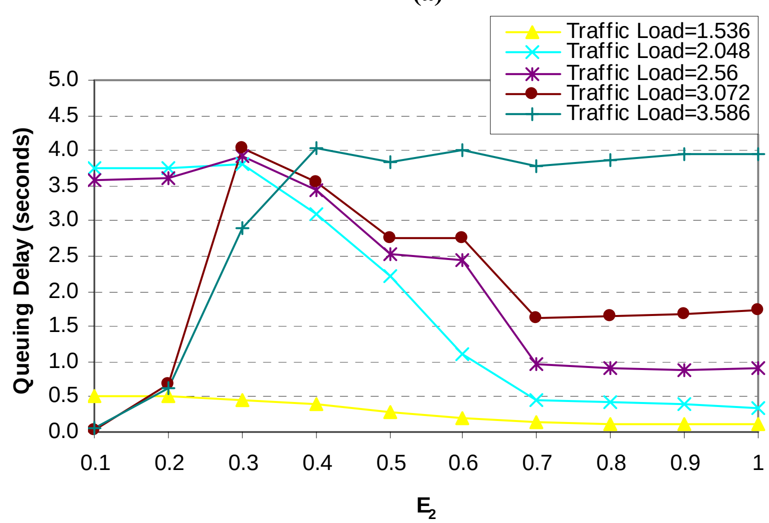
Figure B.1: The effect of Threshold Delays on Throughput for the DBS scheme in the RT traffic environment for varying Traffic Loads
(a) Maximum Threshold Delay of 4 s (b) Maximum Threshold Delay of 8 s

The enhancement in system throughput due to the increase in threshold delay is confirmed by the decrease in drop rate that is shown in Figure B.3(a) and Figure B.3(b). It can be seen from the figures that the fraction of packets dropped clearly decreased for a traffic load of 3.586 Mbps when the maximum threshold delay is doubled from 4 s to 8 s.

Nevertheless, this improvement in system performance is only by a small value which is less than 2%.



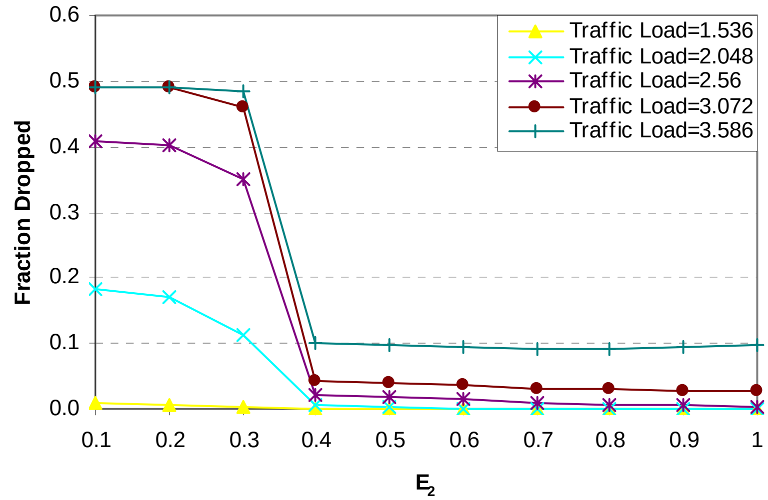
(a)



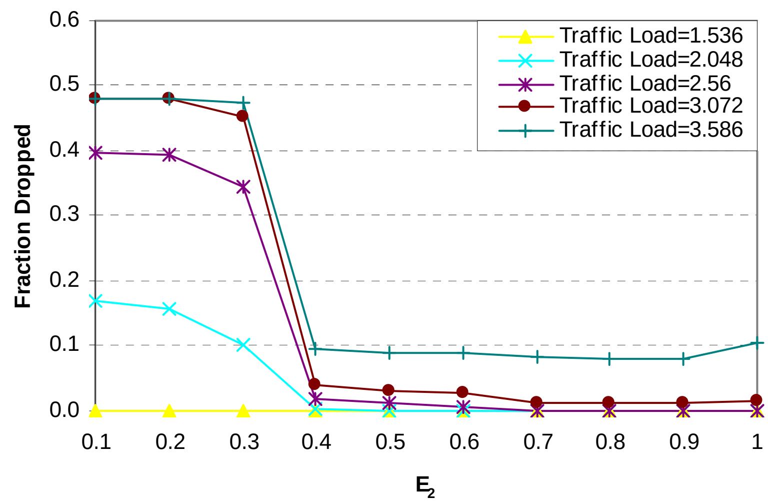
(b)

Figure B.2: The effect of Threshold Delays on average Queuing Delay for the DBS scheme in the RT traffic environment for varying Traffic Loads
(a) Maximum Threshold Delay of 4 s (b) Maximum Threshold Delay of 8 s

From the above discussions it can be concluded that increasing the delay limit of a traffic class improves system performance in terms of throughput and consequently fraction dropped by a small amount but at the cost of a considerable increase in the average queuing delay of the system.



(a)



(b)

Figure B.3: The effect of Threshold Delays on the fraction of packets dropped for the DBS scheme in the RT traffic environment for varying Traffic Loads
(a) Maximum Threshold Delay of 4 s (b) Maximum Threshold Delay of 8 s