Packet Scheduling and Quality of Service in HSDPA

Gutiérrez, Pablo Jose Ameigeiras

Publication date:
2003

Document Version
Publisher's PDF, also known as Version of record

Packet Scheduling And Quality of Service in HSDPA

Pablo José
Ameigeiras Gutiérrez

Ph. D. Thesis
October 2003

R 03 1016
ISSN 0908 1224
ISBN 87 90834 38 0

Department of Communication Technology
Institute of Electronic Systems, Aalborg University
Niels Jernes Vej 12, DK-9220 Aalborg Øst, Denmark

Abstract

Data services are anticipated to have an enormous rate of growth over the next years and will likely become the dominating source of traffic load in 3G mobile cellular networks. The provision of these data services in 3G systems requires spectral efficiency solutions to increase the capacity of the radio access network as well as support for high data rates. In this context, the 3GPP introduces a new feature in the Release 5 specifications denominated High Speed Downlink Packet Access (HSDPA). The target of the HSDPA concept is to increase the peak data rates, improve the quality of service, and enhance the spectral efficiency for downlink packet traffic.

In the early stages of this Ph.D. thesis, the focus concentrated on building understanding on the services that could potentially be conveyed by HSDPA: streaming, interactive, and background services. Relevant characteristics such as the QoS demands of the different traffic classes, the protocols employed in their communication, and the statistical properties of the traffic are investigated. In the case of streaming services, special attention is paid to Constant and Variable Bit Rate encoding and its interaction with QoS bearer attributes such as the Guaranteed Bit Rate. In the case of NRT services, the main interest is in HTTP and its characteristics because Web traffic is expected to be one of the key services in 3G networks.

Since services built over a TCP/IP platform are expected to accumulate a large share of the overall traffic volume in UMTS, the performance of TCP is to be investigated. Connection characteristics such as the downloaded file size, the TCP memory capabilities at the receiver, the round trip time, or the allocated bit rate have strong impact on key performance aspects for both the end-user (e.g. the service response time) and the system capacity (e.g. potential code shortage).

The HSDPA concept appears as an umbrella of features to improve both user and system performance. The concept includes a reduction of the Transmission Time Interval (TTI) to 2 ms, an Adaptive Modulation and Coding (AMC) scheme, and a fast physical layer Hybrid ARQ mechanism. Moreover, the Packet Scheduler functionality is located in the Node B, which enables to track instantaneous variations of the user's channel quality. The investigation has specially concentrated on assessing different Packet Scheduling methods with different degrees of fairness for non delay-sensitive services because it determines the QoS perceived by the users. The results have shown that under minimum throughput guarantees the Proportional Fair algorithm provides the highest cell capacity of all tested algorithms. With this algorithm, HSDPA provides a cell throughput gain of around 150% over WCDMA for a best effort service. The provision of user throughput guarantees of 64, 128 and 384 kbps (at 5% outage) incurs on a HSDPA cell throughput reduction respectively of around 10, 30 and 80% relative to the best effort case. Further analysis of TCP traffic on HSDPA has shown the inherent trade off between cell throughput and Node B queuing delay, and the potentially harmful effect of this queuing on the RLC retransmission process.

Furthermore, the performance of streaming services over HSDPA has been investigated for Constant Bit Rate flows. The Modified Largest Weighted Delay First (M-LWDF) has been selected as the Packet Scheduling algorithm for delay-sensitive traffic because it includes both the instantaneous channel quality and the queuing delay in the user's priority computation. Streaming connections with 128 kbps only represent a minor cell capacity reduction compared to best effort NRT traffic using the Proportional Fair algorithm with a minimum data rate of 128 kbps. The M-LWDF yields an unfair QoS provision among the users in the cell, but other scheduling methods with larger degree of fairness have failed to support larger system capacity when maximum transfer delay requirements are imposed.

Dansk Resumé

Translation: Christian Rom and Rikke Jensen.

Datatjenester forventes at have en enorm vækstrate i løbet af de næste år og vil sandsynligvis blive til en dominerende trafik belastningskilde i 3G mobile netværk. Angivelsen af disse datatjenester i 3G systemer kræver spectrale effektive løsninger for at øge kapaciteten af radio acces netværk såvel som support for høje datarater. I denne sammenhæng introducerer 3GPP et nyt koncept kaldet High Speed Downlink Packet Access (HSDPA) i sin femte specifikationsudgave. Målet for HSDPA er at øge spids data rater, forbedre servicekvaliteten og forøge spektraleffektiviteten for downlink pakketrafik.

I starten af denne Ph.D. these var fokus centreret omkring opbygning af forståelse angående de tjenester, som vil blive overført med HSDPA: Streaming-, Interaktive- og Baggrundstjenester. Relevante karakteristikker såsom; QoS behov af de forskellige trafikklasser, de protokoler der bliver anvendt i deres kommunikation og de statistiske egenskaber af trafikken, bliver undersøgt. I tilfældet af Streaming tjenester, er opmærksomhed specielt givet til Konstant of Variable Bit Rate indkodning og dens interaktion med QoS bærer attributer såsom den Garanterede Bit Rate. I tilfældet af NRT tjenester er hovedinteressen i HTTP og dens karakteristikker, da Webtrafiken forventes at blive en af nøgletjenesterne i 3G netværk.

Da tjenester byggede over at en TCP/IP platform forventes at akkumulere en stor del af den totale trafikvolume i UMTS, er performance af TCP stadig at udforske. Forbindelseskarakteristikker såsom størrelsen på den downloadede fil, TCP hukomelsesmulighederne ved modtageren, round trip tiden eller den allokerede bit rate, har en stor impakt på nøgleegenskaberne for både slutbrugeren (feks. tjeneste respons tid) og systemkapaciteten (feks. potentiel kode mangel).

HSDPA konceptet ser ud som en paraply af funktionaliteter til at forbedre både bruger og system performance. Konceptet inkluderer en reduktion af Transmission Tids Intervalet (TTI) til 2 ms, et Adaptivt Modulation og Kodnings (AMC) skema samt en hurtig fysisk Hybrid ARQ lager mekanisme. Derudover er Pakke fordeler funktionaliteten flyttet til Node B, som tillader at spore øjeblikkelige variationer af brugerens kanalkvalitet. Undersøgelsen har især været centreret omkring vurderingen af forskellige Pakke fordelingsmetoder med forskellige grader af retfærdighed for ikke forsinkelses følsomme tjenester, da de afgør den QoS som opfattes af brugerene. Resultaterne har vist, at under minimalt throughput garanti, giver Proportional Fair algoritmen en maksimal celle kapacitet af alle de testede algoritmer. Dens kapacitet giver omkring 150% forbedring over WCDMA teknologien med samme grad af retfærdighed som i macrocelle-omgivelser. Forsyningen af datarater som garanterer 64, 128, 384 kbps (ved 5% driftsstandsning), fremkommer ved en celle kapacitet reduktion af henholdsvis omkring 10, 30 og 80% relativt til kapaciteten. Videre analyse af TCP trafik af HSDPA har vist det inhærente kompromis mellem celle throughput og Node B kø-forsinkelse, og den potentielle skadelige effekt af denne kø-effekt på RLC retransmission processen.

Derudover er performance af Streaming tjenester over HSDPA blevet undersøgt med Konstant Bit Rater. Det ”Modified Largest Weighted Delay First” (M-LWDF) er blevet valgt som den Pakke fordelings algoritme for forsinkelses-følsom trafik, fordi den indeholder både øjeblikkelig kanal kvalitet og kø forsinkelse i brugerens prioritetsudregning. Streaming fobindelser med 128 kbps repræsenterer kun en mindre cellekapacitet-reduktion i forhold til ”best effort” NRT trafik, som bruger Proportional Fair algoritmen med en minimum datarate på 128 kbps. M-LWDF giver en uretfærdig QoS forsyning iblandt brugerne i en celle, men andre fordelings metoder med højere grader af retfærdighed har ikke kunnet give større system kapacitet når maximum overførsel forsinkelseskrav er givet.

Preface and Acknowledgements

This Ph.D. thesis is the result of a three and a half year research project carried out at the Center for Person Kommunication (CPK), Department of Communication Technology, Aalborg University (AAU). The thesis work has been accomplished in parallel with the mandatory courses and teaching/working obligations required in order to obtain the Ph.D. degree. The investigation has been conducted under the supervision of the Research Professor Ph.D. Preben E. Mogensen (Aalborg University) and the co-supervision of Ph.D. Jeroen Wigard (Nokia Networks).

This thesis investigates the provision of packet data services in the evolution of Third Generation cellular communications. Key aspects of the support of packet data services in cellular networks such as the Quality of Service and the system capacity are widely covered. The focus has concentrated on the so-called High Speed Downlink Packet Access (HSDPA) technology that has been included in the Release 5 of the UMTS Terrestrial Radio Access Network (UTRAN) specifications by the 3rd Generation Partnership Project (3GPP). In the early stages of this Ph.D. thesis, the project focused the investigation on the WCDMA technology, but the scope changed into the HSDPA technology due to the interest of the sponsor. The study is primarily based on computer simulations of a UMTS/HSDPA cellular network. It is assumed that the reader is acquainted with basic knowledge about the system level aspects of UMTS, packet switched services, and radio propagation.

There are many people who have helped me along the duration of this Ph.D. thesis. Without their support this project would have never been possible.

First, I would like to express my gratitude to Preben E. Mogensen, whom I met in Malaga around four years ago, and gave me the opportunity to fulfil my wish of doing a Ph.D. thesis in a world-class research group. He has provided me guidance, support, and countless highlighting discussions. His valuable supervision is directly reflected in the final quality of this thesis. I am very grateful to Jeroen Wigard who has generously devoted to me much of his knowledge and time. He has also encouraged me in many occasions and I deeply acknowledge him for that. I would also like to thank Nokia Networks R&D for the sponsorship of this project.

Many appreciations are paid to the colleagues and former colleagues at the Cellular Systems group of CPK and Nokia Networks Aalborg. Special esteem goes to Laurent Schumacher for his patience with me, and his wise advices. Also distinguished thanks to Troels E. Kolding for his personal interest and innumerable counsels. He and Klaus I. Pedersen have reviewed a significant part of this thesis and I am grateful to them for that. Also remarked recognition to my friend Juan Ramiro, who provided me with a major part of the network level simulator that I have widely used in this investigation. I would also like to thank Frank Frederiksen, Per Henrik Michaelsen, Michael Støttrup and Mika Kolehmainen for their many useful suggestions. In general, all my colleagues at the Cellular Systems group and at Nokia Networks Aalborg, including the secretarial personnel, have helped or assisted me at a certain point and I would like to show them my gratitude for it.

I am grateful to Kari Sipilä for being my host during my four-month stay at Nokia Networks, Helsinki, Finland. I would also like to thank Viktor Varsa and Igor Curcio for their help in the investigation about streaming services.

During these last years many of my friends have particularly supported and encouraged me in the difficult moments. I deeply acknowledge my friends Juan Ramiro, José Outes, Isaías

López, and Juan Pablo Albaladejo, who started with me the adventure of leaving Malaga three and a half years ago, for their attention, care and unlimited support. I am particularly thankful for their friendship and affection to Antonio Carballo, Oscar Rubio, Virginia Sánchez, Christian Rom and Flavia Matíes. I would also like to acknowledge Lucía for the support she gave me for very long time.

I want to thank to my family, especially my aunt Mari, my grandmother Aurea, my cousin María Luisa, and my aunt Olga who have been calling me continuously during these years expressing their interest for me.

Foremost, I am profoundly grateful to my parents who have always expressed me their unconditional support and love.

Table of Contents

Contents

Abbreviations

Abbreviations

3G	Third Generation
AAC	Advanced Audio Coding
AMR	Adaptive Multi Rate (speech codec)
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat Request
AVI	Actual Value Interface
BDP	Bandwidth Delay Product
BLER	Block Erasure Rate
CBR	Constant Bit Rate
CBRP	Constant Bit Rate Packet (Transmission)
CC	Chase Combining
CmCH-PI	Common Transport Channel Priority Indicator
CN	Core Network
CPCH	(Uplink) Common Packet Channel
CQI	Channel Quality Indicator
DCH	Dedicated Channel
DPCH	Dedicated Physical Channel
DSCH	Downlink Shared Channel
DTCH	Dedicated Traffic Channel
EDGE	Enhanced Data Rates for Global Evolution
EGPRS	Enhanced General Packet Radio Service
FACH	Forward Access Channel
FDDI	Fiber Distributed Data Interface
FFQ	Fluid Fair Queueing
FIFO	First In First Out
FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GGSN	Gateway GPRS Support Node
GERAN	GSM/EDGE Radio Access Network
GPRS	General Packet Radio Service
GSM	Global System For Mobile Communications
GTP	GPRS Tunneling Protocol
GTP-U	User Plane Part of GTP
H-ARQ	Hybrid Automatic Repeat Request
HSCSD	High Speed Circuit Switched Data
HSDPA	High Speed Downlink Packet Access
HS-DPCCH	High Speed – Dedicated Physical Control Channel
HS-DSCH	High Speed – Downlink Shared Channel
HS-DSCH FP	High Speed – Downlink Shared Channel Frame Protocol
HS-PDSCH	High Speed – Physical Downlink Shared Channel

HS-SCCH	High Speed – Shared Control Channel
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Procotol
IMAP	Internet Message Access Protocol
IP	Internet Protocol
IR	Incremental Redundancy
IT	Information Technology
MAC	Medium Access Control (layer)
MAC-hs	Medium Access Control (layer) – High Speed
MaxDAT	Maximum Number of Retransmissions of RLC PDU
MBR	Maximum Bit Rate
MCS	Modulation and Coding Scheme
MMS	Multimedia Messaging Service
MPEG	Motion Picture Experts Group
MRC	Maximal Ratio Combining
MSS	Maximum Segment Size
MTU	Maximum Transmission Unit
NRT	Non Real Time
PLMN	Public Land Mobile Network
PDCP	Packet Data Convergence Protocol
PDP	Packet Data Protocol
PDR	Peak Data Rate
PDU	Protocol Data Unit
PS	Packet Scheduling
PSS	Packet Switched Streaming Service
QoS	Quality of Service
RACH	Random Access Channel
RAN	Radio Access Network
RLC	Radio Link Control (layer)
RNC	Radio Network Controller
RRM	Radio Resource Management
RTCP	Real Time Control Protocol
RTO	Retransmission Time Out
RTP	Real Time Protocol
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
RTTVAR	Round Trip Time Variation
SAP	Service Access Point
SAW	Stop And Wait
SDU	Service Data Unit
SF	Spreading Factor
SGSN	Serving GPRS Support Node
SHO	Soft Hand Over
SIR	Signal to Interference Ratio

SISO	Single Input Single Output
SLP	Segment Loss Probability
SMS	Short Message Service
SMTP	Simple Mail Transfer Protocol
SPI	Scheduling Priority Indicator
SRTT	Smoothed Round Trip Time
TCP	Transmission Control Protocol
TFRC	Transport Format and Resource Combination
THP	Traffic Handling Priority
TPC	Transmit Power Control
TTI	Transmission Time Interval
UE	User Entity
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
URL	Uniform Resource Locator
UTRAN	UMTS Terrestrial Radio Access Network
VBR	Variable Bit Rate
VBRP	Variable Bit Rate Packet (Transmission)
WAN	Wide Area Network
WAP	Wireless Application Protocol
WCDMA	Wideband Code Division Multiple Access
WFQ	Weighted Fair Queueing
WTP	Wireless Transaction Protocol

Chapter 1

Introduction

1.1 Third Generation Services

Over the last two decades, Internet has become the largest network for exchange of information worldwide. The volume of Internet Protocol (IP) based traffic has already exceeded that of circuit switched traffic in most fixed networks [1] and it is expected to continue growing at a rapid speed.

Mobile communication forecasts suggest that the average revenue per user for voice traffic is set to decline in cellular networks (e.g. in Western Europe as stated in [2]), which leads operators to introduce new data services to sustain their business growth in the regions where the speech service is already deployed. The provision of IP-based services in mobile communication networks has raised expectations of a major traffic emergence analogous to the one already experienced over the Internet. In [3], the Universal Mobile Telecommunication System (UMTS) Forum predicts a considerable increase of the worldwide demand for wireless data services over the years to come of the present decade. Besides the Internet-like services, the addition of mobility to data communication enables new services not meaningful in a fixed network [4], e.g. location-based services.

The goal of Third Generation (3G) mobile communication systems is to provide users not only with the traditional circuit switched services, but also with new multimedia services with high quality images and video for person-to-person communication, and with access to services and information in private and public networks.

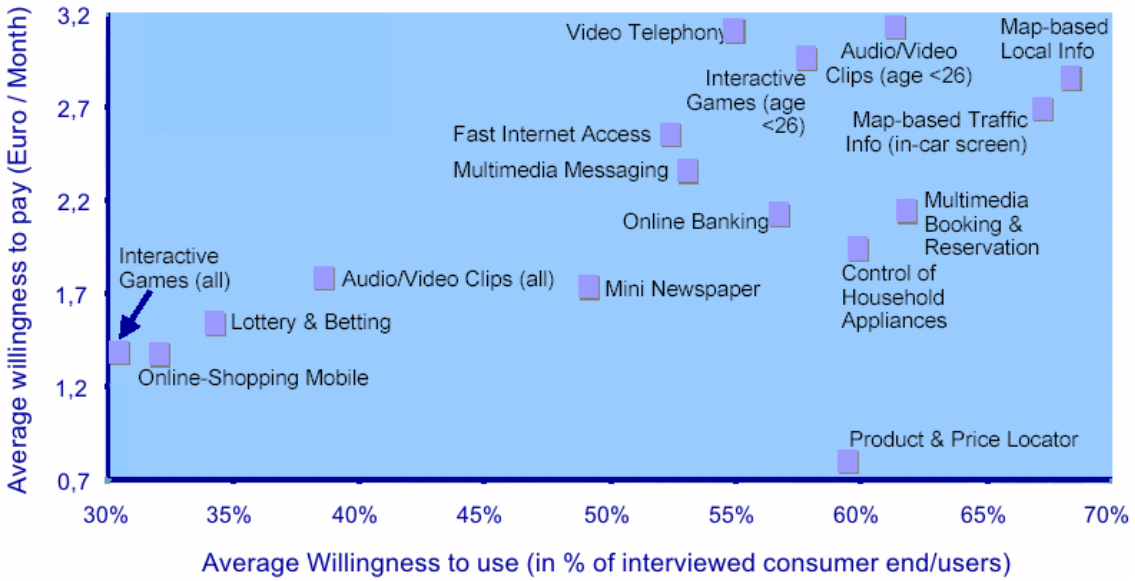


Figure 1.1: Consumer Acceptance of Mobile Data Services for Private Users from [5].

Consumers are expected to acquire mobile data services if their contents add value to the consumer's life by satisfying a concrete necessity or requirement. From the end user's viewpoint, the worth provided by the service contents must contribute to his cost-effectiveness, time-efficiency, or merely his entertainment. Figure 1.1 plots the consumer acceptance of mobile data services for private users in 3G systems from [5]. See in [PI 1.1] the definition of the average user willingness to pay for wireless data services. The average willingness to use and pay for mobile data services varies significantly depending on the service and the consumer characteristics. For example, private users exhibit a high interest for rich content services like video telephony, audio/video clips, or map based information. However, as shown in [5] (not depicted here), business users express much higher interest for fast Internet/Intranet access.

Since the requested information is often located in the Internet, a large majority of the wireless data services (e.g. most of the depicted ones in Figure 1.1) are inherently asymmetrical, causing a heavier, and in some cases much heavier, load in the downlink direction of the communication path.

While the consumer preferences and needs of mobile communication services are the main drivers of the wireless communication market, the technology defines the boundaries of the application and service menu offered by the operators and service providers. A certain technology can only support with a satisfactory user's quality perception (e.g. similar to the one experienced by end user in the fixed network) a limited set of services. Moreover, the provision of those services must not incur a major system cost in terms of spectral efficiency. Figure 1.2 plots the data rate requirements of some popular 3G applications. See in [PI 1.2] the definition of the data rate requirement on 3G applications. Next, it is given a brief overview of the technologies evolved from the Global System for Mobile Communications (GSM) towards the Wideband Code Division Multiple Access (WCDMA).

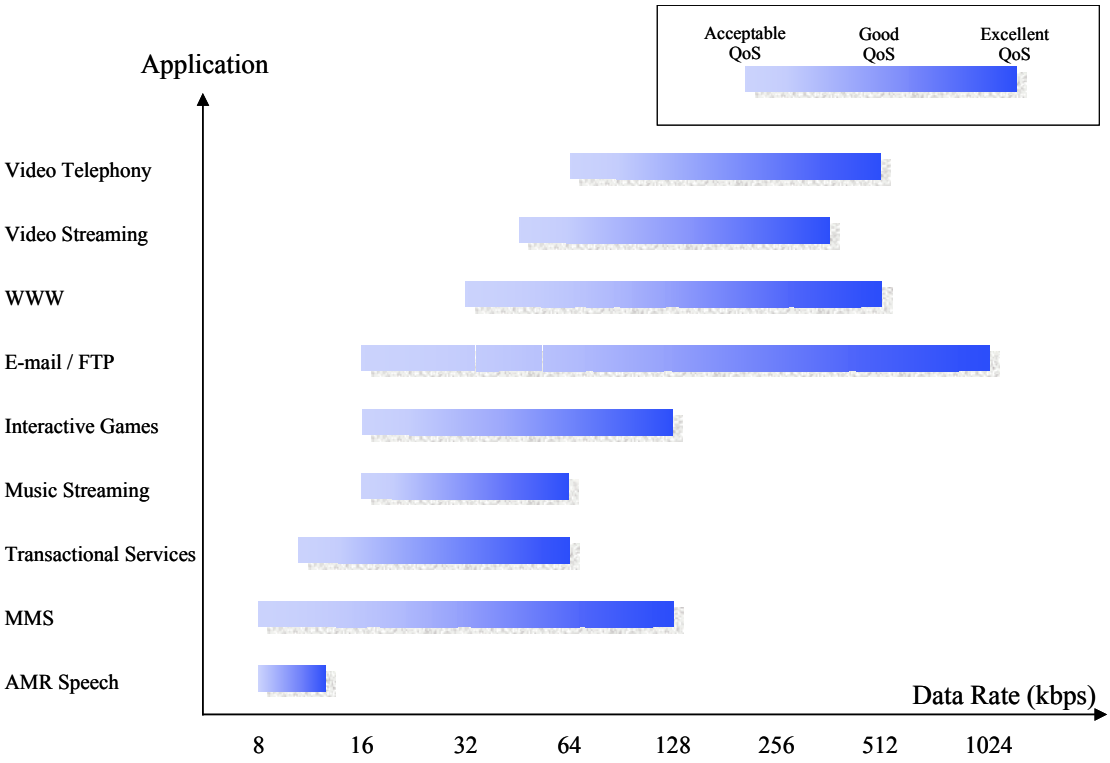


Figure 1.2:Data Rate Requirements of 3G Applications.

1.2 Evolution of Radio Access Technologies

The evolution of the packet switched services in Second Generation mobile communication systems starts with the General Packet Radio Service (GPRS), which introduces the packet switched domain to the Core Network (CN) of GSM. For the users, GPRS offers the possibility to be permanently on line but only paying for the actual retrieval of information. The radio interface of GPRS provides peak data rates of up to 21.4 kbps per time slot, which can be combined with the transmission of multiple time slots per user. See in [PI 1.3] the definition of the peak data rate. With eight time slots a peak data rate of around 171 kbps can be reached with GPRS. However, under more realistic conditions (i.e. loaded network) and assuming a mobile terminal supporting four time slots for downlink and one for uplink, the average user throughput of GPRS could range around 30 – 40 kbps [6]. See in [PI 1.4] the definition of the average user throughput. With such an average user throughput, GPRS can bear data services like multimedia messaging, e-mail download or transaction based commerce applications.

The Enhanced Data Rates for Global Evolution (EDGE) system incorporates the 8PSK modulation together with link adaptation and Incremental Redundancy (IR) across the GPRS air interface to increase the spectral efficiency. The resulting peak data rate per slot is around 60 kbps, and with the support of four time slots, a peak data rate up to 237 kbps can be reached in the downlink. In a loaded network, and with the previous time slot configuration, EDGE can deliver an average user throughput of around 128 kbps [7]. With such capabilities, EDGE enables the provision of e-newspapers, images and sound files, tele-shopping and IP-

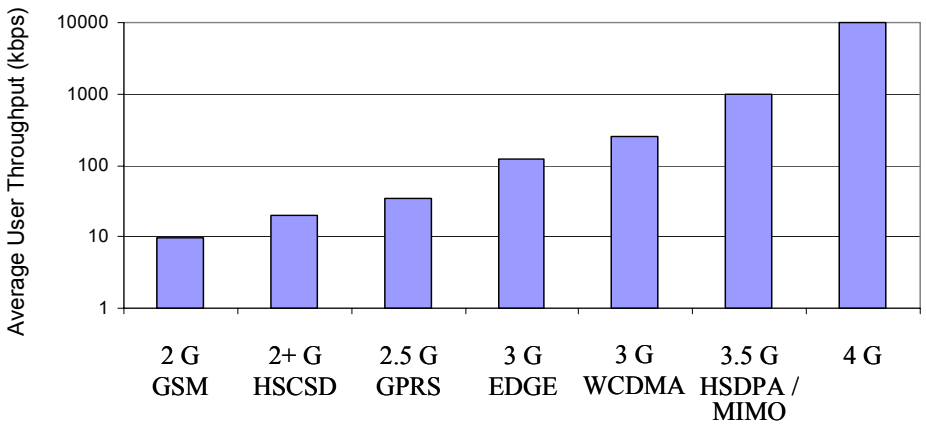


Figure 1.3: Average User Throughput of Mobile Communication Radio Access Technologies.

based video telephony. Moreover, it will improve the user experience for other services such as e-mail or web browsing. Furthermore, to align EDGE with 3G systems, it includes the same traffic classes as the Universal Mobile Telecommunication System (UMTS) and the same interface to the Core Network.

Wideband Code Division Multiple Access (WCDMA) is the most widely adopted air interface for Third Generation systems. It provides peak bit rates of 2 Mbps, variable data rate on demand, a 5 Mhz bandwidth, and a significant reduction of the network round trip time. Though the spectral efficiency of WCDMA may possibly not represent a major benefit over EDGE [7]-[8], it can still provide higher average user data rates under loaded network conditions (in the range 200 – 300 kbps) [7] due to the larger bandwidth. See in [PI 1.6] the definition of the spectral efficiency. The services to be provided by WCDMA are similar to that of EDGE, with potential enhancement of the user's perceived quality for some high data rate services (e.g. video streaming on demand). Figure 1.3 summarizes the average user throughput of the different radio access technologies.

The continuous evolution of the mobile communications technology will allow to increase the user's data rate, thereby enabling new data services and improving the QoS of already existing ones.

1.3 HSDPA

As previously described, the evolution of the mobile communication market is expected to bring a major increase of the data traffic demands combined with high bit rate services. To successfully satisfy these requirements, Third Generation systems must increase their spectral efficiency and support high user data rates especially on the downlink direction of the communication path due to its heavier load. There exists a considerable toolbox of technologies that can improve the downlink performance of the WCDMA radio interface (see [9]). One of these technologies is the High Speed Downlink Packet Access (HSDPA) concept, which has been included by the 3GPP in the specifications of the Release 5 as an evolutionary step to boost the WCDMA performance for downlink packet traffic.

HSDPA appears as an umbrella of technological enhancements that permits to increase user peak data rates up to 10 Mbps, reduce the service response time, and improve the spectral efficiency for downlink packet data services. See in [PI 1.7] the definition of the service response time.

The HSDPA concept consists of a new downlink time shared channel that supports a 2-ms transmission time interval (TTI), adaptive modulation and coding (AMC), multi-code transmission, and fast physical layer hybrid ARQ (H-ARQ). The link adaptation and packet scheduling functionalities are executed directly from the Node B, which enables them to acquire knowledge of the instantaneous radio channel quality of each user. This knowledge allows advanced packet scheduling techniques that can profit from a form of selection (multi-user) diversity.

The main focus of the present Ph.D. thesis is the performance of HSDPA because this concept is expected to become the natural evolution path of the WCDMA radio interface.

1.4 Spectral Efficiency Versus QoS Trade Off

The term Quality of Service (QoS) experienced by the end user of a mobile communication system refers to the collective effect of service performances that determine the degree of satisfaction of this end user of the service. The demands imposed by a certain service on the system to provide a satisfactory user QoS completely depend on the intrinsic characteristics of the service itself. Chapter 2 will provide a detailed description of the QoS requirements of the service classes considered in this Ph.D. thesis. Typically, the service response time is one of key factors with most influence on the user's perceived QoS for Non Real Time services. The service response time of such services is directly determined by the user's data rate. The most crucial constraints of Real Time traffic are the packet transfer delay (average and/or variance) and user's data rate. Fundamentally, the more demanding these two constraints are, the lower spectral efficiency can be achieved by the mobile communication system.

The technological enhancements incorporated by HSDPA may upgrade the system performance of WCDMA as depicted in Figure 1.4. The performance of HSDPA may not only provide a spectral efficiency gain, but also a support for higher user data rates. As shown in Figure 1.4, the potential benefits of the new system can be exploited in several manners. The system may simply be operated at a higher load obtaining the full spectral efficiency profit, thereby allowing to achieve a lower cost per delivered data bit, but without any improvement of the user's QoS. Alternatively, the average throughput of end user could be increased at the cost of maintaining constant the spectral efficiency of the system. Instead, a compromise between previous solutions could be selected.

The optimal preference to the previous alternatives strongly depends on diverse factors such as the requirements (in terms of data rates, etc) to support data services as well as the overall traffic volume. In residential areas, where the average data rates provided by WCDMA could be sufficient to cope with the needs of typical services requested by private users (e.g. e-newspaper or video on demand), HSDPA could accommodate the predicted increase of demand for wireless data services. In business areas, the high peak data rates of HSDPA could enhance common services solicited by corporative users (e.g. fast Internet/Intranet access).

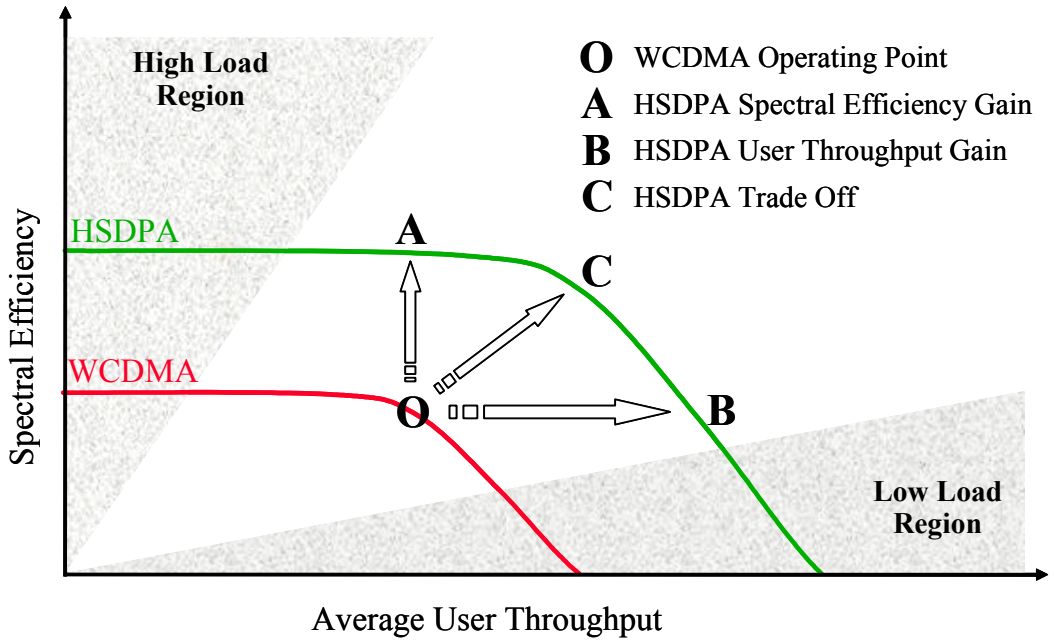


Figure 1.4: Spectral Efficiency Vs Average User Throughput in WCDMA and HSDPA.

1.5 Objectives

The introduction of a new technology such as HSDPA in the Release 5 of the 3GPP specifications arises the question about its performance capabilities: is HSDPA a radio access technology that can facilitate the evolution of the wireless communication market by enabling the support of high data rate services and by increasing the system capacity. See in [PI 1.10] the definition of the network capacity.

The system level performance of HSDPA and the network capacity benefits that this technology can provide to cellular systems have already been addressed in the literature [10], [11]. However, those studies have only considered best effort traffic without taken into consideration the concept of Quality of Service (QoS).

In order to be able to facilitate the evolution of the wireless communication market, the HSDPA technology should be able to increase the system capacity over other compiting technologies (e.g. EDGE or WCDMA) not only for best effort traffic. The HSDPA technology should also be able to meet the QoS demands of third generation mobile data services. Firstly, data services impose minimum QoS requirements on the communication (e.g. delay jitter, guaranteed bit rate, etc) that must be satisfied in order to ensure an acceptable degree of quality to the end user. Secondly, the provision of the service with a degree of quality beyond the minimum acceptable level (Figure 1.2) imposes on the communication more stringent demands (e.g. a bit rates higher than the minimum guaranteed), which typically can be satisfied at the expenses of a reduced network throughput. See in [PI 1.5] the definition of the network throughput.

The concept of QoS is a key factor in the HSDPA technology evaluation because there exists a fundamental trade-off between the QoS provided to the end user and the achievable network

throughput. The main objective of this Ph.D. thesis is to investigate the system level performance of HSDPA, as well as this existing trade-off between the QoS provided to the end user and the achievable network throughput by this technology for different traffic types. This evaluation will enable to set the boundaries of the performance capabilities of the HSDPA technology. The traffic classes targeted in the investigation described above are the UMTS traffic classes to be supported by HSDPA:

- Streaming.
- Interactive.
- Background.

An additional objective of this Ph.D. thesis is to carry out a cell capacity comparison between HSDPA and WCDMA technologies for best effort traffic. This comparison will enable to benchmark HSDPA and conclude if this technology can indeed contribute to accommodate an increase of traffic volume in cellular systems.

Prior to the HSDPA evaluation, it is of major importance to identify the traffic characteristics of the traffic classes under investigation, as well as the QoS requirements that these services impose on the communication, and especially on the radio access network (i.e. on HSDPA).

The investigation will concentrate on the Packet Scheduler because it is the central entity of the HSDPA design. This functionality governs the distribution of the radio resources available in the cell among the users. Due to its function, the Packet Scheduler has a direct impact on the HSDPA system performance. Similarly, it also determines the end user performance, and more specifically the relative performance between the users in the cell. Hence, it will be of paramount interest to find suitable Packet Scheduling algorithms that can optimize the aforementioned trade-off between system capacity and end user performance for the traffic classes specified above. Due to the influence of this functionality on the relative performance between users, special emphasis will be set on the concept of fairness in the distribution of the radio resources among the users in the cell.

Particular attention will be paid to the performance of TCP traffic over HSDPA. A very large share of the overall Non Real Time traffic to be conveyed by future cellular systems is expected to be based on a TCP/IP platform. It is well known that the flow control algorithms of TCP can not only influence the end user throughput, but also the overall system performance [12] due to the traffic inactivity caused by these algorithms. For this reason, the effects of TCP on the performance of HSDPA will be subject of investigation in this Ph.D. thesis. Prior to this analysis, the performance of TCP over Dedicated Channels (DCH) in WCDMA will be evaluated in order to have a more thorough understanding of the flow control mechanisms of this protocol.

1.6 Assessment Methodology

The assessment of the HSDPA technology at network level requires a method that can provide absolute cell capacity figures under realistic conditions. There exist several options to evaluate the performance of a cellular network. Figure 1.5 depicts the different possibilities.

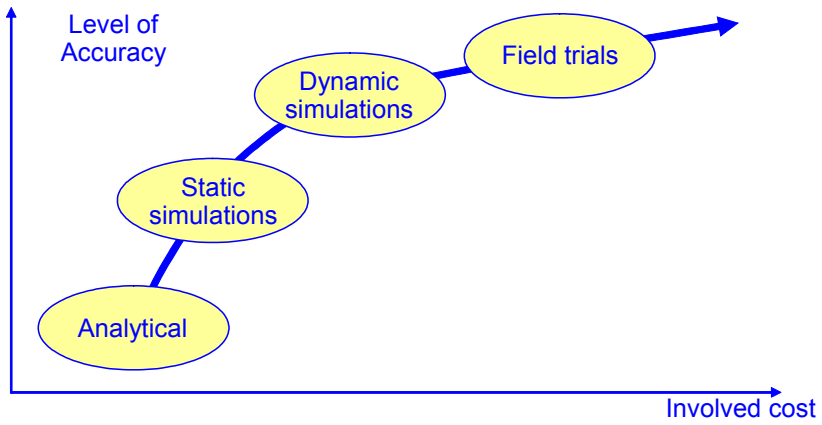


Figure 1.5: Approaches to Assess the Performance of Cellular Networks.

One of those options is to use analytical investigations based on simple mathematical models. This type of studies has the advantages of having a lower cost and requiring less effort than other methods. However, more detailed studies are problematic due to the high complexity of the UMTS system.

Another option is to use computer aided simulations of the cellular network under study. The processing power of today's computers allows including complex models and algorithms in the system evaluation. Computer-based simulations can be divided into three categories: static, dynamic and quasi-dynamic. Static simulations are characterized by not including the time dimension, and the results are obtained by extracting sufficient statistics of statistically independent snapshots of the system (Monte Carlo approach). Static simulations can be used for coverage/capacity analysis or network planning. Dynamic simulations include the time dimension, which adds further complexity to the considered model. However, dynamic simulations are very appropriate for investigating time dependent mechanisms or dynamic algorithms. Radio resource management functionalities, such as the Packet Scheduler, can be properly analyzed with this assessment methodology. A middle step between static and dynamic simulations is the so called quasi-dynamic simulations, which only include a single time dependent process while the rest of time dependent processes are modelled as static. This solution represents a trade-off between the accuracy of fully dynamic simulations and the simplicity of static simulations.

The last alternative is to conduct field tests in a trial or operational network. For a specific network configuration, this method can provide the most accurate assessment of the system performance. However, this option requires the availability of the network itself. Moreover, the results tend to be network specific (i.e. depending on the network configuration and environment) and may therefore differ from network to network.

Considering the need of the Ph.D. objectives (section 1.5) to acquire absolute cell capacity figures under realistic conditions, and the early stage of the HSDPA system design (which implies that operational networks are not available), quasi-dynamic system level simulations will be selected as the assessment methodology in the present Ph.D. thesis.

1.7 Outline of the Dissertation

This thesis report is organized as follows:

Chapter 1: gives a short introduction and outlines the objectives of this Ph.D. thesis.

Chapter 2: presents a description of the most relevant QoS attributes of the UMTS bearers under study, which allows to identify the QoS demands imposed on the conveying networks. The chapter further aims at addressing the traffic characteristics of Streaming and Non Real Time traffic. For this purpose, it is described of an end-to-end framework for video streaming services, as well as the protocol architecture and the statistical traffic characterization of a typical Non Real Time service such as web browsing.

Chapter 3: studies the performance of the TCP traffic over Dedicated Channels in WCDMA. The chapter investigates the effect on the user throughput of the TCP slow start, the receiver's buffer capabilities, or random segments losses. Moreover, the interaction between the link level error corrections and the TCP end to end recovery mechanisms is analyzed.

Chapter 4: provides a general overview of the HSDPA technology that is required to achieve a full comprehension of the HSDPA investigations carried out in this Ph.D. thesis. This chapter provides deeper degree of detail into the aspects that are necessary for the following chapters.

Chapter 5: concentrates on the Packet Scheduling entity of HSDPA. The chapter investigates the performance of various scheduling strategies with different degree of fairness among users, and with and without exploiting the multiuser diversity. The performance of these scheduling algorithms will be assessed under different requirements of minimum throughput guarantees in order to evaluate the trade-off between the network throughput and the QoS provided to the end user.

Chapter 6: analyzes the performance of TCP traffic on HSDPA. The chapter investigates the impact of the TCP flow control algorithms not only the user throughput, but also the HSDPA system throughput. Moreover, the chapter describes the interaction between cell throughput, Node B queuing delay, and end-user throughput for TCP traffic.

Chapter 7: investigates the performance of streaming flows on HSDPA. The chapter discusses different scheduling policies for streaming traffic and evaluates the network level performance of the most interesting ones. The influence of the delay jitter constrains imposed by the streaming service on the system capacity is extensively analysed. The chapter ends comparing the cell capacity results for streaming services on HSDPA with the figures obtained for Non Real Time traffic in Chapter.

Chapter 8: draws the main conclusions of this Ph.D. investigation and discusses future research topics.

1.8 Publications and Patent Applications

The following articles have been published during the Ph.D. study:

- P. Ameigeiras, J. Wigard, and P. Mogensen. "Impact of TCP Flow Control in the Radio Resource Management of WCDMA Networks". IEEE Proceedings of 55th Vehicular Technology Conference, Vol. 2, pp. 977-981, May 2002.

- P. Ameigeiras, I. López, J. Wigard, N. Madsen, and P. Mogensen. "Generic Traffic Models for Packet Data Services in 3G Wireless Networks". Proceedings of the WPMC-01 Conference, Vol. 3, pp. 1521-1526, September 2001.

- P. Ameigeiras, J. Wigard, P. Andersen, H. Damgaar, and P. Mogensen. "Performance of Link Adaptation in GPRS Networks". IEEE Proceedings of 52nd Vehicular Technology Conference, Vol. 2, pp. 492-499, September 2000.

In addition, the following articles have been co-authored:

- J. Wigard, N. Madsen, P. Ameigeiras, I. López, and P. Mogensen. "Packet Scheduling with QoS Differentiation". Kluwer Journal on Wireless Personal Communications, Vol. 23, Issue 1, pp. 147-160, October 2002.

- I. López, P. Ameigeiras, J. Wigard, and P. Mogensen. "Downlink Radio Resource Management for IP Packet Services in UMTS". IEEE Proceedings of 53rd Vehicular Technology Conference, Vol. 4, pp. 2387-2391, May 2001.

- L. Berger, T. Kolding, J. Ramiro-Moreno, L. Schumacher, and P. Mogensen. "Interaction of Transmit Diversity and Proportional Fair Scheduling". IEEE Proceedings of 57th Vehicular Technology Conference, Vol. 4, pp. 2423-2427, April 2003.

- J. Wigard, N. Madsen, P. Ameigeiras, I. López, and P. Mogensen. "Packet Scheduling with QoS Differentiation". Proceedings of the WPMC-01 Conference, Vol. 2, pp. 693-698, September 2001.

During this Ph.D. investigation four patent applications have been filed in the areas of transport protocols and radio access networks.

Chapter 2

Packet Switched Services in Third Generation Systems

2.1 Introduction

In the mobile communication market, where operators and manufacturers need to offer new services to sustain the business growth, customers are expected to demand services that deliver them productivity efficiency for business, entertainment, information and transacting opportunities [3]. As commented in previous chapter, Third Generation systems are designed to provide not only the traditional circuit switched services, but new multimedia services with high quality images and video for person-to-person communication, and with access to services in private and public networks. UMTS, as a system to enable Third Generation services, must support flexible communication capabilities to convey such as multimedia services and any other services that result of the continuous evolution process of the IT society.

To provide such flexibility, UMTS allows a user/application to negotiate the bearer characteristics that are the most appropriate for carrying the information [10]. The 3GPP standardized different classes of bearers attending to the application QoS requirements. The number of classes is limited to four to ensure that the system is capable of providing

reasonable QoS resolution [13]. The general characteristics of the services significantly differ from class to class: the QoS requirements imposed on the network (and thereby the applicable UMTS bearer parameters and the parameter values), the protocols employed in the communications, the statistical properties of the incoming traffic (e.g. the amount of data to be downloaded), etc. The system design and its performance analysis compel deep understanding of these general service characteristics. The present chapter aims at highlighting these service properties. Within the four service classes, the chapter concentrates on those ones that are the scope of HSDPA.

Due to the relevance of the topic, the literature dealing with Third Generation services and their traffic characteristics is relatively extensive. In [14], Reguera addresses aspects of classification, characterisation and deployment scenarios for services and applications in the evolution of UMTS. In [15], Montes gives a thorough description of an end-to-end QoS framework for streaming services in 3G networks and the involved signalling procedures. The literature related to Web traffic is even wider due to its investigation by the Internet research community [16]-[22]. Perhaps [20] is one of the most relevant publications dealing with the characterization of Web traffic in Wide Area Networks (WAN). In [20], Crovella gives evidence of the self-similarity of this type of traffic, being the heavy tailed distribution of the file sizes likely the primary determiner of that behaviour. In [22], Casilari provides a general review of HTTP traffic characteristics, and the author already points out that the behaviour of Web users strongly affects the nature of TCP connections. Particularly, Casilari advances that the reading time between consecutive pages determines the reusability of the TCP connections. In [19], Molina presents a compilation of the main measurement statistics of HTTP connections, and concludes that the lognormal is a good approximation for the distribution of the page sizes and reading times. In [23], Stuckmann also considers models for Simple Mail Transfer Protocol (SMTP) email. In [24], it is proposed a model for File Transfer Protocol (FTP) traffic.

The chapter is organized as follows: section 2.2 describes the UMTS QoS classes. Section 2.3 describes the protocols and transmission of streaming sessions over UMTS. Section 2.4 deals with the creation of the multimedia content and its traffic characteristics, while section 2.5 analyses the interaction between those traffic properties and the attributes of the UMTS QoS streaming class. Section 2.7 concentrates on one of the interactive services, HTTP, since it is expected to be one of the services with larger share of the overall traffic volume in 3G networks. Finally section 2.8 gives the concluding remarks of the present chapter.

2.2 The UMTS QoS Classes

According to [14], “an application is defined as a task that requires communication of one or more information streams, between two or more parties that are geographically separated”, while a set of applications with similar characteristics is classified as a service.

There are many ways to categorize services depending on the classification criterion, such as the directionality (unidirectional or bi-directional), symmetry of communications (symmetric or asymmetric), etc [25]. The 3GPP created a traffic classification attending to their general QoS requirements. The QoS refers to the collective effect of service performances that determine the degree of satisfaction of the end-user of the service. The four classes are [13]:

- *Conversational class.*
- *Streaming class.*
- *Interactive class.*
- *Background class.*

The main distinguishing factor between the classes is how delay sensitive the traffic is. Conversational class is meant for traffic very delay sensitive whereas background class is the most delay insensitive.

The traffic corresponding to the conversational class refers to real time conversation where the time relation between information entities of the stream must be preserved. The conversational pattern of this type of communication requires a low end-to-end delay to satisfy the stringent requirements of human perception. A service example is telephony speech, voice over IP or video conferencing. According to [26], HSDPA focuses on streaming, interactive, and background traffic classes but not on conversational traffic. For this reason, only the three least delay sensitive traffic classes will be within the scope of this Ph.D. thesis. The main characteristics of these classes are described in following subsections.

2.2.1 Description of Streaming UMTS QoS Class

According to [27], when the user is looking at (listening to) real time video (audio) the scheme of real time stream applies. The real time data flow always aims at a live (human) destination. This scheme is a one-way transport. The fundamental characteristic of this QoS class is that the communication has to preserve the time relation (or variation) between information entities (i.e. samples or packets) of the stream, although it does not have any requirements on low end-to-end transfer delay. In order to allow end-to-end delay variations larger than accepted by the human perception, today's streaming applications apply time alignment prior to decoding at the receiving end. The time alignment function initially delays by buffering the received stream before starting the decoding process, which enables to cope with delay variations up to the limits provided by the buffering capabilities. Then, the client can start playing out the data before the entire file has been transmitted. Figure 2.1 depicts the variable network delay removal through initial buffering delay for a constant bit rate video application.

The QoS attributes of a UMTS streaming bearer with most relevance for this thesis are [27]:

- *Transfer Delay*: Regarding the statistical characteristics of this group of services, for audio streaming, the generated traffic is rather non-bursty, whereas the video traffic may be of a more bursty nature. In either case, it is meaningful to guarantee a transfer delay of an arbitrary Service Data Unit (SDU). According to [27], "the transfer delay indicates the maximum delay for 95th percentile of the distribution of the delay for all delivered SDUs during the lifetime of the bearer service, where delay for an SDU is defined as the time from a request to transfer an SDU at one Service Access Point (SAP) to its delivery at the other SAP". It is worthy to note in [27] that the application's requirement on delay variation through the UMTS network is solely expressed through the transfer delay. This implies that both the maximum UMTS end-

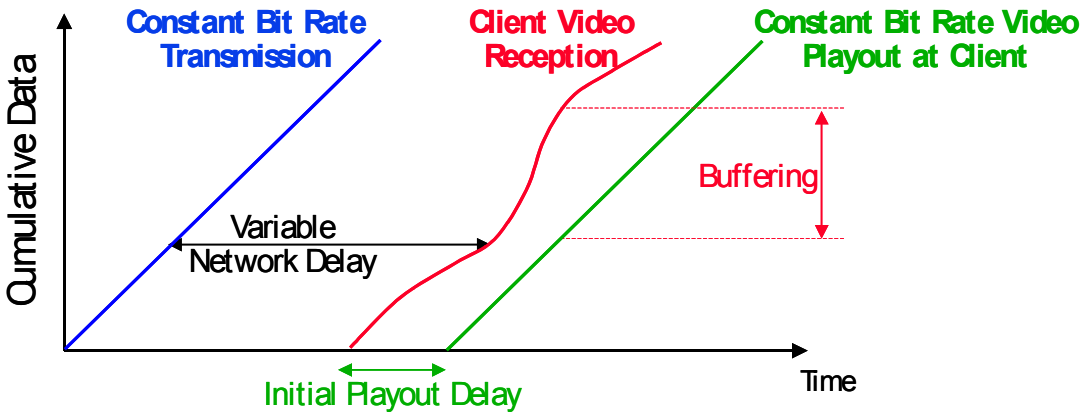


Figure 2.1: Variable Network Delay Removal Through Initial Play out Delay for a Constant Bit Rate Video Application.

to-end delay and the maximum UMTS network jitter are upper bounded by the transfer delay attribute.

- *Guaranteed Bit Rate (GBR)*: It defines the bit rate the bearer shall guarantee to the user or the application. Other bearer service attributes (e.g. transfer delay, etc.) are guaranteed for traffic up to the guaranteed bit rate.

- *Maximum Bit Rate (MBR)*: According to [27], "it describes the upper limit of the bit rate that a user or application can provide or accept". This attribute is defined for applications that are able to operate with different bit rates, for example video applications with real-time encoders. It is relevant to note that network is not required to carry traffic exceeding the GBR, but it may transfer traffic up to the MBR depending on the network conditions (such as load). Section 2.5 will treat the QoS mapping from application attributes to the GBR & MBR attributes and their implications for different encoding possibilities for video streaming services.

- *SDU Error Ratio*: As mentioned in [27], "it indicates the fraction of SDUs lost or detected as erroneous". The SDU error ratio requirements may have implications on the operating mode of the Radio Link Control (RLC) layer. If the Physical layer can not provide a sufficiently reliable connection to meet the SDU error ratio requirement, the RLC layer will have to operate in acknowledged mode to satisfy this bearer attribute. See in section 7.2 the discussion about the RLC operating mode for streaming services on HSDPA.

- *Residual Bit Error Ratio*: "is the undetected bit error ratio in the delivered SDUs" [27].

- *Delivery of Erroneous SDUs:* it determines if SDUs with detected errors must be delivered.

By using the SDU Error Ratio, the Residual Bit Error Ratio, and the Delivery of Erroneous SDUs attributes, the application demands on error rate can be met. Moreover, these three attributes must be considered in combination with the Transfer Delay to fulfil the overall application requirements of delivery a certain percentage of all transmitted packets (e.g. 99%) within a certain time.

Bearer Attribute	Streaming Class	Interactive Class	Background Class
Maximum Bit Rate (kbps)	<= 2048	<= 2048 – overhead	<= 2048 – overhead
Delivery of Erroneous SDUs	Yes / No / -	Yes / No / -	Yes / No / -
Residual Bit Error Ratio	$5 \cdot 10^{-2}$, 10^{-2}, $5 \cdot 10^{-3}$, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}	$4 \cdot 10^{-3}$, 10^{-5}, $6 \cdot 10^{-8}$	$4 \cdot 10^{-3}$, 10^{-5}, $6 \cdot 10^{-8}$
SDU Error Ratio	10^{-1}, 10^{-2}, $7 \cdot 10^{-3}$, 10^{-3}, 10^{-4}, 10^{-5}	10^{-3}, 10^{-4}, 10^{-6}	10^{-3}, 10^{-4}, 10^{-6}
Transfer Delay (ms)	280 – maximum value		
Guaranteed Bit Rate (kbps)	<= 2048		
Traffic Handling Priority		1, 2, 3	
Allocation / Retention Priority	1, 2, 3	1, 2, 3	1, 2, 3

Table 2.1: Value Ranges for UTMS Bearer Service Attributes from [13].

Table 2.1 gives the attribute ranges of the streaming bearers from [13]. Examples of streaming services are audio or video streaming on demand, or telemetry (monitoring).

2.2.2 Description of Interactive and Background UMTS QoS Classes

According to [27], the interactive class is a type of Non Real Time communication where an on-line end user requests data from a remote equipment. The communication is characterised by the request response pattern of the end user. Hence, one of the key properties of this scheme is the service response time. The service response time can be defined as the period elapsed since the instant of the data request until the end of the message reception, which determines the degree of satisfaction perceived by the end user. According to [13], interactive users expect "the message within a certain time". Another characteristic of the interactive services is that the message must be transparently transferred. Moreover, interactive traffic can be bursty. Typical services of interactive class are web browsing, wap, e-mail service (server access), data base retrieval, e-commerce, network games, etc.

The background class is another type of Non Real Time communication, which, according to [27], "is optimised for machine-to-machine communication that is not delay sensitive". It is relevant to note that this class is targeted for applications that do not require an immediate action, and message delays in the order of seconds, tens of second, or even minutes may be acceptable because the "destination is not expecting the data within a certain time" [13]. This scheme is the most delay insensitive of all the UMTS QoS classes. As for interactive services, the message must be transparently transferred. Some services examples are background delivery of e-mails (server to server), Short Message Service (SMS), Multimedia Messaging Service (MMS), FTP file transfers, fax, etc.

The QoS attributes of UMTS interactive and background bearers with most relevance for this study are [27]:

- *SDU Error Ratio*: defined as in section 2.2.1. As the message of these types of bearers must be transparently transferred, their actual SDU error ratio is typically very exigent

(in the range 10^{-3} - 10^{-6} as shown in Table 2.1). With the purpose of achieving such an exigent SDU error ratio, the RLC layer is usually operated in acknowledged mode. With this operation mode, the RLC layer executes a selective repeat ARQ protocol that retransmits every unsuccessfully decoded PDU. The SDU error ratio requirements of the UMTS and the Block Erasure Rate (BLER) of the Physical layer are the key parameters to determine the maximum number of retransmissions of the RLC layer. Section 3.3.5 will analyse the influence of the number of retransmissions on the throughput performance of the RLC layer.

- *Residual Bit Error Ratio*: defined as in section 2.2.1.

- *Traffic Handling Priority (THP)*: It is defined only for interactive services and it denotes the relative priority when handling the data belonging to these bearers.

- *Allocation and Retention Priority (ARP)*: This attribute is defined for all the UMTS QoS classes, and its purpose is to prioritise between different bearers when doing admission control. Note that this is a subscription parameter.

Table 2.1 gives the attribute ranges of the interactive and background bearers.

One of the main conclusions from the attributes description above is that the specification does not set any absolute quality guarantees on delay, or bit rate for interactive bearers. Instead, a bearer priority attribute is defined (THP) to allow relative traffic priority differentiation, though the final bearer performance depends on the actual load of the network and its admission control policy. Since the ARP attribute already permits subscription distinction, the THP parameter may be used to differentiate services with different QoS requirements within the interactive class. Furthermore, regarding background bearers, according to [27], the network is to transfer data only “when there is definite spare capacity in the network”.

2.2.3 Real And Non Real Time Traffic

Traditionally, when classifying services attending to their delivery requirements, they have been primarily divided between Real and Non Real Time services. Usually, Real Time (RT) services have been considered to impose strict delay requirements on the end-to-end communication, so that any of the network nodes involved in the Real Time connection is to transfer the packets of the RT flow within a maximum tolerable delay. Due to these severe delay constraints, the error correction possibilities of this type of communication are very limited, and, while link level retransmissions have been difficultly feasible for Real Time services, end-to-end recovery mechanisms have been completely dismissed. This implies that the Real Time flow delivered at the terminating entity is subject to transmission errors, and the final application must be able to cope with them.

On the other hand, Non Real Time traffic has been commonly considered as error sensitive, though with less demanding delay constraints than RT traffic. These characteristics of NRT traffic allow for link and end-to-end level recovery mechanisms that enable the error free delivery of the payload. Another typical difference between RT and NRT services is that the receiving end of RT applications usually consumes the information before the end of the data transaction.

The 3GPP QoS specification [27] also follows the RT/NRT classification by identifying the conversational and streaming traffic with Real Time services, whereas the interactive and background services are claimed to respond to a Non Real Time traffic pattern. However, as soon as certain delay delivery requirements are imposed on background and mainly interactive users, their frontier with the Real Time services becomes diffuse [28]. Network gaming is usually considered as a good example of this situation [10]. While this service responds to a typical interactive pattern, it may have stringent delay requirements that could only be satisfied by a UMTS bearer of the conversational class.

For clarity and structuring purposes, the remaining of this thesis will still assume the division between NRT and RT traffic. Non Real Time services will be considered those whose payload must be transferred error free, and whose transmission delay requirements still allow for end-to-end error recovery mechanisms such as carried out by the Transmission Control Protocol (TCP). On the other hand, Real Time services will be assumed to be those whose transmission delay requirements do not permit end-to-end retransmissions mechanisms, and are to be conveyed by unreliable transport protocols such as the User Datagram Protocol (UDP).

2.3 PSS Streaming Session in 3G Networks. Protocols and Codecs.

2.3.1 PSS Streaming Session in 3G Networks

A Packet Switched Streaming Session (PSS) is enabled by a set of protocols in charge of the session establishment, set-up, session control, data transport and session release. Figure 2.2 depicts a schematic view of a simple streaming session [29].

For the PSS session establishment, the application in the UE initially requests a primary Packet Data Protocol (PDP) context using the interactive traffic class [30], which will convey the session control signalling (Real Time Streaming Protocol, RTSP). Through RTSP signalling, the client obtains the session description, which can be e.g. a presentation description, the scene description (spatial layout and temporal behaviour of a presentation), or just an URL to the contents. RTSP behaves as a session control protocol that sets up and controls the individual media streams [31].

Afterwards, a secondary PDP context is set-up with a UMTS QoS profile for streaming class, which will convey the streaming data flow (Real Time Protocol, RTP) and feedback signalling with the quality of the distributed data (Real Time Control Protocol, RTCP) [30]. A *play* command, transmitted via the RTSP protocol, triggers the data contents to be played out by the client's application. After the data flow retrieval, a *teardown* request stops the stream delivery from the media server, freeing the resources associated with it.

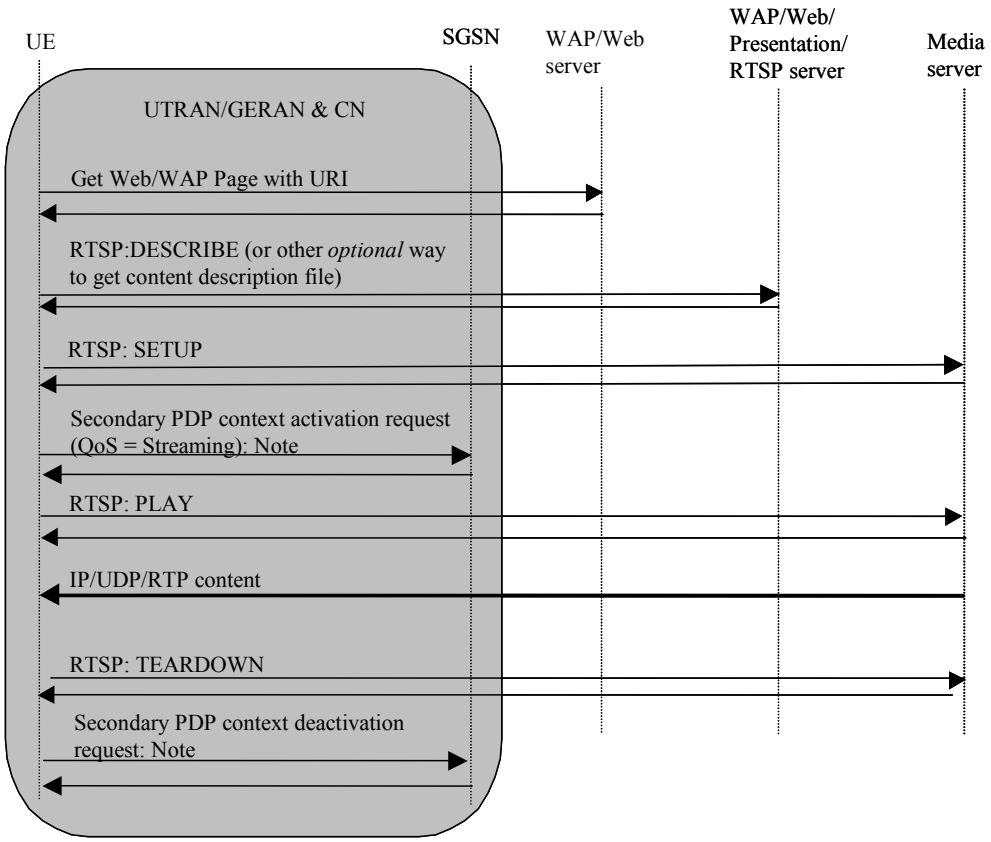


Figure 2.2: Schematic View of a Simple Streaming Session from [29].

2.3.2 Multimedia Content Transmission in UMTS Networks

Figure 2.3 represents the transport of a multimedia component (e.g. the video component) of a PSS over UMTS. At first, the codec generates an encoded sequence from the source stream (e.g. a video clip). Then, the server may apply bit rate control techniques to the released streaming packets in order to limit the bit rate fluctuations of the transmitted flow (see section 2.4.2). These bit rate control techniques make the data stream more suitable for transmission over a bottleneck link (e.g. dial-up connections to the Internet, radio links, etc) [32].

The Internet and the Core Network will route the IP datagrams towards the appropriate Serving GPRS Support Node (SGSN), although a variable network delay is introduced in the conveyance process. Then, the SGSN will forward the data flow to the appropriate Radio Network Controller (RNC). Once in the UMTS Terrestrial Radio Access Network (UTRAN), the flow control function of the Iub interface will rule the data transfer from the RNC to the Node B [33]. At this point, a significant difference between WCDMA and HSDPA architecture arises. In the former case the scheduler is located in the RNC, while in the latter case the scheduler is situated in the Node B. This implies that the main data buffers are located differently in both architectures, which has strong implications on the transport process within UTRAN.

Under a DCH Release 99 configuration, the data transmission is governed by the RNC, and both the Iub and the Node B merely transfer the information. In HSDPA the flow control

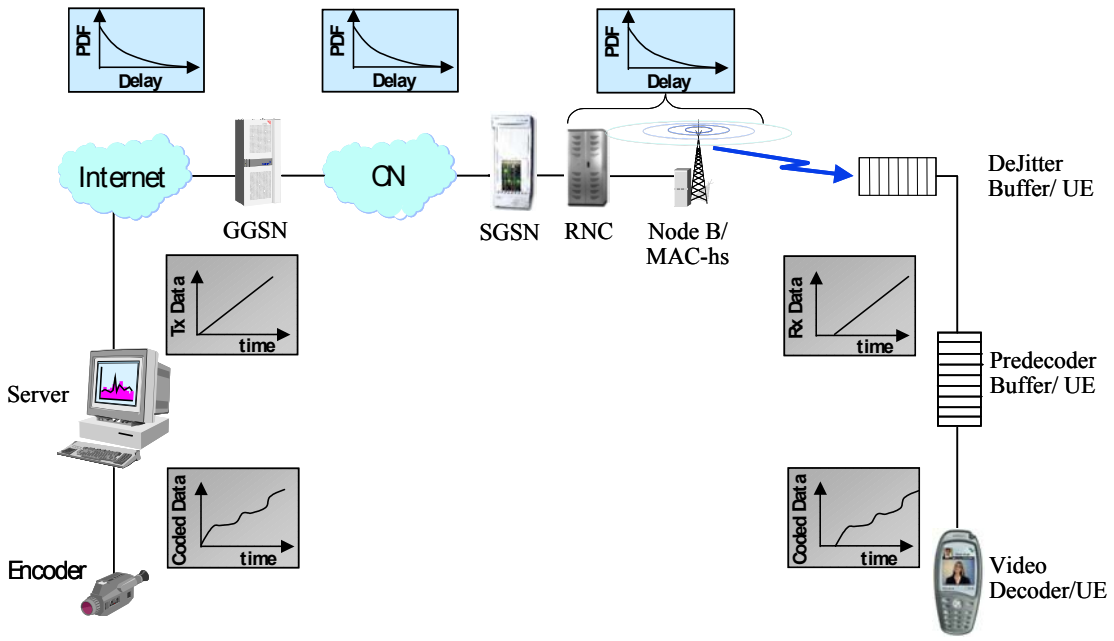


Figure 2.3: End-To-End Video Stream Content Transport.

functionality of the Iub is in charge of carrying the data from the RNC to the Node B buffers. This flow control functionality must avoid the starvation of the Node B buffers associated to the streaming bearers, which might not be an easy task due to the variety of bit rates of the HSDPA users. Finally, in HSDPA, if the cell capacity allows, the Node B will schedule the streaming data flow before its packets reach their due time (see section 7.3), or get rid of them otherwise.

At the UE, the *Dejitter buffer* will compensate for the end-to-end delay variations introduced by the fixed and wireless networks, while the *Predecoder buffer* will restore the time relation between the streaming packets previously distorted in the bit rate control process of the streaming server. Although those two buffers are represented as two different logical entities, they can be implemented in the same physical entity, which can potentially improve their performance. Finally[1], after the RTP packets time alignment is recovered, the decoder can generate the final decoded sequence (e.g. video clip) that is to be presented by the application of the mobile terminal (e.g. in the graphical display).

It is relevant to note that the various networks crossed by the streaming flow introduce their own delay variations without any further information of the jitter generated by the remaining networks, while its removal can only be carried out at the end terminal. That implies that the overall jitter introduced along the entire connection equals the sum of the jitters introduced by all the conveying networks, which should not overcome the Dejitter buffering capabilities if the RTP packet time alignment is to be fully restored. Hence, the total delay jitter must be budgeted among the conveying networks, which limits the maximum delay jitter that any of them can introduce. In the case of UMTS, as explained in section 2.2.1, that delay jitter upper bound is expressed through the transfer delay QoS attribute.

[1] The Post decoder buffering applied after the decoding process is omitted here for simplicity purposes.

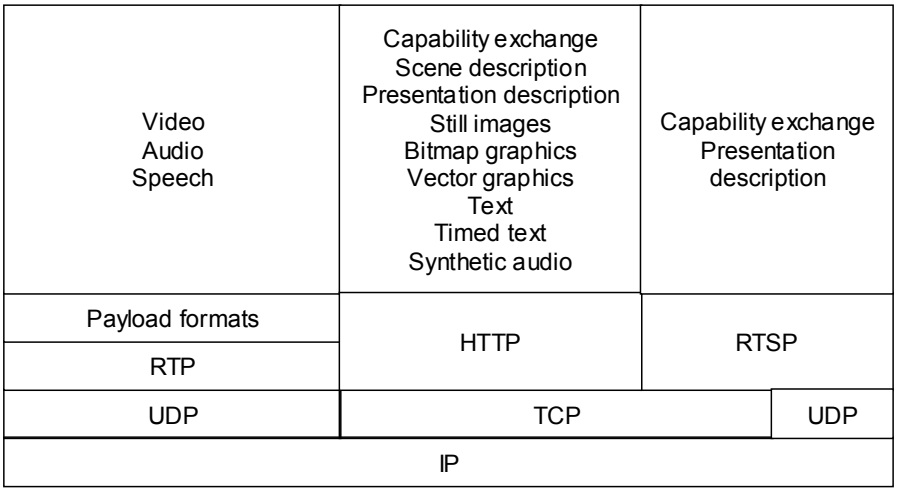


Figure 2.4: Protocol Stack of a PSS Session from [31].

2.3.3 Protocols and Codecs in a PSS Session

The contents of a PSS session can be divided into continuous media, which includes audio, video and speech data streams, and discrete media, which includes still images, text, bitmap graphics, etc [31]. The former category is transported through RTP/UDP/IP, and uses RTSP to set-up and control the individual media streams. The latter one adopts a HTTP/TCP structure and does not need a session set-up and control protocol since this is already built into HTTP [31]. See the protocol stack of a PSS session in Figure 2.4 from [31].

Focusing on the conveyance of continuous media, the transport layer protocol to be employed is the User Datagram Protocol (UDP) because the flow control and the retransmission at end-to-end level performed by TCP is undesirable for RT applications.

UDP, unlike TCP, is unreliable, connectionless, and used for individual message transfer as given by the application (neither SDU segmentation nor concatenation) [34]. There are no sequence numbers involved in UDP, and thereby, no recovery mechanisms for lost datagrams (lack of reliability). It is half-duplex. The main function of a receiving UDP entity is to forward the data to a higher layer process identified by a port. The UDP header contains a checksum field to cover the UDP header and the UDP field. These characteristics make UDP suitable for preserving the time relation between the entities of the RT streaming flows as required by the application. However, UDP does not provide some functionalities, such as framing, sequence numbering, or time stamping, required for the data transfer of continuous media streams.

The Real Time Protocol (RTP) can be considered as a logical extension of UDP to support RT services. It is in charge of time stamping the encoded media frames with the sampling instant of the first data byte of the RTP packet, allowing the synchronization at the receiving side, and hence the removal of the delay jitter introduced by the network. RTP includes a sequence number field to the RTP packets in order to detect packet losses and to restore the packet sequence. The RTP protocol also identifies the format of the RTP payload and determines its interpretation by the application enabling packeting for different types of encoding [35].

The specified codecs in [31] for continuous media streams are:

- *Speech*: Adaptive Multi Rate (AMR) for narrowband speech.
- *Audio*: Motion Picture Experts Group (MPEG)-4 Advanced Audio Coding (AAC) Low Complexity with maximum sampling rate up to 48 kHz.
- *Video*: The mandatory codec to be supported is the H.263 profile 0, level 10.

2.4 Video Streaming Content Creation and Transmission

The speech component of a multimedia stream can be considered as bursty, since it consists of talk spurts and silent periods, and during those silent periods the codec produces none or very few frames. The audio component (e.g. web broadcast from a remote radio station) produces a regular data stream transmission without silent periods with constant length packets. On the other hand, the video component has an inherent variable rate nature, which combined with a bandwidth requirement larger than the other two components, determines the session performance when the video data flow is present in the multimedia stream. Thus, henceforth, the present study will only focus on the video component of the PSS service.

2.4.1 Video Streaming Content Creation

Concerning the bit rate characteristics of the resulting encoded data stream, the video data source encoding can be classified [32] in:

- *Constant Bit Rate (CBR) Encoding*: this type of video encoding generates a constant bit rate encoded data flow by effectively smoothing out the bit rate variations from the video data source. This encoding process limits the number of bits to be used for compression regardless of the difficulty to compress the picture. This technique was originally designed for conversational services because it virtually does not require any Pre decoder or Post decoder buffering at the receiver, which is very attractive for low end-to-end delay communications. However, the final quality of the encoded sequence depends on the complexity of the video content, and arbitrary source data sequences containing scenes with difficult image compression (e.g. image motion) yield a degraded user's perceived quality.
- *Variable Bit Rate (VBR) Encoding*: It provides a constant image quality regardless of the encoding complexity of the video sequence at the cost of generating a encoded data stream with variable bit rate.

CBR encoding is very suitable for the transmission of a variable bit rate data source, like video streaming, over bandwidth limited connections (e.g. dial up access to the Internet or mobile communication systems) [36]. Those connections are characterized not only by a low available bandwidth, but also by the impossibility to temporarily increase the channel bit rate that could accommodate the bit rate peaks from the source.

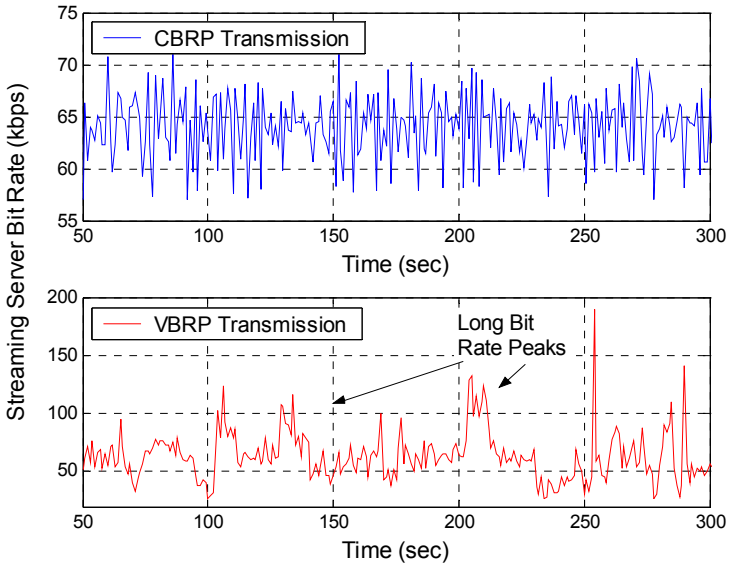


Figure 2.5: Realization of Streaming Server Bit Rate for CBRP & VBRP Transmission.

2.4.2 Streaming Server Transmission of VBR Encoded Video

The streaming server can transmit a VBR encoded video sequence with two different techniques [32]:

- *Constant Bit Rate Packet (CBRP) Transmission*: The streaming server can still transmit a VBR encoded video sequence at a constant bit rate by adjusting the delay between consecutive RTP packets to maintain a near constant bit rate. This transmission technique distorts the time relation between the RTP packets as generated by the codec, and therefore has to be restored at the receiver prior to decoding. The restitution of that time relation can be achieved as a part of the process of the RTP packets time alignment at the receiver (described in section 2.2.1), but requires further buffering capabilities (here denominated *Predecoder buffer*).

- *Variable Bit Rate Packet (VBRP) Transmission*: During the transmission, the server maintains the time relation between consecutive RTP packets as generated by the codec. This implies that the bit rate variations at the encoder output are directly reflected in the channel. Hence, ideally, no Predecoder buffer is required at the client.

If VBRP transmission is to be supported by UMTS, the variations of the data flow bit rate over the average, and the duration of those bit rate increases appear as key factors to determine the required capabilities of the conveying radio bearer.

In order to extract some conclusions about those required capabilities, it is collected a realization from a streaming server when applying CBRP and VBRP transmission from a VBR encoded video trace of a NASA documentary with a target average sequence bit rate equal to 64 kbps (without headers). The video trace is assumed to be a typical video clip (i.e. not a worst case scenario). The full video sequence is longer than 14 minutes, and applies a soft video rate control scheme described in [36] that reduces partially the temporal source rate

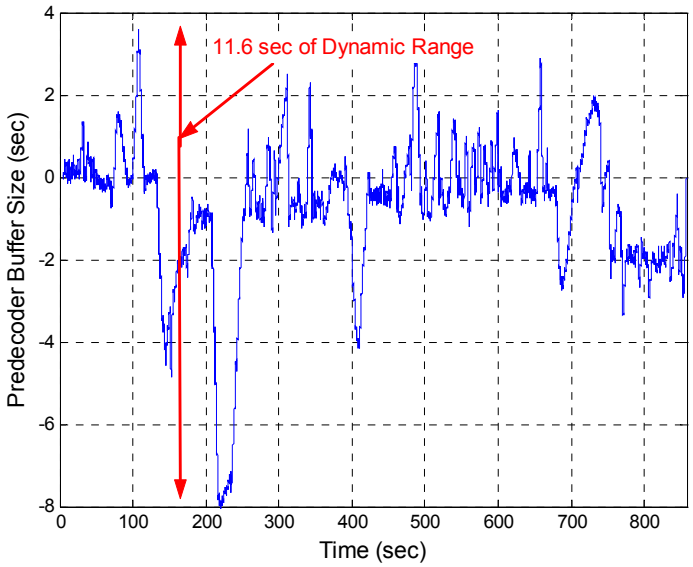


Figure 2.6: Realization of Predecoder Buffer Size Variations for a CBRP Server Transmission.

fluctuations. Figure 2.5 represents around four minutes of the realization. In the figure, the bit rate computation has applied a one second averaging window.

The figure shows that the bit rate of the VBRP transmission fluctuates with long periods that require significant bandwidth capabilities over the average sequence bit rate (up to 10 seconds with bit rates close to double the average). Note that the conveying radio bearer has to be able to support those long bit rate increases if the end quality is to be maintained throughout the whole session. Section 2.5 will deal with the radio bearer support of these bit rate fluctuations. The figure also plots the CBRP transmission corresponding to the same video encoded trace, whereas the long bit rate variations have been smoothed out into a near constant bit rate. However, this conversion from VBR encoding to CBRP transmission is at the expenses of further queuing requirements (Predecoder buffering), which will be quantified in section 2.4.3.

2.4.3 Predecoder Buffer Dimensioning for CBRP Transmission

As described in section 2.4.2, the streaming server can still adapt a VBR encoded sequence to bandwidth limited communications such as dial-up connections to the Internet or mobile communication networks that may not be able to temporarily increase the bandwidth resources to accommodate the bit rate peaks in the range shown in Figure 2.5.

However, the Predecoder buffer required as a consequence of the CBRP transmission process, impose further constraints in the buffering capabilities of the receiver terminal. Figure 2.6 depicts the Predecoder buffer size variations for the CBRP server transmission of the video trace described in section 2.4.2. This computation of the Predecoder buffer variations assumes no delay jitter introduced by the conveying networks.

Comparing Figure 2.5 and Figure 2.6, it can be identified that the period when the VBR encoded sequence[2] generates a bit rate lower than the CBRP transmission causes a Predecoder buffering increase, while a higher bit rate one provokes a Predecoder buffer decrease. Note that the Predecoder buffer may decrease up to negative values, which implies that the corresponding data amount should be cumulated in advance (as a part of the initial playout/buffering delay).

For the analyzed video trace, the CBRP transmission requires a Predecoder buffer dynamic range of up to 11.6 seconds[3] to absorb the bit rate variations from the VBR encoded data stream. Considering that the UE has also to accommodate the Dejitter buffering, these Predecoder buffering requirements are probably too demanding for mobile terminals where the memory capabilities are very limited.

2.5 Interaction between the GBR &MBR UMTS QoS Attributes and the Multimedia Server Transmission

One of the key challenges to successfully accomplish the PSS services through UMTS is to find the appropriate UMTS QoS attributes that will contribute to meet the end-to-end requirements of the streaming sessions.

Knowing that the transfer delay attribute is mainly determined by the Dejitter buffering capabilities of the UE and the maximum delay variation in the fixed Internet, the main focus concentrates on how to set the GBR and the MBR attributes to face with the bit rate variations of a video streaming flow. According to [27], the mapping of application attributes to UMTS bearer service attributes is an operator and/or implementation issue.

In the case of CBR encoding or CBRP transmission, the solution is simple since a constant bit rate flow can be conveyed through UMTS by setting the GBR equal to MBR and equal to the target average video bit rate. Perhaps, a small increase of the aforementioned attributes over the target average video bit rate is convenient to cope with the instantaneous curling (see top part of Figure 2.5).

In the case of VBRP transmission of a VBR encoded sequence, the UMTS bearer has to be able to temporarily increase its bandwidth resources to cope with the transitory bit rate peaks, so that they do not cause a major RTP packet loss during those transitory periods that would lead to a user's perceived quality degradation. However, the bearer capacity is only guaranteed by UMTS up to the GBR attribute, and the remaining bandwidth up to the MBR attribute may only be allocated depending on the network conditions (such as load). Hence, assuming a highly loaded network, there are no guarantees that the temporary bit rate peaks inherent to VBR video encoding will be successfully transferred if they exceed the GBR. At this stage, from a network perspective, two potential solutions are identified for conveying VBR encoded sequences:

[2] Note that the VBRP transmission matches the VBR encoded sequence.
[3] 11.6 seconds is equivalent to 93 kbytes buffering requirements for 64 kbps bearers, and 186 kbytes for 128 kbps bearers.

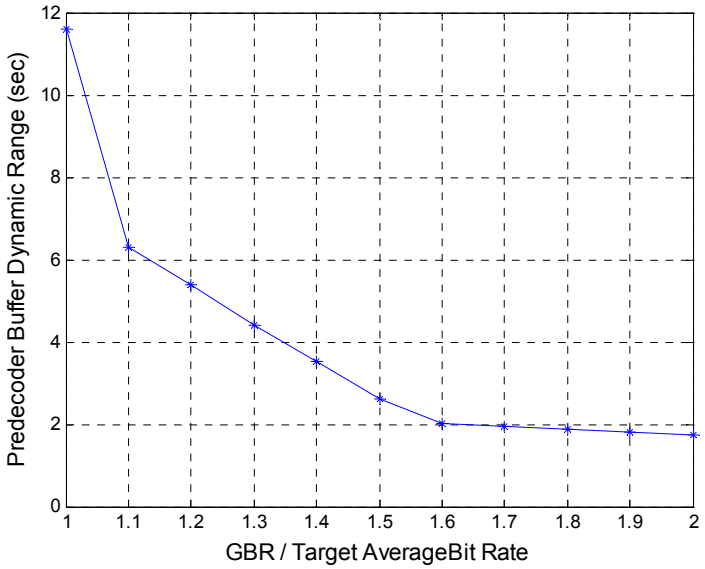


Figure 2.7: Predecoder Buffer Dynamic Range Vs GBRs for the Video Clip Example Described in Section 2.4.2.

- *Statistical Solution*: Setting the GBR and the MBR equal to the target average bit rate and the maximum bit rate indicated by the application respectively, and implementing a statistical admission control algorithm that would allow the multiplexing of the transitory bit rate peaks from multiple streaming users in the cell. This admission control algorithm should statistically guarantee the successful transfer of traffic exceeding the GBR. This solution is beyond the scope of this study.

- *Deterministic Solution*: Limiting the maximum bit rate transmitted by the streaming server and setting the GBR and the MBR equal to this upper limit. The corresponding admission control algorithm should ensure the successful transfer under simultaneous transmission of all users in the cell up to the GBR. Note that this solution represents a system capacity waste during the periods the user bit rate decreases below the GBR.

Focusing on the second solution, by increasing of the GBR (and the MBR) above the target average bit rate of the video sequence, the bearer can partially absorb the bit rate peaks created by the VBR encoder. However, the Predecoder buffer at the receiver can still compensate for the imposition of an upper limit to the bearable bit rate. This represents a hybrid solution between CBRP and VBRP transmission, where a limited increase of the GBR can be translated into a reduction of the Predecoder buffer dynamic range.

Figure 2.7 plots the dynamic range of the Predecoder buffer as a function of the ratio between the GBR and the target average bit rate for the video clip sequence described in section 2.4.2. The result clearly shows that the reduction of the Predecoder buffer capabilities is very significant for small increments of the GBR over the average (e.g. a 15% increase of the GBR implies around 50% reduction of the Predecoder buffer dynamic range), but it does not represent a major benefit to increase the GBR 50 or 60% beyond the average. Although the outcome of Figure 2.7 has been obtained from the specific video clip example described in section 2.4.2, the resulting trend is expected to be general.

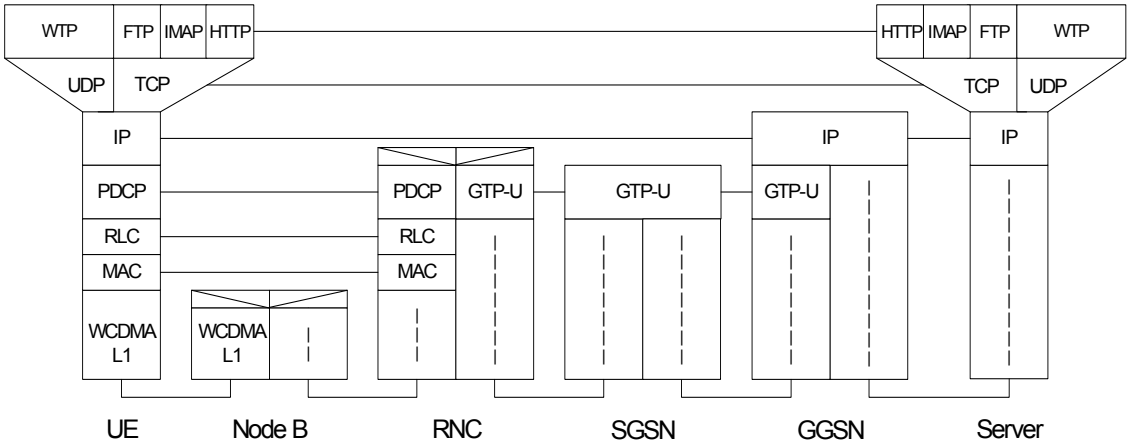


Figure 2.8: Protocol Stack of a NRT Packet Switched Session in UMTS Release 99 Architecture.

2.6 NRT Packet Switched Session

2.6.1 Protocol Architecture of a NRT Session

In contrast to the transparent Packet Switched Streaming service, the 3GPP does not specifically standardize other IP multimedia services. These services are usually based on protocols standardized outside 3GPP, and use 3G networks as bearers to support them.

Figure 2.8 plots the protocol architecture for some typical Non Real Time services based on a TCP/IP Reference Model. On the application layer, four protocols are considered for four different service examples: Hyper Text Transfer Protocol, used for fetching pages on the World Wide Web. The Internet Message Access Protocol (IMAP) that allows the client to access and manipulate mail messages on a server [37]. The File Transfer Protocol (FTP), which is the Internet standard for file transfer [34]. And the Wireless Transaction Protocol (WTP) that provides reliable transport for the Wireless Application Protocol (WAP) datagram service [38].

The WTP protocol runs on top of the UDP transport protocol due to better suitability of the latter protocol for low bandwidth mobile stations. The WTP protocol is in charge of functions such as the information transaction handling, the acknowledgement process, or the error recovery, thus providing the reliability missed by the UDP protocol. The WTP protocol operates in combination with other protocols such as the Wireless Session Protocol (WSP), which is in charge of providing a connection mode WAP service [39] (note that the WSP can also provide a connectionless service operating above a secure or non-secure datagram transport service). See [39] for more information about the WAP protocol stack.

On the other hand, HTTP, IMAP, or FTP require a reliable communication such as offered by the Transmission Control Protocol (TCP). This protocol was designed to provide a reliable end-to-end and full duplex exchange of information over an unreliable Internet network. TCP communications are connection oriented, so there are two end points communicating with each other. Since the Internet Protocol (IP) layer does not guarantee any reliability, TCP provides strategies as retransmission, sequence numbering, flow control, etc. to fulfil the

proper deliver of the data packets as sent by the sender. Next chapter will analyze the effect on TCP flow control on the user and system performance in WCDMA. TCP also supports multiplexing of different applications simultaneously over a single IP address through the use of port numbers.

The Internet routes the IP datagrams conveying the TCP packets towards the UMTS Public Land Mobile Network (PLMN), where the datagrams are processed by the Gateway GPRS Support Node (GGSN), whose function is to connect the Packet Switched core network to the Internet. The Packet Data Protocol (PDP), terminated at the GGSN, is then in charge of conveying the data to the UE. A PDP context is established at the beginning of the packet session between the UE and the GGSN for the data transfer, and it keeps relevant session information such as the allocated IP addresses, the related network elements, or the QoS attributes corresponding to the UMTS QoS class (i.e. interactive or background). After the PDP context is set-up, packet traffic is tunnelled between the GGSN and the RNC with the user plane of the GPRS Tunnelling Protocol (GTP-U). Physically this tunnelling consists of two parts [40], the PS domain backbone bearer between the GGSN and the Serving GPRS Support Node (SGSN), and the Iu bearer over the Iu interface[4]. The SGSN is mainly responsible for mobility related issues.

The data reach UTRAN from the SGSN, where the radio access protocols are in charge of transmitting it through the air interface. The Packet Data Convergence Protocol (PDCP) is responsible for compressing redundant protocol control information such as TCP/IP or RTP/UDP/IP headers that only represent superfluous overhead [41]. The Radio Link Layer (RLC) provides segmentation and link level retransmission services [42]. The Medium Access Control (MAC) layer holds very diverse functionality. Here, it is only cited the functionality of unacknowledged data transfer between MAC entities without data segmentation, and the control of the HS-DSCH transmission and reception. See [43] for more details about the functionality of the MAC layer.

2.7 HTTP

The present section intends to describe the characteristics of interactive services. As the behaviour varies significantly among different services, it is decided to select HTTP for this purpose since it is expected to be one of the key services in UMTS.

HTTP (hypertext transfer protocol) is the protocol used by Web browsing applications. It is a request response protocol designed to make up the parts of the Web documents. HTTP communications usually take place over TCP/IP, though the standard does not preclude HTTP from being implemented on top of any other protocol architecture. HTTP only presumes a reliable transport protocol.

Web pages are written using the Hypertext Markup Language (HTML) [44] as a plain text file commonly referred to as the primary HTML file. In such a file, the textual contents of the page are written as ASCII text, while the in line images are stored as separated objects in different files. These separated objects are commonly named secondary objects. The primary HTML file provides references to such image files as well as references to other Web pages.

[4] The Iu interface is terminated at the RNC and the SGSN.

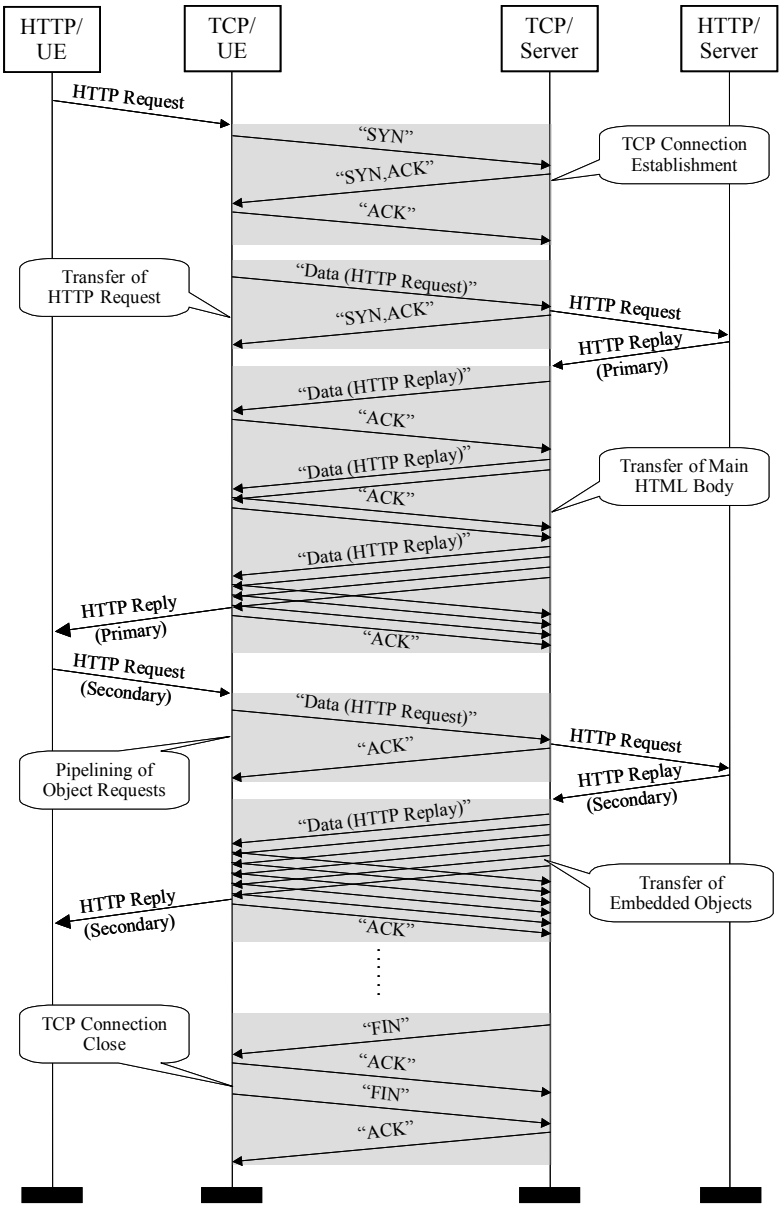


Figure 2.9: Packet Exchange Between UE and Server during Web Page Download.

There exist two versions of the http protocol: HTTP 1.0, and its evolved version: HTTP 1.1. There exist several differences between both versions, but one of the main differences is that in HTTP 1.1 the *persistent connection* is the default behaviour of any HTTP connection [44]. With HTTP 1.0, the retrieval of every file (primary or secondary) requires a TCP connection, which implies a one-to-one mapping between objects and TCP connections. This scheme results in a lot of open and closed connections transferring a small amount of data, which obviously leads to an overhead increase. The earliest Web browsers only utilized a single TCP connection at a time, though later browsers allowed opening several connections in parallel to improve performance. But the utility of this approach is limited by the possibility of increased congestion.

The persistent connection feature keeps alive the TCP connection and allows the retrieval of multiple objects with a single TCP connection. With HTTP 1.1 clients, servers and proxies assume that a connection will be kept open after the transmission of a request and its response. This way, the latency of the subsequent in line objects downloads is reduced since there is no time spent in multiple TCP connection establishments. Moreover, a client employing HTTP 1.1 can pipeline its requests (i.e. send multiple object requests without waiting for each response), which results in a lower page download time.

Figure 2.9 depicts the packet exchange between the UE and the server during web page download process. A persistent connection is assumed in the picture. Whenever the user wants to retrieve a web page, it transmits an HTTP request that specifies the HTML file to be downloaded. The request contains the Uniform Resource Locator (URL) of the page, which effectively serves as the page's worldwide name. Prior to the HTTP request transmission, the UE has to establish the TCP connection with the server, which comprises the three-way handshake between them. The *SYN* TCP segments and their corresponding acknowledgements involved in the connection set-up procedure only include the TCP/IP headers, i.e. they are 40 bytes long. In UTRAN, assuming no dedicated channel (DCH) was previously allocated for this user, such small data packets could be transmitted on common transport channels such as the Random Access Channel (RACH), the Forward Access Channel (FACH), or the Uplink Common Packet Channel (CPCH). The duration of the three-way handshake is mainly dependent on the round trip time of the connection. After the connection establishment, the UE can proceed to transmit the HTTP request. Since the request only carries the URL address of the primary HTML file, its size is no longer than a few hundred bytes (see Table 2.2).

Proposed Distribution	Mean	Reference	Year Of Publication
Bimodal	320 B	[16]	1997
Log Normal	300 B	[17]	1997

Table 2.2: HTTP Request Size. The Publications in the Table Obtain their Parameters Through Statistical Measurements.

The server reads the primary HTML file from its disk and begins transferring the data to the client. During this phase, the file is segmented into TCP packets whose transmission is ruled by the slow start algorithm of the TCP protocol. See the analysis of the slow start algorithm in Chapter 3. At this stage, the downlink transmission of data is too heavy to be conveyed by common transport channels, and the Radio Resource Management (RRM) triggers the establishment of a dedicated channel for this user (note that the transfer procedure could also be done over the DSCH or the HS-DSCH).

After receiving the HTML main body, the UE parses the file and extracts the URL of the embedded objects. The average parsing time was measured in [24] to be 130 ms. After parsing the primary file, the client pipelines all the requests of the secondary objects of the page without waiting to receive every HTTP reply. Figure 2.9 only includes a single secondary HTTP request for simplicity purposes. The average number of objects in a Web page is relatively low (below seven objects per page). The number of objects also depends on the considered scenario because, for example, it is common for corporative users to retrieve pages with larger number of in line objects than for researchers who usually require text-oriented pages. See Table 2.3.

Proposed Distribution	Mean	Reference	Year Of Publication
Not Offered	2.8-3.2	[16]	1997
Delta + Lognormal	6.7	[19]	2001
Gamma	2.2	[17]	1997
Truncated Pareto	6.55	[24]	2001
Geometric	2.5	[23]	2001

Table 2.3: Total Number of Objects Per Page (Including Primary and Secondary Objects). Publications in Shaded Cells Obtain their Parameters Through Statistical Measurements, while Publications in Non Shaded Cells Are the Result of a Simulation Model Proposal.

Regarding the size of the objects, the primary file is commonly larger than the embedded ones. The difference is due to the dissimilar type of files being transferred. Primary files are usually HTML text files, whereas secondary files are commonly images. In [20], it is shown that the files available in the Web are heavy-tail distributed, which is the main reason behind the probability mass in the tail of the distribution of the HTML transferred files. This is the primary determiner of the self-similar nature of aggregate Web traffic [20]. This means that significant traffic variance (burstiness) is present within all time scales (larger than 1 second according to [20]). The packet loss probability and delay behaviour of HTTP traffic is radically different under the presence of such long-range traffic correlation. Then, the object sizes (and ultimately the Web page sizes) are commonly described through heavy-tailed distributions as the Pareto function. However, for convergence reasons, simulation models usually truncate heavy-tail distributions, or use fat-tail functions such as the Lognormal or the Weibull distributions. See Table 2.4.

Proposed Distribution	Mean	Reference	Year Of Publication
Lognormal	5 kbytes	[19]	2001
Lognormal	7.7-10.7 kbytes	[45]	1999
Lognormal (Main Body)	10.7 kbytes	[24]	2001
Lognormal (In Line)	7.8 kbytes	[24]	2001
Geometric	3.7 kbytes	[23]	2001

Table 2.4: Characterization of the Objects Size. Publications in Shaded Cells Obtain their Parameters Through Statistical Measurements, while Publications in Non Shaded Cells Are the Result of a Simulation Model Proposal.

When all the objects are finally downloaded, the end user (human) reads the contents of the page before making the request for the next web page. See the description of the reading time in Figure 2.10. This parameter is completely determined by human behaviour and not by the network, and according to [20] this behaviour is also heavy-tail distributed. In order to assess the statistics of the reading times, experimental investigations face the difficulty of distinguishing between the interval the user is examining the downloaded page and the period when the user is not utilizing the browser at all. Hence, most of the experimental investigations filter out the reading times larger than a given threshold (e.g. 3 minutes in [17]).

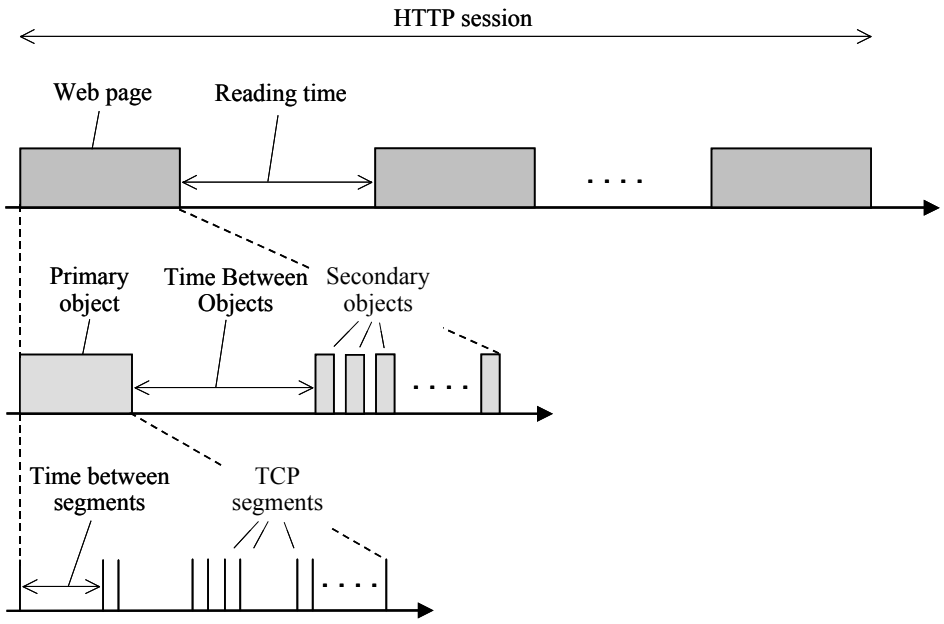


Figure 2.10: Contents of a HTTP session.

However, the selected threshold has significant impact on the resulting average reading time of the proposed distribution.

On the other hand, simulation models are usually tailored to reduce the simulation time. For example, in [46], it is proposed to reduce the average reading time down to 5 seconds. See Table 2.5.

Within a Web browsing session, users commonly retrieve more than a single Web page. In the reading time between two consecutive page downloads, the HTTP persistent connection is assumed to be kept open until either the client or the server explicitly closes it. As commented before, persistent connections reduce latencies and overhead, but on the other hand it consumes resources at both the server and the client (such as buffer memory or a socket). These idle resources can cause the detriment of the server's throughput. Hence, there exists the need for policies that manage the duration of such idle connections.

Proposed Distribution	Mean	Reference	Year Of Publication
Lognormal	61 sec	[19]	2001
Weibull	21 sec	[17]	1997
Not Offered	81 sec	[21]	1997
Geometric	5 sec	[46]	2000
Exponential	30 sec	[24]	2001
Exponential	12 sec	[23]	2001

Table 2.5: Characterization of the Reading Time. Publications in Shaded Cells Obtain their Parameters Through Statistical Measurements, while Publications in Non Shaded Cells Are the Result of a Simulation Model Proposal.

Proposed Distribution	Mean	Reference	Year Of Publication
Not Offered	19.6	[21]	1997
Lognormal	23-26	[45]	1999
Geometric	5	[48]	1998
Geometric	5	[23]	2001

Table 2.6: Characterization of the Number of Pages Per Browsing Session. Publications in Shaded Cells Obtain their Parameters Through Statistical Measurements, while Publications in Non Shaded Cells Are the Result of a Simulation Model Proposal.

According to [47], the version 1.3 of the Apache HTTP server limits the holding time of a connection to 15 seconds. The holding time starts when the connection is established or when a new HTTP request arrives, and the server closes the connection when the holding time expires. Figure 2.9 includes the TCP termination procedure. With the model proposed in [19], around 50% of the reading times have a duration longer than 15 seconds. This implies that, in around 50% of the cases, this server will close the TCP connection before the user finishes the reading time and selects the subsequent Web page (even if the user selects both URL located in the same server). This has strong implications at transport level because new TCP connections will have to be established with the subsequent delay and with the effect of the TCP's slow start algorithm. Table 2.6 gives the average number of Web page downloads in a browsing session.

2.8 Round Trip Time Delay Budget

The round trip time (RTT) is a merit figure of any connection because it gives an indication on how fast the transmitter can react to any event that occurs in the connection. It could be defined as the elapsed period since the server transmits a packet until it receives the corresponding acknowledgement. With the purpose of accelerating such transmitter response time, the round trip time should be minimized as much as possible.

The round trip time is a very relevant factor for TCP connections, whose performance completely depends on this figure (as it will be shown in Chapter 3). For this reason, and due to the special focus paid to the TCP protocol in various chapters of this Ph.D. thesis, this section provides an estimate of the potential range of the round trip time between a server located in the Internet and a mobile terminal in UTRAN. The selected transport channel for data conveyance within UTRAN is the Dedicated Channel (DCH). This estimation should be considered as a first order approximation, because the actual magnitude is heavily dependent on factors such as implementation, load, physical distance between transmitter and receiver, etc.

The RTT can be decomposed on the delay components introduced by the Internet, the Core Network, UTRAN, and the UE. Table 2.7 describes the one-way downlink UMTS delay budget for a 40 bytes packet over a DCH with a bit rate of at least 32kbps. The components of the Table 2.7 do not include RLC retransmission delays caused by the unsuccessful decoding of the coded block. The Internet delay estimate is obtained from [49], the CN and RNC delays are obtained from [50], whereas the Node B and UE delays are obtained from [51]. Soft

Network	Element	Delay	Remarks
Internet		25 – 125 ms	Highly dependent on number of hops. Considerably higher in overseas connections.
CN	GGSN	1 – 2 ms	Switching and processing delay.
	Network Transport	18 – 41 ms	Propagation and queuing delay across the network.
	SGSN	1.5 – 4 ms	Switching, processing, and Iu contribution.
UTRAN	RNC	3 – 10 ms	Processing, packetization, and Iub delay.
	Node B	12 – 14 ms	CRC attachment, encoding, 10 ms interleaving, spreading and modulation.
	UE	15 ms	Demodulation, 10 ms deinterleaving, decoding, and higher layers processing.

Table 2.7: One-way downlink delay budget over Dedicated Channels in UMTS for a 40 bytes IP datagram. Assumed at least 32 kbps Bit Rate for DCH, and no SHO. RLC retransmissions not Included.

handover case is not considered in this estimate, and an interleaving depth of 10ms has been assumed.

As it can be seen in Table 2.7, the network transport through the CN, and the interleaving and deinterleaving in UTRAN are substantial sources of delay. However, the conveyance through Internet is likely to be one of the main contributors to the overall budget. One potential solution to eliminate this contribution is to include a Proxy server inside the operator's Core Network. Such Proxy server would contain the user's requested information (for example by storing the most visited web pages), which would suppress the user's need for accessing a remote URL in the Internet.

Connection	RTT
With Proxy Server	114 – 189 ms
Without Proxy Server	164 – 439 ms

Table 2.8: Round Trip Time Range Estimation in UMTS for a 40 bytes IP datagram. Assumed at least 32 kbps Bit Rate and 10% BLER for both Uplink and Downlink DCH.

The overall round trip time estimate can be obtained from Table 2.7 by adding the average RLC retransmission delay, and by considering the effect of both downlink and uplink. The RLC retransmission delay is simply computed as the RNC, Node B and UE delays (uplink and downlink) multiplied by the average number of RLC retransmissions. Note that the contribution of the RLC polling mechanisms is not included. It will be further assumed that the one-way uplink budget can also be taken from Table 2.7. The resulting estimation for a 32 kbps and 10% BLER for both uplink and downlink DCH is depicted in Table 2.8.

The results of Table 2.8 show how for a 40 bytes IP datagram, the Core Network and UTRAN round trip time ranges in the order of 115 – 190 ms, which is in relative accordance with [10] that states "UMTS network round trip time without radio transmission equals 150 milliseconds". However, the introduction of Internet delays significantly raises the overall round trip time magnitude (especially in overseas connections).

2.9 Concluding Remarks

The present chapter has aimed at building understanding in key areas such as the QoS demands of the different traffic classes, the protocols employed in their communication, and the statistical properties of the traffic. The presented characteristics are fundamental for the RRM system design and performance analysis carried out in the following chapters.

The main pattern of the streaming, interactive and background services have been described. There exist relevant QoS attributes of the streaming bearers such as the transfer delay, guaranteed bit rate, SDU error ratio, that impose stringent service guarantees on the network to support such services. On the other hand, the 3GPP does not set any absolute service guarantees on delay or bit rate for interactive bearers and their absolute performance depend on the actual network load and admission control policy of the network. However, regarding interactive bearers, the QoS specification [13] states that "at the destination there is an entity expecting the message within a certain time". Similar absolute service guarantees have not been defined for background bearers, and the network is to transfer data only "when there is definite spare capacity" [13].

The protocol stack of streaming, interactive and background services has been described too. Regarding the Packet Switched Streaming Session, two different types of video encoding and have been presented: (i) Constant Bit Rate (CBR) video encoding that generates a constant bit rate data flow, which is very suitable for transmission over bandwidth limited connections that can not temporarily increase the channel bit rate to accommodate the bit rate peaks from the source. However, this bit rate smoothing is at expenses of the degradation of the final image quality. (ii) Variable Bit Rate (VBR) encoding that provides a constant image quality at the cost of generating an encoded data stream with variable bit rate. For the video trace example considered in this study, the bit rate peaks of the VBR encoded flow reached up to 10 seconds with bit rates close to double the average. It has also been shown that the server can still transmit a VBR encoded video stream at a near constant bit rate through the Constant Bit Rate Packet (CBRP) transmission technique, but it seems to require too large Predecoder buffering (around 11 seconds for the video trace example analysed in this chapter) to be a viable solution.

The implications of these streaming traffic characteristics to the GBR and MBR bearer attributes have been exposed, and two alternative attribute settings have been identified: (i) Constant Bit Rate (CBR) encoded flows can be simply transmitted through UMTS by setting the GBR and the MBR equal to the encoder bit rate and it does not require long Predecoder buffering. (ii) VBR encoding streams can still be transferred by a hybrid solution between CBRP and VBRP transmission by setting the GBR equal to the MBR and relatively higher (e.g. 15-60%) than the encoder average bit rate. With this hybrid solution, the limited increase of the GBR is directly translated into a reduction of the Predecoder buffer size. For example, for the video trace case considered in this study, a 15% increase of the GBR implies around 50% reduction of the Predecoder buffer dynamic range.

Regarding interactive services, the focus has turn to Web browsing applications. It has been presented that every browsing session is composed by several page downloads separated by periods when the end user (human) analyses the retrieved information. With HTTP 1.1, multiple objects within a Web page (usually pages are composed by no more than seven objects on average) can be downloaded in the same TCP connection, which reduces the delay and overhead of establishing new TCP connections, and avoids new slow starts. However, a

single TCP connection can not always be employed for multiple page downloads from the same server because the server's resource management policy limits the duration of idle persistent connections. If the reading time outlasts the server's holding time of the server (e.g. 15 seconds for a common Internet server), then new TCP connections will be needed for subsequent pages with the negative effect the TCP connection set-up. It is also interesting to observe that the download is accompanied with significant amount of small packets such as TCP/IP headers for connection establishment and close, acknowledgements and HTTP requests. The RRM must ensure that, unless a Dedicated Channel is already set-up, those packets are transmitted via common transport channels that reduce the transmission time (because they do not need to be explicitly configured as DCH) and avoid channelization code shortage (caused by inactive DCH channels).

Chapter 3

TCP Performance over WCDMA

3.1 Introduction

As stated in previous chapter, packet data services are the distinctive characteristic of Third Generation wireless networks. In many cases, the information provided by these services will be accessible on the Internet. According to measurements carried out by Miller in April 1998 [52], 95% of the overall data traffic that crossed a middle Internet backbone was TCP traffic. The majority of Internet's traffic is implemented over TCP because this is a stable, mature, and probably the most thoroughly tested transport protocol of its kind. Most of the interactive and background services to be addressed by WCDMA networks are expected to belong to this vast group of applications employing a TCP/IP platform, which defies UMTS to properly handle the TCP/IP traffic. As also described in previous chapter, the properties and characteristics of the incoming data to the WCDMA radio network completely depend on the protocol stack of every service. This chapter analyses the performance of TCP over UMTS due to the wide application of this protocol and the strong impact of its flow control on the behaviour of the traffic.

It is well known that the design of the TCP protocol has been optimized to cope with the typical packet loss pattern of wired networks and not for wireless ones. In wired networks, the packet losses caused by link level errors are typically very infrequent (less than 1% packet loss rate according to [34]). However, the congestion in wired networks may easily provoke

the buffer overflow of any router in the network, with the ultimate consequence of multiple packet drops at a time.

In wireless networks the main source of packet losses is the link level errors of the radio channel, which may seriously degrade the achievable throughput of the TCP protocol [53]. Nevertheless, the Radio Link Layer (RLC) of UMTS includes an Automatic Repeat reQuest (ARQ) mechanism to recover from the packet losses, which can hide the unsuccessful transmissions of the wireless link. According to [54]-[55], with a maximum number of retransmissions at the ARQ mechanism larger than three (and a BLER lower than ~ 30%) a TCP throughput close to the maximum achievable can be obtained in CDMA systems. However, the studies of the aforementioned papers assume unrealistically low round trip time delays.

The operation of the ARQ mechanism of the RLC layer represents a trade off between error correction capabilities and variable transmission delay. The mitigation of the radio link errors by RLC retransmissions causes delay spikes of the TCP segments involved in those link level retransmissions. The interaction between these delay spikes and the TCP retransmission timer must be addressed in order to evaluate the overall impact of the radio link transmission errors in the performance of TCP.

Besides the effect of the radio link errors, the present chapter also assesses the influence on the end user throughput of the TCP flow control mechanisms such as the slow start algorithm and the memory capabilities of the mobile terminal.

The chapter is organized as follows: section 3.2 gives a general description of the TCP protocol focusing on the slow start and the congestion avoidance algorithms, and the TCP retransmission mechanisms. The bandwidth delay product of TCP connections is notably covered in section 3.2.4 due to its relevance on the overall performance of the protocol. Section 0 assesses the influence of the initial slow start for small files, whereas section 3.3.2 describes the channelization code efficiency for different channel bit rates. Section 3.3.3 highlights the influence of the bandwidth delay product and the receiver window on the TCP performance. Section 3.3.4 describes the impact of random segment losses on the protocol behaviour, and section 3.3.5 analyses the interaction between the reliability of the RLC layer and the TCP performance. Section 3.4 finally draws the main conclusions of the chapter.

3.2 TCP Description

As commented in the previous chapter, the TCP protocol was designed to provide a reliable end-to-end and full duplex exchange of information over an unreliable Internet network. TCP communications are connection oriented, so there are two end points communicating with each other. TCP provides the reliability by means of flow control, sequence numbering, and retransmission of lost data.

In TCP the data is split into what the protocol considers the optimum size chunks to transmit. The chunks are denominated *segments,* and each end can announce at the establishment of the connection the maximum segment size (*Maximum Segment Size* or *MSS)* expected to be received. The resulting IP datagram is 40 bytes longer due to the TCP and IP headers. If one TCP end does not receive a MSS option from the other end, a default value of 536 bytes is assumed [34]. In general, the larger MSS the less overhead is included. But a too large

Network	MTU (bytes)
Official Maximum	65535
16 Mb. IBM Token Ring	17914
FDDI	4352
Ethernet	1500
IEEE 802.3	1492
Arpanet	1006
X.25	576
Point-to-Point (low delay)	296

Table 3.1: Common MTUs in the Internet from [56].

segment may produce the fragmentation in any of the different networks crossed through the Internet. The *Maximum Transmission Unit* (*MTU*) is the largest packet that a network can transmit without fragmentation. Table 3.1 shows the MTU of typical data networks [56]. Therefore, it is recommended to set the MSS lower than the lowest MTU of the crossed networks. Note that the MTU describes the full packet size (i.e. including the datagram headers). Typical values of MSS are 256, 512, 536 or 1460 bytes.

From a wireless performance point of view, one of the most relevant characteristics of TCP is the flow control. The flow control is in charge of determining the load offered by the sender to achieve maximum connection throughput while preventing network congestion or receiver's buffer overflow. The flow control in TCP is accomplished by means of two windows: the *congestion window* (*cwnd*) and the *advertised window* (*awnd*). The congestion window is controlled by the *slow start* and the *congestion avoidance* algorithms that try to optimize the amount of data injected to the load situation and the capacity of the Internet. The awnd is a window with the size the receiver has buffer for, and the receiver advertises it in every segment transmitted to the server. The advertised window is therefore flow control imposed by the receiver and it is upper bounded by the buffer size of the UE. Typical TCP buffer sizes in mobile terminals are expected to be 8, 16, and 32 kbytes. At any time, the minimum of the congestion window and the advertised window determines the amount of data the server can have outstanding.

3.2.1 Slow Start

After the TCP connection establishment, the *slow start* algorithm controls the congestion window, and therefore the initial segment transmission pace. In 1988, Jacobson [57] proposed this algorithm that gradually increases the amount of data on the fly. More specifically, in slow start, the congestion window is initialised to one MSS, and upon the reception at the server of an acknowledgement that acknowledges new data, the congestion window increases one MSS. Despite its name, this algorithm produces an aggressive exponential increase of the sender transmission rate with the objective of filling the bottleneck rapidly, and hence, having a negligible impact on the throughput performance. Moreover, the algorithm guarantees that a connection will source data into the network at a rate at most twice the maximum of the limiting link [57]. Recent enhancements permit to initialise the congestion window to two or

four MSS at the beginning of the TCP connection [58], which contributes to fill the connection bottleneck even faster.

The TCP receiver can not generate acknowledgements faster than the segments get through the network, which is to be used by the server to identify the arrival rate of the data at the receiver. This is the so-called *self clocking* behaviour of TCP, which permits the sender to automatically adapt to bandwidth and delay variations of the connection. More specifically, an acknowledgement should be generated for at least every second full-sized segment, and must be generated within 500 ms of the arrival of the first unacknowledged packet [59]. Most implementations use a 200 ms delay [34].

However, upon reception of an out of order segment, the client should deliver a duplicate acknowledgement immediately, because the reception of such an out of order segment could be the result of a segment loss in the network. Hence, with the purpose of informing swiftly the server about the out of order segment, this duplicated acknowledgement must not be delayed. As it will be described in section 3.2.3, upon the reception of a number of consecutive duplicated acknowledgements, the server considers the segment as lost and triggers a fast retransmission of the missed segment.

Assuming the generation of an immediate acknowledgement after the reception of every data segment, a constant round trip time, and an error free connection, the congestion window evolves during the slow start phase as described in Table 3.2. In the table, *RTT* denotes the connection round trip time, and μ represents the average bandwidth of the bottleneck link (expressed in segments per second).

Cycle	Time	Cwnd (MSS)
0	0	1
1	RTT	2
2	$2 \cdot RTT$	3
	$2 \cdot RTT + 1/\mu$	4
	….	
n	$n \cdot RTT$	$2^{n-1}+1$
	$n \cdot RTT + 1/\mu$	$2^{n-1}+2$
	….	
	$n \cdot RTT + (2^{n-1}-1)/\mu$	$2^{n-1}+2^{n-1} = 2^n$

Table 3.2: Evolution of the Congestion Window During Slow Start.

3.2.2 Congestion Avoidance

While the slow start algorithm opens the congestion window quickly to reach the capacity of the limiting link rapidly, the congestion avoidance algorithm is conceived to transmit at a safe operating point and increase the congestion window slowly to probe the network for more bandwidth becoming available.

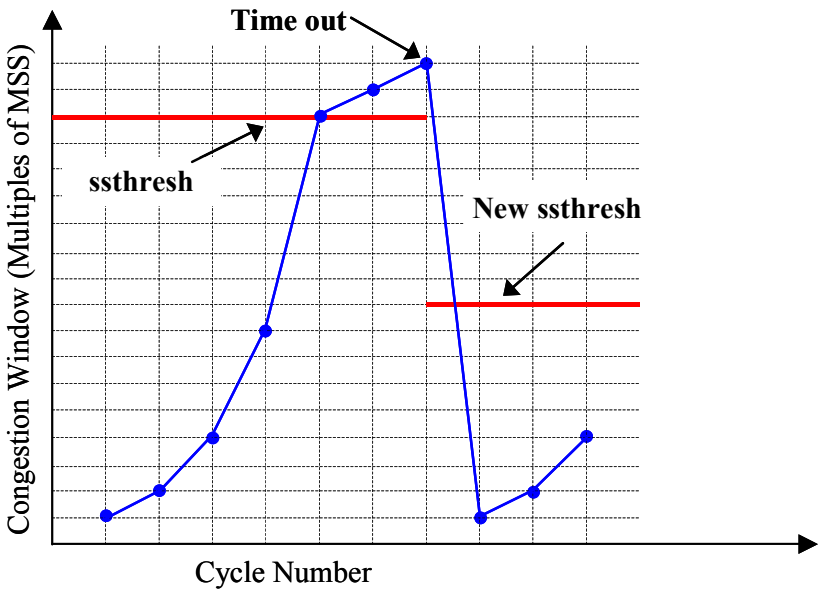


Figure 3.1: Congestion Window Evolution During Slow Start And Congestion Avoidance.

The congestion avoidance takes over the control of the evolution of the congestion window when cwnd overcomes the so-called slow start threshold *ssthresh*. According to [59], "the initial value of ssthresh may be arbitrarily high […], but it may be reduced in response to congestion". Some implementations simply initialise ssthresh equal to the size of the window advertised by the receiver. After congestion, the slow start threshold is normally reduced as it will be described in following subsections. During congestion avoidance upon the reception of an ack that acknowledges new data, the server updates the congestion window W as:

$$W = W + {}^{1}\!/_{\text{int}\lfloor W \rfloor} \tag{3.1}$$

where W is measured in segments. This congestion window update represents a linear enlargement because the cwnd is increased by one segment in every cycle. See in Figure 3.1 the congestion window evolution during slow start and congestion avoidance.

3.2.3 TCP Retransmissions

TCP has two mechanisms to deal with packet losses: a retransmission timer, and the fast retransmission and fast recovery algorithms. Early implementations (TCP Tahoe) only included the retransmission timer, while modern versions (TCP Reno and newer) also incorporate the fast retransmission and fast recovery algorithms.

3.2.3.1 TCP Retransmission Time Out

The TCP transmitter uses a retransmission timer to ensure the data delivery. The duration of this timer is denominated RTO (retransmission time out). The timer is started whenever it is not running and the sender delivers a data segment. The timer is reset whenever an ack that acknowledges new data is received, and it is turned off when all the outstanding data has been acknowledged [60]. When the retransmission timer expires, the sender retransmits the earliest

segment that has not been acknowledged and restarts the timer with an RTO double of previous value. The packet loss is considered as a signal of congestion in the network, and after the retransmission of the first segment the congestion window is set to the initial number of segments, the slow start threshold is set to half the actual window size, and the slow start takes over the control [59].

The RTO is set to cope with variations of the round trip time (both in terms of average and standard deviation). For that purpose, the server keeps track of the smoothed RTT (estimator of the mean) and the smoothed mean deviation (estimator of the standard deviation). Whenever a new RTT measurement *R* is carried out, the smoothed round trip time (SRTT) estimator and the round trip time variation (RTTVAR) are updated as follows [60]:

$$SRTT = (1-\alpha)\cdot SRTT + \alpha \cdot R \tag{3.2}$$

$$RTTVAR = (1-\beta)\cdot RTTVAR + \beta \cdot |SRTT - R| \tag{3.3}$$

where the values of alpha and beta should be set equal to 1/8 and 1/4 respectively. Note that both estimators are updated with a low pass filter, which challenges the possibility of tracking sudden changes of the RTT. When the first RTT measurement is made, the SRTT is set equal to the measurement and the RTTVAR is set to half of it. Finally, after every RTT measurement, the RTO is computed as:

$$RTO = SRTT + \max(G, 4\cdot RTTVAR) \tag{3.4}$$

where *G* indicates the granularity of the clock at the server. The resulting RTO should be rounded up to one second if the result of equation (3.4) is lower than one second.

A TCP implementation must take at least one RTT measurement per cycle [60], and in any case the RTT samples must not be made using segments that were retransmitted (Karn's algorithm).

3.2.3.2 TCP Fast Retransmission and Fast Recovery

The second mechanism for dealing with segment losses is the Fast Retransmist and Fast Recovery algorithms, which are typically implemented together as described in [34], though they are two different algorithms. The TCP Fast Retransmit algorithm aims at detecting any segment loss without waiting for the retransmission timer to expire. With this algorithm, after receiving a small number of duplicate algorithms (typically three) the sender infers that a segment is lost and retransmits it. The Fast Recovery algorithm makes use of the fact that the reception of several duplicate acknowledgements indicates that there is still data flowing between both ends, and therefore an abrupt reduction of the congestion window size is not convenient. Hence, this algorithm prevents the pipe going empty, thereby avoiding the later need for the slow start. These two algorithms are typically implemented as described in Table 3.3.

Upon reception of three duplicated acks
Retransmit the missing segment
Set ssthresh to max((segments on the fly)/2 , 2 · MSS)
Set cwnd to ssthresh + 3
Every new duplicate ack increases cwnd by one, & transmit packet if allowed by new value of cwnd
When the new ack arrives
Set cwnd to ssthresh
Execute congestion avoidance

Table 3.3 Combined Implementation of Fast Retransmission and Fast Recovery Algorithms.

3.2.4 Bandwidth Delay Product (BDP)

The bandwidth-delay product is a merit figure of the amount of outstanding data in the ideal steady state of a TCP connection. It can be computed by simply:

$$BDP = \mu \cdot RTT, \qquad RTT = rtt + \frac{1}{\mu} \Rightarrow \quad BDP = \mu \cdot rtt + 1 \tag{3.5}$$

where μ represents the bandwidth of the limiting link of the connection measured in segments per second, and RTT is the connection round trip time measured in seconds. The BDP in equation (3.5) is measured in segments. The round trip time in previous equation has been further decomposed into the average delay to transmit a segment $1/\mu$ by the radio link and the remaining round trip time rtt (composed by the delay in Internet, Core Network, RNC processing delay, Iub interface, UE processing delay, etc). The factor rtt could also be interpreted as the round trip time of a small packet with a negligible transmission delay in the TCP connection bottleneck.

Ideally the congestion window should be as close as possible to this product because lower windows cause the limiting link (radio in this case) to be underused. On the other hand, higher congestion windows imply more data than what the limiting link can handle, which causes the data excess to be cumulated in the buffer of the bottleneck link:

$$W = \mu \cdot RTT + B \tag{3.6}$$

where W denotes the congestion window, and B represents the amount of queued data (both metrics measured in segments)[5]. Large amounts of data excess to be queued in the RNC may cause buffer overflow, with the subsequent segments losses.

The UE's buffer size has to be at least as large as the BDP or otherwise the receiver will limit the connection window with the consequent radio resource waste. Table 3.4 includes several examples of the BDP for various channel bit rates and round trip times. See section 2.8 about round trip estimates for UMTS.

[5] Note that the RTT metric in equation (3.6) excludes the queuing delay at the buffer of the bottleneck link.

Radio Channel Bit Rate	rtt		
	150 ms	300 ms	450 ms
32 kbps	1.05 kbytes	1.58 kbytes	2.11 kbytes
64 kbps	1.58 kbytes	2.63 kbytes	3.69 kbytes
128 kbps	2.63 kbytes	4.74 kbytes	6.85 kbytes
256 kbps	4.74 kbytes	8.96 kbytes	13.18 kbytes
384 kbps	6.85 kbytes	13.18 kbytes	19.51 kbytes

Table 3.4: Bandwidth Delay Products for Various WCDMA/DCH Bit Rates and Round Trip Time of Small Packets rtt (Assuming FER Target of 0.1 and a MSS 536 bytes).

In [65], the term high bandwidth-delay product is employed to describe those BDPs that cause "that the buffering on the bottleneck link is typically of the same order of magnitude as, or smaller than, the bandwidth delay product". According to this definition it could be stated that, in general, the BDP in WCDMA is small (for example compared to WANs), so that buffering on the radio link can be larger or sometimes much larger than the BDP. However, the buffering on the limiting link also depends on other diverse factors such as the actual receiver's buffer, the congestion situation of the network, the probability of segment losses, the size of the file to be downloaded, etc.

As previously stated, during the periods of the connection when the congestion window is lower than the BDP, the radio channel resources are not fully utilized. This has disadvantages for the end user who experiences a connection throughput below the possibilities of the radio link. But it also has disadvantages for the system, because code (and possibly power) resources are reserved for the allocation of the conveying Dedicated Channel without fully exploiting them. The allocation of high bit rates DCHs to WWW applications has been reported in [12] to lead to code shortage situations in non-time dispersive environments. This code shortage situation may impose a more restricting limit to the cell capacity than the overall network interference. This effect may impede the allocation of high bit rate DCHs to bursty applications in microcells (low time dispersive environments).

In macrocell environments the channelization code shortage caused by the traffic inactivity derived from low congestion windows is likely to have less impact on the cell throughput because the time dispersion typical of such environments causes interpath interference, which seriously impairs the cell capacity. This code limitation impact may still easily be non negligible. For example, the maximum downlink capacity with one scrambling code, 1/3 of coding rate and a 20% of soft handover overhead is described in [10] to be 2.5 Mbps, whereas the same author reports a cell throughput of 915 kbps for macrocell Vehicular A environments with a fair distribution of the resources in the cell. In any case, due to the aforementioned disadvantages not only for the system but also for the end user, the link utilization will be an aspect of special consideration throughout this chapter.

In order to assess the period required by the slow start to reach the BDP, let assume as in section 3.2.1 an error free connection and the generation of an immediate acknowledgement after the reception of every data segment. Let further assume that the radio link capacity is reached in cycle N. That means that all acknowledgements corresponding to segments released by the server in cycle N will not be able to arrive within RTT after the start of cycle N+1. This condition can be expressed as:

$$\frac{2^N}{\mu} \geq RTT \Rightarrow \quad N \geq \text{int}\lfloor \log_2(rtt \cdot \mu + 1) \rfloor \tag{3.7}$$

where N represents cycle number when the 100% radio link utilization is reached.

Table 3.5 evaluates the cycle that produces 100% radio link utilization for different segment sizes, radio link bit rates and round trip times. The BLER target assumed for the Table 3.5 is 10 %, and the header size is assumed to be totally compressed by the PDCP layer of UMTS. The results show that the lower bandwidth delay product, the earlier cycle reaches the radio link capacity limit. However, the allocation of high bit rates produces poorer radio link utilization during the slow start phase. A larger maximum segment size also contributes to a better link utilization during slow start. Note as well that after the capacity of the radio link has been reached, the overall connection round trip time evolves as the left most part of (3.7), which represents an exponential increase of the round trip time in slow start.

MSS	**536 Bytes**		**1460 Bytes**	
rtt	**150 ms**	**300 ms**	**150 ms**	**300 ms**
Rb = 32kbps	2	2	1	1
Rb = 64kbps	2	3	1	2
Rb = 128kbps	3	4	2	2
Rb = 256kbps	4	5	2	3
Rb = 384kbps	4	5	3	4

Table 3.5: Cycle Number that produces 100% Link Utilization (Assuming FER Target of 0.1).

3.3 Performance Evaluation of TCP Over DCH

The present section evaluates the performance of TCP connections over Dedicated Channels based on a MATLAB™ simulator whose description and default parameters are given in Table 3.6. The main indicator of the TCP connection performance will be the RLC throughput. See its definition in [PI 3.1].

Parameter		Description
Physical Layer	Transport Channel Type	DCH
	TTI	10 ms
	Channel Bit Rate	Constant = [64 128 256] kbps
	Closed Loop Power Control	Ideal (Block erasures independent)
	Outer Loop Power Control	BLER Target = 10%
RLC Layer	Operation Mode	Acknowledged
	PDU Delivery	In-Sequence
	PDU Size	320 bits
	RLC_Tx Window Size	1024 PDUs
	RLC_Rx Window Size	1024 PDUs
	RNC Buffering Capabilities	Unlimited
	SDU Discard Mode	After MaxDAT
	MaxDAT[6] (definition in 3.3.5)	10
	Polling Mechanism	RLC status report every PDU
	PDU Retransmission Delay	90 ms
PDCP Layer	TCP/IP Header Compression	Header Fully Compressed
TCP Layer	Version	Reno
	MSS	536 B
	awnd	32000 B
	Initial cwnd	1 MSS
	Initial ssthresh	awnd
	Clock Tick	10 ms
	ACK Generation	Maximum 1 per clock tick
	Initial RTO	3 sec
	Maximum RTO	60 sec
	Minimum RTO	1 sec
	DUPACKS for Fast Retransmit	3
	ACKs Loss Probability	0 %
	Internet + CN + UE Processing + Uplink Delay	Constant = [140 300] ms
Application Layer	FTP	100 kB

Table 3.6: Default Parameters for TCP Simulations.

[6] It is relevant to note that the RLC parameter MaxDAT indicates the maximum number of times an RLC PDU can be retransmitted minus two. See section 3.3.5.

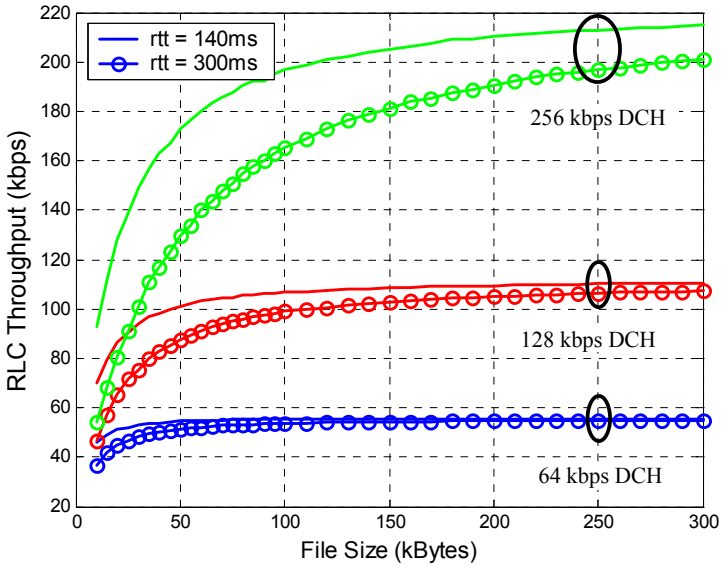


Figure 3.2: RLC Throughput as Function of the File Size.

3.3.1 Influence of the Slow Start

From an efficiency point of view, the impact of the initial slow start after the TCP connection establishment depends on the overall amount of downloaded data. For small files (like WWW pages) the document transmission is affected by the poor radio link utilization during the initial slow start, with the subsequent throughput degradation. On the other hand, for larger files (e.g. ftp downloads) the effect is expected to be minor.

Figure 3.2 shows the RLC throughput as a function of the file size in a dedicated channel (DCH) for three different constant channel bit rates 64kbps, 128 kbps and 256 kbps. The results assume an error free connection in the wired network, and the MaxDAT parameter of the RLC layer equals 10, which ensures a negligible RLC SDU error ratio. The results further assume an initial congestion window equal to one (see in section 2.7 the discussion about the influence of the reading times on the TCP connection set-up and close).

As shown in the graphic, small web page sizes significantly degrade the RLC throughput of users allocated high channel bit rates. For example, a user that is allocated a 256 kbps channel with a round trip time of 140 ms experiences a 25% RLC throughput degradation with a 50 kbytes file, whereas a 150 kbytes file reduces the degradation down to 11%. As expected from section 3.2.4, larger round trip times increase the BDP, and also influence negatively the user's performance. For example, for the previously mentioned 50 kbytes file example, a round trip time of 300 ms increases the throughput degradation to 44%. The throughput of users allocated low channel bit rates is less sensible to the page size or the round trip time, but on the other hand these users perceive worse QoS due to larger page transmission delays. For example, a user that is allocated a 64 kbps channel and downloads a 50 kbytes file suffers an RLC throughput degradation of only 10%, but only experiences a user throughput of 51 kbps.

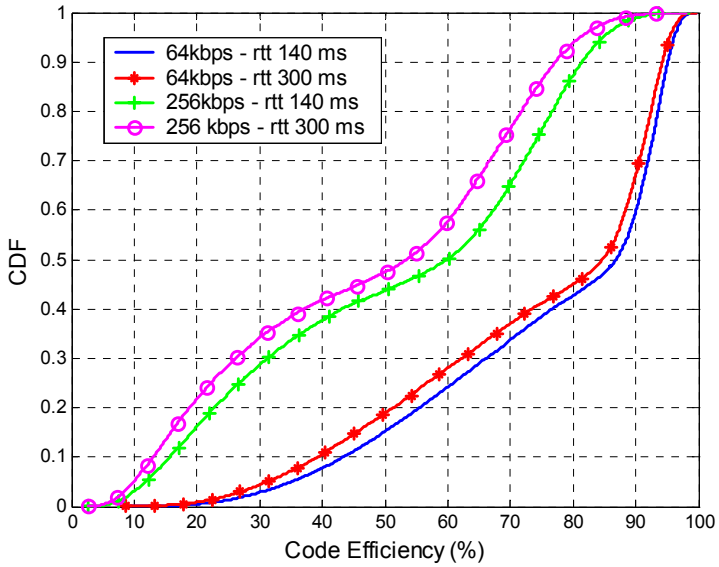


Figure 3.3: Channelization Code Eficiency of WWW applications for Various Channel Bit Rates and DCHs. The Packet Call Size Distribution from [24]. DCH Release Timer of 2 sec.

3.3.2 Code Utilization

In WCDMA, the RRM allocates the channel bit rates to the users according to the availability of the two main resources in the cell: the total Node B transmission power and the channelization codes. High bit rates DCHs consume a significant part of the code tree. The strategy of allocating high bit rate DCHs to traffic flows with low activity may produce an imbalance between average Node B transmission power and code consumption, possibly leading to a code shortage situation in which the capacity is limited due to the lack of codes.

Figure 3.3 depicts the cdf of the code efficiency of the DCH for web browsing applications. The code efficiency was computed as the time the link was being utilized for transmission divided by the DCH allocation period. The channel is allocated when the first data segment starts to be transmitted, and it is released when the last data bit is successfully transmitted and the DCH release timer goes off. The release timer was set to 2 seconds. The packet call size model is taken from [24], and as described in section 2.7, its distribution is heavy tailed. The resulting average packet call size equals around 140 kbytes, which is considerably large due to the probability mass in the distribution tail. The code inactivity caused by effects of the HTTP layer, such as the parsing time, is disregarded. It is further assumed that the initial congestion window is equal to one. Note from Chapter 2 that the same TCP connection may be reused for consecutives downloads in the reading time is sufficiently low.

As shown by the figure, the code efficiency is highly dependent on the allocated channel bit rate. Around 50% of the web page downloads in 256kbps DCH have a code efficiency below 50-60%. In the case of the 64kbps DCH, 50% of the web page downloads achieve a code efficiency higher than around 85%. It is interesting to observe that the round trip time has a minor influence on the resulting distribution of the code utilization. The strong dependence of the code utilization on the channel bit rate suggests a radio resource allocation strategy that could adapt the bit rate (step wise) to the offered load of the application [61].

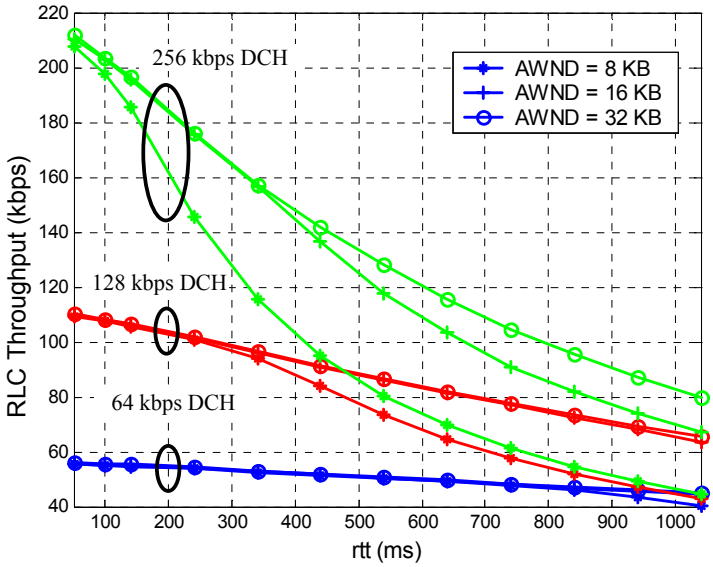


Figure 3.4: RLC Throughput as Function of the Round Trip Time (rtt) For Various Receiver's Buffering Capabilities and Bit Rates.

3.3.3 Influence of the TCP Receiver's Buffering Capabilities

As mentioned in section 3.2, the TCP receiver's buffer size imposes an upper bound on the achievable TCP connection throughput. If the window advertised by the receiver limits the TCP transmitting window to a value below the bandwidth delay product, the total amount of outstanding data will not be enough to reach the bottleneck capacity and the resources allocated to the TCP connection will not be fully utilized.

Figure 3.4 shows the RLC throughput as a function of the round trip time (rtt) in a dedicated channel (DCH) for a 100 kbytes file, for three different constant channel bit rates 64kbps, 128 kbps and 256 kbps, and TCP receiver's windows (awnd). The results again assume an error free connection in the wired network, and the MaxDAT parameter of the RLC layer equal to 10.

The results show that the effect of the TCP receiver's window completely varies depending on both the available bandwidth and the round trip time. It can be seen that there exists no throughput performance difference between the tested awnds for 64 kbps dedicated channels and for realistic round trip times. The 8 kbytes receiver's window only limits the throughput for 128 kbps channels for round trip times larger than 300-400 ms. Larger windows are definitely required for higher bit rates.

It is interesting to observe in Figure 3.4 that there exists no abrupt operation point that determines the throughput limitation, but the transition is rather smooth. Moreover, the throughput limitation starts for round trip times slightly lower than predicted by the BDP. For example, Table 3.4 describes that a 128 kbps DCH with a round trip time of 450 ms has a 6.85 kbytes, and therefore an 8 kbytes advertised window should not limit the throughput. However, Figure 3.4 already shows a small throughput degradation for that round trip time. The reason behind this effect is the variability of the overall round trip time, which can rise

over its average and cause some small transmission gaps if the advertised window is exactly equal to the BDP.

It is clear from Table 3.4 and Figure 3.4 that the larger the TCP receiver's window, the wider it is the range of bit rates and round trip times can enjoy a full radio link utilization. However, having very large receiver's window has a relevant disadvantage besides the cost of the memory in the mobile terminal. The TCP transmitter does not have information of the actual BDP, so the congestion will grow over the BDP until there is an error or it reaches the awnd. Hence, as expressed in equation (3.6), the data excess over the BDP will be cumulated in the buffer of the RLC. With a relative segment loss free connection, if the round trip time is low and the allocated channel bit rate is low too, the amount of data to be queued may cause the RLC buffer to overflow. Note that the RNC buffering capabilities must be sufficient to store all the information of all the DCH/DSCH connections in all the Node Bs served by the RNC. The RLC buffer overflow is a very undesired situation because it will drop multiple TCP segments that will trigger a retransmission timer expiration with its subsequent negative effects on the congestion window and ultimately in the user throughput.

3.3.4 Influence of Random Segment Losses

Previous sections have assumed an error free TCP connection. Under those conditions, it has been described that the main parameters that determine the TCP behaviour are the bandwidth delay product and the evolution of the congestion window during the slow start. However, the appearance of segment losses, either single random drops or multiple packet losses caused by network congestion, triggers the flow control mechanisms that completely modify the overall connection performance. As discussed in [62], the different TCP phases are directly reflected on the RLC buffer occupancy, and thus on the link utilization. This section will analyse the effect of the random segment losses, assuming that these losses occur with a certain probability and are fully independent of each other.

Let briefly describe the events that take place when a single segment loss occurs. Figure 3.5 depicts an example of the RLC buffer size evolution during a file download in such a situation. When the packet loss occurs, the receiver sends a duplicate acknowledgement for every out of sequence data segment, which will trigger the fast retransmit algorithm at the server. After the lost segment is retransmitted, the fast recovery algorithm may send a few segments if permitted by the congestion window, and when the new acknowledgement arrives a large amount of new data is usually delivered. Later, cwnd evolves linearly as dictated by the congestion avoidance algorithm. If the congestion window is large enough when the random loss takes place, the retransmission and the process of halving the window may fortunately represent a minor source of degradation.

The frequency of the random losses and the actual value of the congestion window when the errors occur are key factors in the resulting performance. If the fast retransmission and fast recovery algorithms are triggered too frequently, the congestion window is halved too many times, and the congestion avoidance will not have cycles enough to raise the cwnd over the BDP between consecutive losses. Moreover, if more than one packet (burst of packets) are lost within the same congestion window, it will be difficult to carry out all the retransmissions and receive their corresponding acknowledgements before the RTO timer expires. This performance problem of TCP Reno when multiple packets are dropped from a window of data is well known [63].

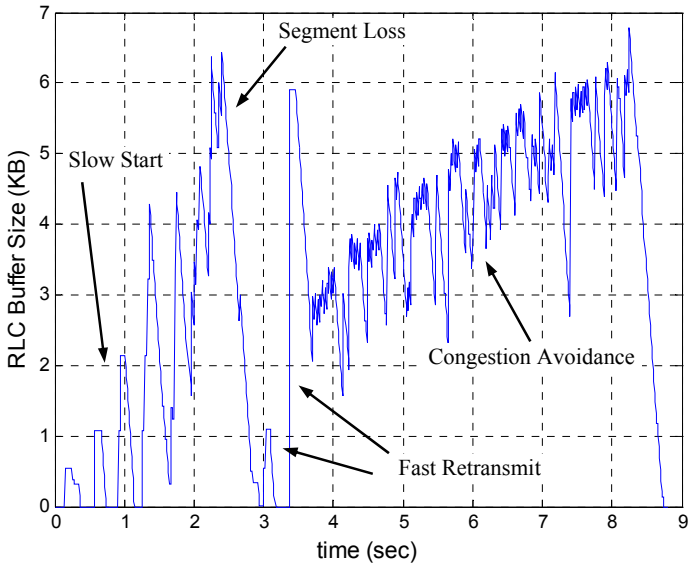


Figure 3.5: Realization of the Evolution of the RLC Buffer Size During File Download With a Single Random Segment Loss on a 128 kbps Channel.

Furthermore, during the initial slow start, if a segment loss occurs when the congestion window is grown-up but the RTO has still not adapted enough to the window growth (see left most part of equation (3.7)), then a time out is likely to happen. Let briefly explain the triggering of the time out. When the segment loss takes place, the fast retransmission algorithm will carry out the retransmission of the missing segment. The retransmitted segment may experience an overall round trip time larger than the RTO (due to the queuing at bottleneck link) and trigger the time out. Unluckily, if the random loss probability is high enough, this event is not so rare. This belongs to the so-called spurious time outs, i.e. a time out that could have been avoided if the sender had waited "long enough".

The spurious timer expirations are very unfortunate because the sender must reduce the ssthresh (i.e. the load) and it is forced to execute slow start to retransmit all the segments that had already been transmitted but not acknowledged (which might not even require the retransmission). Besides, the double transmission of the packets can trigger duplicated acknowledgements, which may activate the fast retransmission. The ultimate result is that the congestion window is completely exhausted. This type of time outs has been dealt with in [64].

There exist some other cases when a segment loss can cause a time out, for example when the congestion window is smaller than five. In such a case, the number of outstanding segments can not exceed four, and hence, the loss of any of them would imply that there would be not enough duplicate acknowledgements to trigger a fast retransmission. Nonetheless, these subtle cases are omitted here because they are much less statistically relevant.

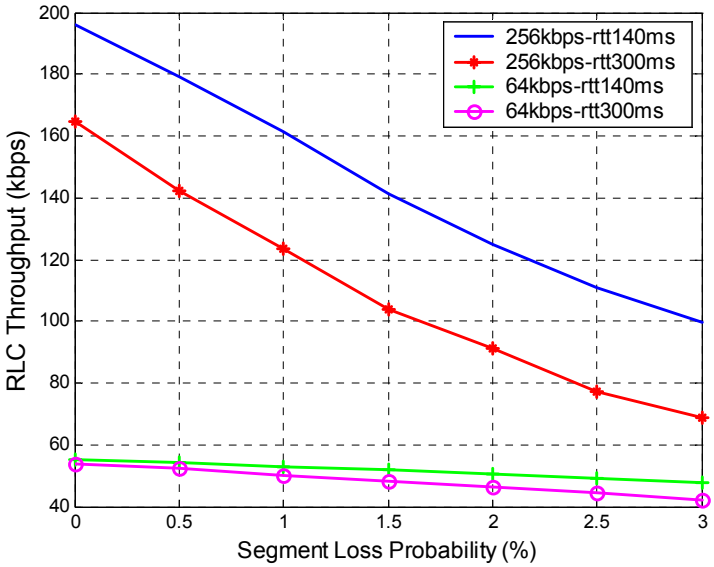


Figure 3.6: RLC Throughput as Function of the Random Segment Loss Probability for Various BDPs.

Figure 3.6 depicts the RLC layer throughput as a function of the random loss probability for various channel bit rates and round trip times. The results show a significant performance difference between low and high bit rate channels. Firstly, the segment transmission gap caused by either fast retransmissions or time outs causes larger deterioration for higher bit rates. Secondly, as commented above, high bit rates require larger BDPs to fully utilize the radio link. In [65], Lakshman illustrated that the effect of the random losses degrades significantly the TCP throughput if the loss probability exceeds a threshold that equals the inverse of the square of the BDP (expressed in segments). For example, for a 64 kbps channel with a 300 ms rtt, the probability threshold is equal to 4%, whereas for the same round trip time and for a 256 kbps channel the threshold equals 0.34%, which approximates pretty well to the results of Figure 3.6. Table 3.7 gives the average number of time outs and the average number of times the fast retransmit and fast recovery algorithms are triggered as a function of the segment loss probability (SLP) for the cases simulated for Figure 3.6.

Rb	**64 kbps**		**256 kbps**	
rtt	**140 ms**	**300 ms**	**140 ms**	**300 ms**
SLP = 0.5 %	0.44/0.94	0.43/0.94	0.23/0.77	0.21/0.83
SLP = 1 %	0.74/1.76	0.75/1.67	0.42/1.44	0.41/1.48
SLP = 1.5 %	0.98/2.38	1.00/2.35	0.72/2.05	0.64/2.19
SLP = 2 %	1.22/2.94	1.26/3.00	1.03/2.62	0.93/2.67
SLP = 2.5 %	1.46/3.55	1.49/3.60	1.42/3.09	1.23/3.36
SLP = 3 %	1.77/4.08	1.77/4.16	1.81/3.59	1.61/3.78

Table 3.7: Average Number of Time Outs/ Average Number of Times Fast Retransmit is Triggered as function of Segment Loss Probability (SLP).

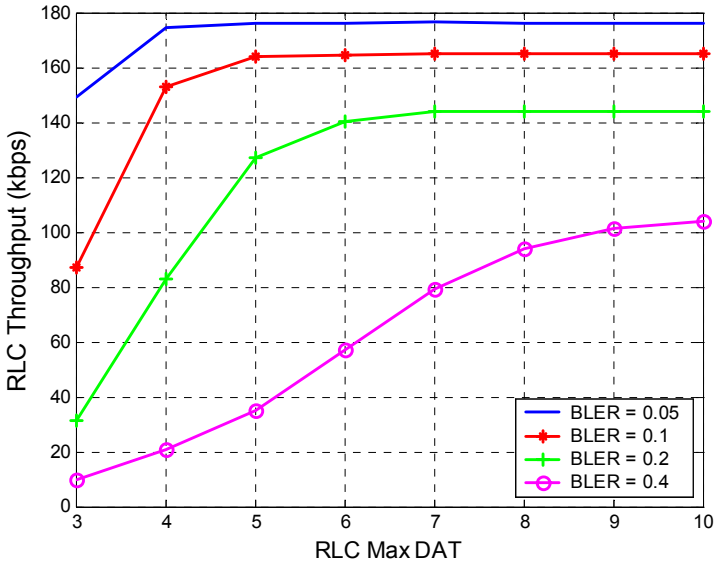


Figure 3.7: RLC Throughput as Function of the RLC MaxDAT for a 256 kbps Channel and 300 ms Round Trip Time (rtt).

3.3.5 Influence of RLC MaxDAT

UMTS introduces the RLC layer to recover the block erasures caused by the link level errors of the radio channel. The acknowledged mode of the RLC layer aims at providing an error free data transmission service by making use of a selective repeat ARQ mechanism that retransmits the erroneously decoded PDUs. With a sufficiently high number of RLC retransmissions, the SDU error ratio provided by the RLC layer can be negligible. However, this reliability benefit is at expenses of a variable link level transmission delay, which can potentially lead to TCP time outs if the retransmission timer can not cope with the round trip time peaks caused by a large number of RLC retransmissions.

The RLC specification [42] enables several modes to discard an RLC SDU, though the one considered in the present study is the discarding after reaching MaxDAT number of retransmissions. In this case, the network keeps a counter (called VT(DAT)) that is increased by one whenever an RLC PDU is fetched for transmission. The RLC layer starts the SDU discard procedure (or resets its operation) when the counter reaches MaxDAT [42]. More specifically, as the counter VT(DAT) is initialised to zero [42] and it is typically updated before forwarding the PDUs to the radio channel, it must be concluded that the RLC parameter MaxDAT indicates the maximum number of RLC retransmissions minus two.

Figure 3.7 plots the RLC layer throughput as function of the RLC parameter MaxDAT for a 256 kbps DCH and a 300 ms round trip time (rtt). The results clearly show that the more reliable the RLC, the better performance can be experienced by the end user. In order to reach a high degree of reliability (marginal SDU error ratio), different MaxDAT are required for the different BLER targets. So, while a BLER target of 5% reaches the maximum performance with MaxDAT equal to 5, a BLER target of 40% requires a MaxDAT of 9-10 for minimum throughput degradation.

However, the most relevant conclusion from Figure 3.7 is that the link level error correction mechanisms are absolutely preferable than the TCP end-to-end mechanisms, because the latter ones halve the congestion window whenever triggered. Even for a large number of retransmissions, the round up of the RTO to one second and the reset of the TCP retransmission timer whenever a new acknowledgement is received[60], contribute to cope with the round trip time spikes. However, had the block not been retransmitted by the RLC layer enough times, it would have triggered either a fast retransmit and fast recovery or a time out at TCP layer anyway. Simulations with RLC retransmissions delays of 180 ms yield very analogous results as depicted by Figure 3.7. Furthermore, results for a 64 kbps channel also showed very similar trend.

Table 3.8 gives the average number of time outs and average number of times the fast retransmit and fast recovery algorithms are triggered. The results indicate that for relatively low RLC SDU error ratios, the main mechanism to deal with the retransmissions is the fast retransmissions.

Initially, it could be expected that the fast retransmit and fast recovery algorithms are sufficient to cope with the link level errors not corrected by the RLC layer. However, there exist situations when more than a single segment is lost within the same congestion window. These situations trigger an initial fast retransmit, and possibly a time out too because the second retransmission is not successfully acknowledged before the timer expiration. For example, if the erased block comprises PDUs of two distinct SDUs, the RLC layer discards both SDUs (i.e. two TCP segments). Similarly, when the SDU error ratio increases significantly due to a very low value of MaxDAT (3-4), the probability of having more than one segment lost within the same congestion window raises. In these situations, the slow start threshold is halved twice, and when the retransmission episode finishes, the congestion window evolves from the halved threshold linearly (congestion avoidance) to the detriment of the throughput performance. Furthermore, with an exhaust congestion window lower than five segments, any new segment loss can only be recovered with a new time out.

	BLER Target			
	5 %	**10 %**	**20 %**	**40 %**
MaxDAT = 10	0.0/0.0	0.0/0.0	0.0/0.0	0.06/0.14
MaxDAT = 9	0.0/0.0	0.0/0.0	0.001/0.004	0.13/0.33
MaxDAT = 8	0.0/0.0	0.0/0.0	0.002/0.004	0.27/0.73
MaxDAT = 7	0.0/0.0	0.0/0.001	0.007/0.03	0.63/1.57
MaxDAT = 6	0.0/0.001	0.0/0.002	0.06/0.16	1.58/3.16
MaxDAT = 5	0.001/0.001	0.001/0.04	0.23/0.64	4.69/5.61
MaxDAT = 4	0.017/0.056	0.11/0.39	1.04/2.51	14.19/6.92
MaxDAT = 3	0.26/0.89	1.37/2.88	7.75/6.51	-

Table 3.8: Average Number of Time Outs/ Average Number of Times Fast Retransmit is Triggered.

3.4 Conclusions

The present chapter has analysed the performance of the TCP protocol over WCDMA/DCH. On a general perspective, it can be concluded that there exists a significant performance difference between low and high bit rate channels for realistic round trip times (115-450 ms). The higher bit rates allocated by the Packet Scheduler to the user, the bigger is the potential risk of the congestion window not reaching the bandwidth delay product. This risk can lead to the user throughput being degraded below the allocated channel bit rate, as well as to low traffic activity factors. The low traffic activity is also an undesired effect because it enhances the variance of the overall Node B transmission power (i.e. risk of network instability), and it can even lead to channelization code shortages especially in microcell environments.

For small file sizes, such as web pages, the initial slow start can harm the user throughput of users allocated high channel bit rates. As described above, a user that is allocated a 256 kbps channel with a round trip time (rtt) of 140ms, experiences 25% RLC throughput degradation with a 50 kbytes file, whereas a 150 kbytes file reduces the degradation down to 11%. Users assigned lower bit rates channels, such as a 64 kbps experience a higher file download delay, but hardly suffer any degradation from the initial slow start for round trip times (rtt) of 150 or 300 ms. Accordingly, terminals of users allocated bit rates lower or equal to 256 kbps should be equipped with at least 16 kB of TCP receiver's buffer for realistic round trip times. For higher bit rates, such as 384 kbps, they require at least 32 kB. Very large receiver's buffer sizes (e.g. 64 kB) have the disadvantage that if users are allocated low bit rate channels (e.g. 32-64 kbps), then a large amount data will be cumulated in the RLC buffer, which may cause the RNC buffer to overflow due to the large number of users served by a single RNC.

The introduction of segment losses significantly changes the behaviour of the protocol. Again, their impact on the performance is completely dependent on the bit rate allocated to the user. Low bit rate channels, such as 64 kbps, hardly suffer any throughput degradation for 1% loss probability, whereas for a 256 kbps channel and a round trip time of 300 ms the same segment loss rate already degrades the throughput by 25%.

Regarding the interaction between the RLC and the TCP layer, it has been shown that the more reliable is the RLC layer, the more benefits are obtained in terms of user throughput because the end-to-end TCP recovery mechanisms have negative long-term effects on the congestion window. This conclusion is obtained even considering RLC retransmission delays up to 180 ms, because the round up of the RTO to one second and the reset of the TCP retransmission timer whenever a new acknowledgement is received, offer considerable protection against spurious time outs caused by a large number of RLC retransmissions.

Chapter 4

WCDMA Evolution: High Speed Downlink Packet Access

4.1 Introduction

As commented in Chapter 1, the evolution of the mobile communication market brings demands for both larger system capacity and higher data rates. To boost the support for the packet switched services, the 3GPP has standardized in the Release 5 a new technology denominated High Speed Downlink Packet Access (HSDPA) that represents an evolution of the WCDMA radio interface. HSDPA appears as an umbrella of features whose combination improves the network capacity, and increases the peak data rates up to theoretically 10 Mbps for downlink packet traffic. These technological enhancements can allow operators to enable new high data rate services, improve the QoS of already existing services, and achieve a lower cost per delivered data bit.

The HSDPA concept relies on a new transport channel, the High Speed Downlink Shared Channel (HS-DSCH), where a large amount of power and code resources are assigned to a single user at a certain TTI in a time and/or code multiplex manner. The time-shared nature of the HS-DSCH provides significant trunking benefits over DCH for bursty high data rate traffic [12]. In addition, HSDPA uses Adaptive Modulation and Coding, fast Physical Layer

Hybrid ARQ, and fast Packet Scheduling. These features are tightly coupled and permit a per-TTI adaptation of the transmission parameters to the instantaneous variations of the radio channel quality.

The objective of the present chapter is not to give a complete description of the HSDPA concept and its performance. It will rather provide a general overview of HSDPA that will be required to achieve a full comprehension for the analysis/design of the Packet Scheduler functionality.

The Packet Scheduler is a key element of the overall HSDPA concept and the central scope of the following chapters and will, therefore, be treated in detail there. This way, this chapter will not cover specific HSDPA aspects that are not subject of analysis in the rest of this Ph.D. thesis like for example the UE capabilities. In the rest of the thesis it will be assumed that the penetration of HSDPA capable terminals in the network is 100%. Other overviews of the HSDPA concept are given in [66], [11], [9].

Section 4.2 gives a general description of the specific features included and excluded in HSDPA compared to basic WCDMA technology. Section 4.3 gives an overview of the major architectural modifications introduced by HSDPA, while section 4.4 reviews the HSDPA channel structure. Section 4.5 describes the AMC functionality, while section 4.6 deals with one of the most important features of HSDPA: the link adaptation functionality. Section 4.7 highlights the powerful HS-DSCH retransmission mechanism: Hybrid ARQ. Finally section 4.8 draws the main concluding remarks. Note that the Packet Scheduler is not covered in this chapter but in the Chapter 5 due to the special attention paid to it.

The 3GPP Technical Specification 25.308 [67], that provides the overall description of HSDPA, is the main source for the following sections unless otherwise explicitly stated.

4.2 General HSDPA Concept Description

Figure 4.1 plots the fundamental features to be included and excluded in the HS-DSCH of HSDPA. For this new transport channel, two of the most main features of the WCDMA technology such as closed loop power control and variable spreading factor have been deactivated.

In WCDMA, fast power control stabilizes the received signal quality (EsNo) by increasing the transmission power during the fades of the received signal level. See definition of EsNo in [PI 4.1]. This causes peaks in the transmission power and subsequent power rise, which reduces the total network capacity. However, delay tolerant traffic may be served only under favourable radio channel conditions, avoiding the transmission during the inefficient signal fading periods. Moreover, the operation of power control imposes the need of certain headroom in the total Node B transmission power to accommodate its variations. The elimination of power control avoids the aforementioned power rise as well as the cell transmission power headroom. But due to the exclusion of power control, HSDPA requires other link adaptation mechanisms to adapt the transmitted signal parameters to the continuously varying channel conditions.

One of these techniques is denominated Adaptive Modulation and Coding (AMC). With it, the modulation and the coding rate are adapted to the instantaneous channel quality instead of

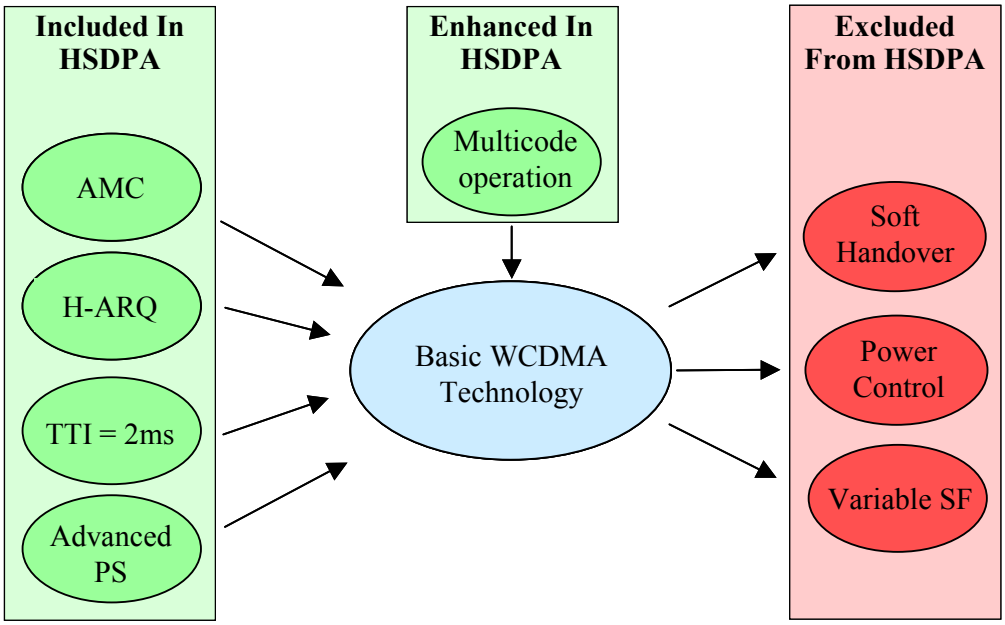


Figure 4.1: Fundamental Features To Be Included And Excluded in the HS-DSCH of HSDPA.

adjusting the power. The transmission of multiple Walsh codes is also used in the link adaptation process. Since the combination of these two mechanisms already plays the link adaptation role in HSDPA, the variable spreading factor is deactivated because its long-term adjustment to the average propagation conditions is not required anymore.

As closed power control is not present, the channel quality variations must be minimized across the TTI, which it is accomplished by reducing its duration from the minimum 10 ms in WCDMA down to 2 ms. The fast Hybrid ARQ technique is added, which rapidly retransmits the missing transport blocks and combines the soft information from the original transmission with any subsequent retransmission before the decoding process. The network may include additional redundant information that is incrementally transmitted in subsequent retransmissions (i.e. Incremental Redundancy).

To obtain recent channel quality information that permits the link adaptation and the Packet Scheduling entities to track the user's instantaneous radio conditions, the MAC functionality in charge of the HS-DSCH channel is moved from the RNC to the Node B. The fast channel quality information allows the Packet Scheduler to serve the user only when his conditions are favourable. This fast Packet Scheduling and the time-shared nature of the HS-DSCH enable a form of multiuser selection diversity with important benefits for the cell throughput. The move of the scheduler to the Node B is a major architecture modification compared to the Release 99 architecture.

4.3 HSDPA Architecture

Unlike all the transport channels belonging to the Release 99 architecture, which are terminated at the RNC, the HS-DSCH is directly terminated at the Node B. With the purpose of controlling this channel, the MAC layer controlling the resources of this channel (so called

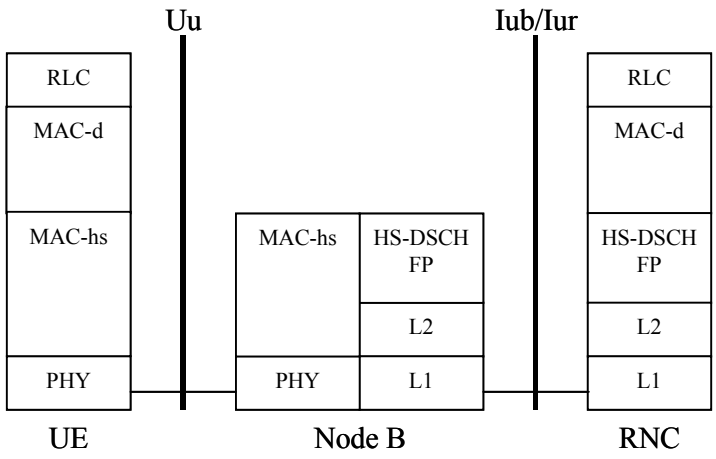


Figure 4.2: Radio Interface Protocol Architecture of the HS-DSCH Transport Channel (Configuration without MAC c-sh) from [67].

MAC-hs) is directly located in the Node B (see Figure 4.2), thereby allowing the acquisition of recent channel quality reports that enable the tracking of the instantaneous signal quality for low speed mobiles. This location of the MAC-hs in the Node B also enables to execute the HARQ protocol from the physical layer, which permits faster retransmissions.

More specifically, the MAC-hs layer [43] is in charge of handling the HARQ functionality of every HSDPA user, distributing the HS-DSCH resources between all the MAC-d flows according to their priority (i.e. Packet Scheduling), and selecting the appropriate transport format for every TTI (i.e. link adaptation). The radio interface layers above the MAC are not modified from the Release 99 architecture because HSDPA is intended for transport of logical channels. Nonetheless, the RLC can only operate in either acknowledged or unacknowledged mode, but not in transparent mode due to ciphering [26]. This is because for the transparent mode the ciphering is done in the MAC-d[7], not in the RLC layer, and MAC-c/sh and MAC-hs do not support ciphering [43].

The MAC-hs also stores the user data to be transmitted across the air interface, which imposes some constraints on the minimum buffering capabilities of the Node B. Chapter 7 will study the effect of limited Node B memory on the performance of streaming bearers on HSDPA. The move of the data queues to the Node B creates the need of a flow control mechanism (HS-DSCH Frame Protocol) that aims at keeping the buffers full. The HS-DSCH FP handles the data transport from the serving RNC to the controlling RNC (if the Iur interface is involved) and between the controlling RNC and the Node B. The design of such flow control is a non-trivial task, because this functionality in cooperation with the Packet Scheduler is to ultimately regulate the user's perceived service, which must fulfil the QoS attributes according to the user's subscription (e.g. the guaranteed bit rate or the transfer delay for streaming bearers or the traffic handling priority and the allocation/retention priority for interactive users).

In addition, the HS-DSCH does not support soft handover due to the complexity of synchronizing the transmission from various cells. The HS-DSCH may optionally provide full or partial coverage in the cell, though the rest of this thesis will only consider full HSDPA cell

[7] Note that the ciphering function is only performed in MAC-d for the DCH channel [43].

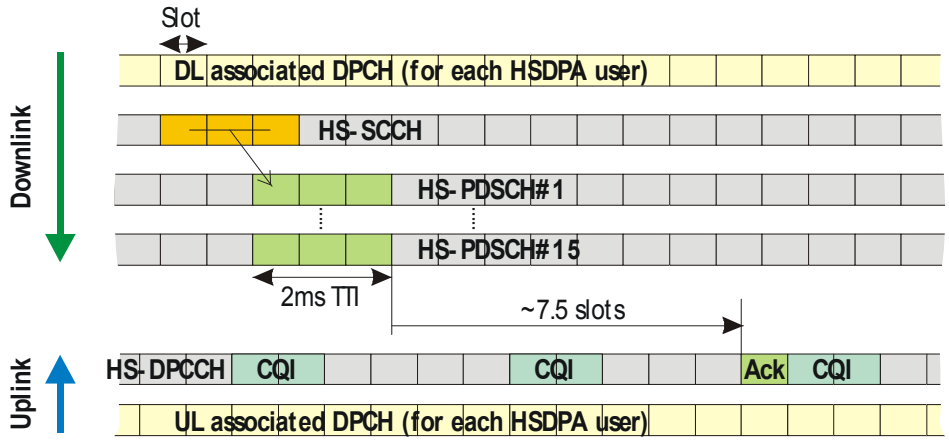


Figure 4.3: Uplink and Downlink Physical Layer Structure in HSDPA.

coverage. Note, however, that the associated DPCH (see its description in section 4.4) can still operate in soft handover mode.

4.4 HSDPA Channel Structure

There exists in WCDMA a transport channel particularly suited for downlink packet bursty traffic: the Downlink Shared Channel (DSCH). The DSCH provides common code resources that can be shared by several users in a time multiplex manner. It has the potential to improve the capacity for bursty packet traffic, since the sharing of the resources reduces the potential channelization code shortage that might occur if every user was allocated a DCH.

As mentioned above, the HSDPA concept relies on a new transport channel, the HS-DSCH, which can be seen as an evolution of the DSCH channel. The HS-DSCH is mapped onto a pool of physical channels (i.e. channelization codes) denominated HS-PDSCHs (High Speed Physical Downlink Shared Channel) to be shared among all the HSDPA users on a time multiplexed manner. The spreading factor of the HS-PDSCHs is fixed to 16, and the MAC-hs can use one or several codes, up to a maximum of 15. Moreover, the scheduler may apply code multiplexing by transmitting separate HS-PDSCHs to different users in the same TTI. However, code multiplexing will not be subject of investigation in this Ph.D. thesis. The uplink and downlink channel structure of HSDPA is described in Figure 4.3. The HSDPA concept includes a Shared Control CHannel (HS-SCCH) to signal the users when they are to be served as well as the necessary information for the decoding process. The HS-SCCH carries the following information [67]:

- *UE Id Mask*: to identify the user to be served in the next TTI.
- *Transport Format Related Information*: specifies the set of channelization codes, and the modulation. The actual coding rate is derived from the transport block size and other transport format parameters [10].
- *Hybrid ARQ Related Information*: such as if the next transmission is a new one or a retransmission and if it should be combined, the associated ARQ process, and information about the redundancy version.

This control information solely applies to the UE to be served in the next TTI, which permits this signalling channel to be a shared one.

The RNC can specify the recommended power of the HS-SCCH (offset relative to the pilots bits of the associated DPCH) [33]. The HS-SCCH transmit power may be constant (possibly inefficient) or time varying according to a certain power control strategy though the 3GPP specifications do not set any closed loop power control modes for the HS-SCCH.

An uplink High Speed Dedicated Physical Control Channel (HS-DPCCH) carries the necessary control information in the uplink, namely, the ARQ acknowledgements, and the Channel Quality Indicator (CQI) reports. The CQI reports are described in section 4.6.1. To aid the power control operation of the HS-DPCCH an associated Dedicated Physical Channel (DPCH) is run for every user.

According to [33], the RNC may set the maximum transmission power on all the codes of the HS-DSCH and HS-SCCH channels in the cell. Otherwise, the Node B may utilize all unused Node B transmission power for these two channels, though this option will not be considered in the following chapters. Likewise, the RNC determines the maximum number of channelization codes to be used by the HS-DSCH channel.

4.5 AMC and Multi-code Transmission

As mentioned in section 4.2, HSDPA utilizes other link adaptation techniques to substitute power control and variable spreading factor. To cope with the dynamic range of the EsNo at the UE, HSDPA adapts the modulation, the coding rate and number of channelization codes to the instantaneous radio conditions. The combination of the first two mechanisms is denominated Adaptive Modulation and Coding (AMC).

Besides QPSK, HSDPA incorporates the 16QAM modulation to increase the peak data rates for users served under favourable radio conditions. Support for QPSK is mandatory for the mobile, though the support of 16QAM is optional for the network and the UE [67]. The inclusion of this high order modulation introduces some complexity challenges for the receiver terminal, which needs to estimate the relative amplitude of the received symbols, whereas it only requires the detection of the signal phase in the QPSK case. A turbo encoder is in charge of the data protection. The encoder is based on the release 99 turbo encoder with a rate of 1/3, though other effective coding rates within the range (approximately from 1/6 to 1/1 [11]) can be achieved by means of rate matching, i.e. puncturing and repetition. The resulting coding rate resolution has 64 steps geometrically distributed. The combination of a modulation and a coding rate will be denominated here as *Modulation and Coding Scheme*. Table 4.1 shows the set example of Modulation and Coding Schemes (MCS) to be used in the following chapters.

Besides AMC, multi-code transmission can also be considered as a tool for link adaptation purposes. If the user enjoys good channel conditions, the Node B can exploit the situation by transmitting multiple parallel codes, reaching significant peak throughputs. For example, with the MCS 5 and a set of 15 multi-codes, a maximum peak data rate of 10.8 Mbps can be obtained. Such high peak data rates are expected to be used under favourable instantaneous signal quality conditions.

MCS	Modulation	Eff. Coding Rate	Bits per TTI	Peak Rate (1 Code)
1	QPSK	¼	240	120 kbps
2		½	480	240 kbps
3		¾	720	360 kbps
4	16 QAM	½	960	480 kbps
5		¾	1440	720 kbps

Table 4.1: MCS Set Example for HSDPA and Available Peak Bit Rate per Code.

With the multi-code transmission, the overall dynamic range of the AMC can be increased by $10 \cdot \log_{10}(15) \cong 12$ dBs. The overall link adaptation dynamic range coped with the combination of the AMC and multi-code transmission is around 30 dB. Note that the dynamic range of the variable spreading factor of WCDMA is around 20 dB, which yields a dynamic range around 10 dB larger in HSDPA. Nonetheless, the dynamic range of HSDPA is slightly shifted upwards, which might indeed have some coverage implications. The most protecting coding rate (1/6) may not be sufficient to serve a user close to the cell edge during a signal level fade with a reasonable BLER.

The HSDPA dynamic range to be obtained with the MCSs of Table 4.1 can be seen in the Figure 4.5, and further in the Chapter 5 with the Figure 5.2 b).

4.6 Link Adaptation

The link adaptation functionality of the Node B is in charge of adapting the modulation, the coding format, and the number of multi-codes to the instantaneous radio conditions. In order to understand the principles that should rule this functionality, first the spectral efficiency of the different MCSs is analysed.

The bar plot of Figure 4.4 depicts the received EbNo per data bit (per channelization code of SF = 16) at BLER 10% from link level simulations for the MCSs described in Table 4.1 in Pedestrian A at 3 km/h. See definition of the received EbNo per data bit in [PI 4.2]. The figure adds the EbNo lower bound as a function of the peak data rate (PDR) for a channelization code of SF = 16. This lower bound has been computed according to the channel capacity of a band limited AWGN channel derived by Shannon [68]. Note that the theoretical link capacity represents the optimum performance limit. From the figure it can be concluded that the usage of the most protective transport formats (i.e. MCS 1 for the set described in Table 4.1) represents the lowest cost in terms of received EbNo per data bit.

To include the effect of users location, a geometry factor (G factor) is based on the ratio between the total instantaneous own cell power received by the UE, I_{or}, and the total instantaneous received interference from other cells plus noise perceived by the UE, I_{oc}. Then, the G factor is computed taking statistical averaging of the ratio I_{or}/I_{oc} over a local area. See [PI 4.3] for further indications about the G factor. Section B.7 gives the distribution of the G Factor in the cell obtained under the modelling of section B.2.

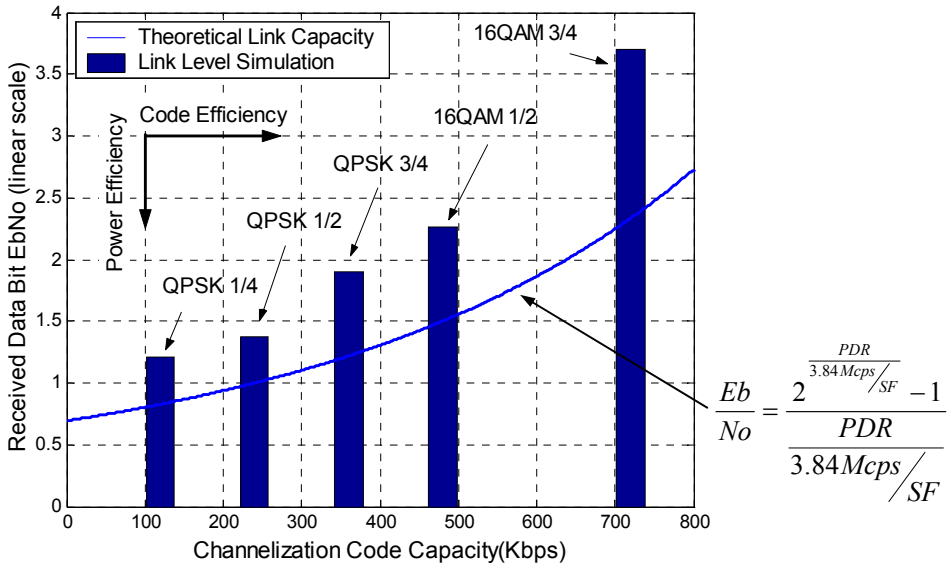


Figure 4.4: Received Data Bit Energy to Noise Spectral Density vs the Peak Data Rate (PDR) Per Code. The Figure includes the Shannon Capacity and Link Level Simulations Results at BLER = 10%, Pedestrian A, 3km/h.

For low G factor users, who are typically very interference limited with poorly received instantaneous signal quality, the allocation of the most robust MCS appears as the most power efficient solution. On the other hand, the usage of higher order MCSs is attractive in code shortage allocations where the served user can afford a higher cost in terms of bit energy.

Figure 4.5 plots the combination of the number of multi-codes and the MCS that provides highest first transmission throughput. The result assumes that the set of available MCSs is the one given in Table 4.1. With a very fine resolution of the coding rate, the most spectrally efficient allocation would only resort to higher order MCSs when all the available multi-codes are already used. Note however, in Figure 4.5 that higher order MCSs (i.e. MCS 2-5 of Table 4.1) could achieve a highest first transmission throughput with a number of multi-codes lower than the maximum available. This is due to the finite resolution of the coding rate in the MCS set of Table 4.1.

4.6.1 Methods for Link Adaptation

As described above, the link adaptation functionality must select the MCS and the number of multi-codes to adapt them to the instantaneous EsNo. However, the selection criterion can be based on various sources:

- *Channel Quality Indicator (CQI)*: the UE sends in the uplink a report denominated CQI that provides implicit information about the instantaneous signal quality received by the user. The CQI specifies the transport block size, number of codes and modulation from a set of reference ones that the UE is capable of supporting with a detection error no higher than 10% in the first transmission for a reference HS-PDSCH power. The RNC commands the UE to report the CQI with a certain periodicity from the set [2, 4, 8, 10, 20, 40, 80, 160] ms [69], and can possibly disable the report. The table including the set of reference CQI reports can be found in [70].

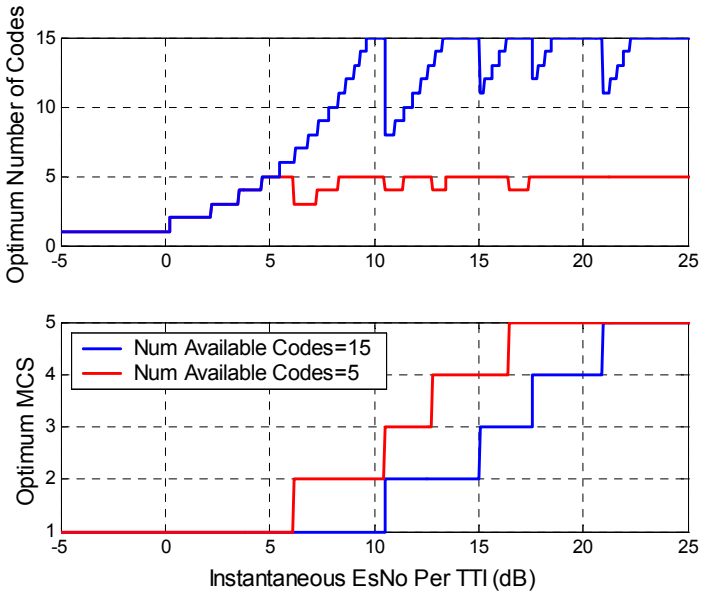


Figure 4.5: Optimum Number of Codes and MCS as a function of the Instantaneous EsNo Per TTI. PedA and 3km/h. Assumed Ideal Channel Quality Estimation.

- *Power Measurements on the Associated DPCH*: every user to be mapped on to HS-DSCH runs a parallel DPCH for signalling purposes, whose transmission power can be used to gain knowledge about the instantaneous[8] status of the user's channel quality. This information may be employed for link adaptation [71] as well as Packet Scheduling. With this solution, the Node B requires a table with the relative EbNo offset between the DPCH and the HS-DSCH for the different MCSs for a given BLER target. The advantages of utilizing this information are that no additional signalling is required, and that it is available on a slot basis. However, it is limited to the case when the HS-DSCH and the DPCH apply the same type of detector (e.g. a conventional Rake), and can not be used when the associated DPCH enters soft handover.

- *Hybrid ARQ Acknowledgements*: The acknowledgement corresponding to the H-ARQ protocol may provide an estimation of the user's channel quality too, although this information is expected to be less frequent than previous ones because it is only received when the user is served. Hence, it does not provide instantaneous channel quality information. Note that it also lacks the channel quality resolution provided by the two previous metrics since a single information bit is reported.

- *Buffer Size*: the amount of data in the MAC-hs buffer could also be applied in combination with previous information to select the transmission parameters.

For an optimum implementation of the link adaptation functionality, probably a combination of all the previous information sources would be need. If only one of them is to be selected, the CQI report possibly appears as the most attractive solution due to its simplicity for the network, its accuracy and its frequent report. However, during the analysis phase of this Ph.D. thesis, the exact definition of this report was not agreed, so it was decided to use the power

[8] Note that this may only be possible for low mobile speeds.

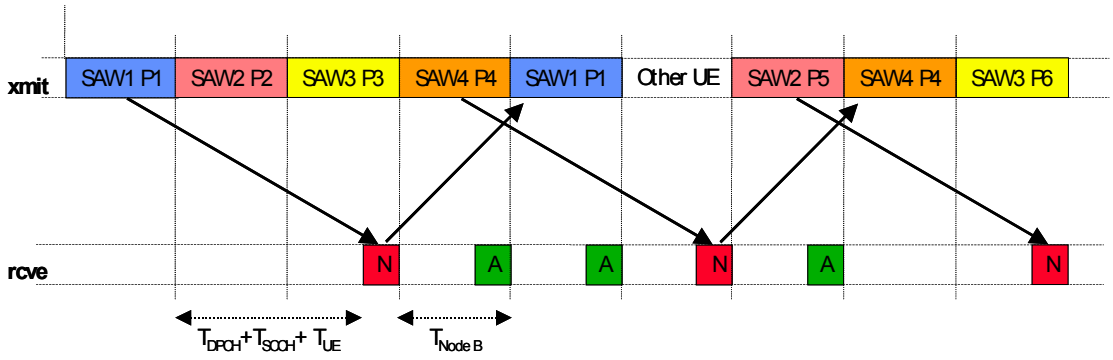


Figure 4.6: Four Channels SAW Protocol Operation. (Timing not in Scale).

measurements on the associated DPCH as the criterion for link adaptation purposes. This will be the underlying assumption in the following chapters.

4.7 Fast Hybrid ARQ

HSDPA incorporates a physical layer retransmission functionality that significantly improves the performance and adds robustness against link adaptation errors. Since the Hybrid ARQ functionality is located in the MAC-hs entity of the Node B, the transport block retransmission process is considerable faster than RLC layer retransmissions because the RNC or the Iub are not involved. This benefit is directly reflected on a lower UTRAN transfer delay (both in terms of average and standard deviation), which has obvious pay offs at end-to-end level (e.g. for TCP).

The retransmission protocol selected in HSDPA is the Stop And Wait (SAW) due to the simplicity of this form of ARQ. In SAW, the transmitter persists on the transmission of the current transport block until it has been successfully received before initiating the transmission of the next one. Since the continuous transmission to a certain UE should be possible [46], N SAW-ARQ processes may be set for the UE in parallel, so that different processes transmit in separate TTIs. The maximum number of processes for a single UE is 8. According to L1 round trip time estimates [72], the delay between transmission and first retransmission is around 12 ms, which requires 6 SAW processes for continuous transmission to a single UE. The SAW protocol is based on an asynchronous downlink and synchronous uplink [67]. That implies that in downlink, the HS-SCCH must specify the HARQ process that is transmitting on the HS-DSCH, while in uplink the SAW process acknowledgements are tied to the timing. Figure 4.6 shows an example of how the N channels SAW protocol transmits a sequence of packets P1…P6.

The Hybrid ARQ technique is fundamentally different from the WCDMA retransmissions because the UE decoder combines the soft information of multiple transmissions of a transport block at bit level. Note that this technique imposes some memory requirements on the mobile terminal, which must store the soft information of unsuccessfully decoded transmissions. There exist different Hybrid ARQ strategies:

- *Chase Combining (CC)*: proposed in [73], every retransmission is simply a replica of the coded word employed for the first transmission. The decoder at the receiver combines these multiple copies of the transmitted packet weighted by the received SNR prior to decoding. This type of combining provides time diversity and soft combining gain at a low complexity cost and imposes the least demanding UE memory requirements of all Hybrid ARQ strategies. The combination process incurs a minor combining loss as reported in [74] to be around 0.2-0.3 dB per retransmission.

- *Incremental Redundancy (IR)*: The retransmissions include additional redundant information that is incrementally transmitted if the decoding fails on the first attempt. That causes that the effective coding rate increases with the number of retransmissions. Incremental Redundancy can be further classified in Partial IR and Full IR. Partial IR includes the systematic bits in every coded word, which implies that every retransmission is self-decodable, whereas Full IR only includes parity bits, and therefore its retransmissions are not self-decodable. IR, and particularly Full IR, imposes demanding requirements on the UE memory capabilities, and the standard [67] only compels the UE soft memory to support the needs for Chase Combining. According to [74], Full IR only provides a significant coding gain for effective coding rates higher than 0.4-0.5, because for lower coding rates the additional coding rate is negligible since the coding scheme is based on a 1/3 coding structure. On the other hand, for higher effective coding rates the coding gain can be significant, for example a coding rate of 0.8 provides around 2 dB gain in Vehicular A, 3km/h, QPSK modulation.

Due to the simplicity of the Chase Combining method and its straightforward modelling, which only requires the addition of the signal qualities of the multiple retransmissions and the subtraction of the combining loss [74], the remaining chapters of this Ph.D. thesis will only consider this Hybrid ARQ scheme.

4.8 Summary

This chapter has given a general overview of the technologies comprised by the HSDPA concept. Though every technology by itself provides a significant network performance enhancement, it is their complementary characteristics that make HSDPA a key step in the evolution of WCDMA. The introduction of the AMC technique permits to exploit the high throughput available for users under favourable instantaneous signal quality conditions. The multi-code operation combined with the AMC extends the operating signal dynamic range up to around 30 dB. Note however, that this range is still smaller than the one given by the combination of power control and variable spreading factor. The most protective effective coding rate (around 1/6) already yields a data rate of 80 kbps, which might be not sufficiently robust to serve a user close to the cell edge during a signal fade with a reasonable BLER. Regarding the link adaptation operation, the multi-code allocation is more spectrally efficient than the usage of a less robust Modulation and Coding Scheme in power limited conditions. With a fine resolution of the coding rate in the set of available MCSs, the assignment of a less protecting Modulation and Coding Scheme should be restricted to code shortage situations.

The Hybrid ARQ represents a fast and efficient mechanism that reduces the average retransmission delay in UTRAN. Hybrid ARQ provides protection against measurement

errors and delays that can induce link adaptation errors. In Hybrid ARQ, the soft combining of the multiple transmissions of a certain packet provides time diversity and soft combining gain, which reduces the average number of transmissions for successfully decoding with obvious benefits in terms of spectral efficiency. With Incremental Redundancy, the retransmissions can decrease the effective coding rate and increase the soft combining gain, though this is expected to be significant only for high coding rates in the first transmission.

The architectural modifications of HSDPA are introduced to take advantage of the knowledge of the instantaneous signal quality in the radio interface. It enables to modify the classical CDMA concept of maintaining a certain channel bit rate and vary the transmission power to compensate for the channel fades, towards adaptation of the user's data rate while keeping the transmission power about constant. But even more relevant, the combination of the time multiplexing nature of the HS-DSCH and the knowledge of the instantaneous EsNo footprint make the Packet Scheduling a key element of the HSDPA design. This will be the subject of the following chapter.

From a complexity point of view, the deployment of HSDPA requires a Node B software update and possibly a hardware upgrade to increase the processing power of the base station. This extra processing power will allow the Node B to execute the new HSDPA functionalities (e.g. Packet Scheduling and link adaptation). Additionally, HSDPA imposes some new requirements on the mobile terminal. HSDPA capable terminals have to incorporate functionalities for the Hybrid ARQ soft combining, multi-code reception, 16QAM detection (optional), etc. The migration from WCDMA to HSDPA may initially be done only in certain areas of the network such as hot spots (indoor offices, airports, etc). Operators may gradually upgrade the rest of the network as the traffic demand evolves.

Chapter 5

Packet Scheduling for NRT Services on HSDPA

5.1 Introduction

HSDPA represents a new Radio Access Network concept that introduces new adaptation and control mechanisms to enhance downlink peak data rates, spectral efficiency and the user's QoS [66]. As commented in previous chapter, the Packet Scheduling functionality plays a key role in HSDPA. The HSDPA enhancing features and the location of the scheduler in the Node B open a new space of possibilities for the design of this functionality for the evolution of WCDMA/UTRAN.

The goal of the Packet Scheduler can be specified as to maximize the network throughput while satisfying the QoS of the users. With the purpose of enhancing the cell throughput, the HSDPA scheduling algorithm can take advantage of the instantaneous channel variations and temporarily raise the priority of the favourable users. Since the users' channel quality vary asynchronously, the time-shared nature of HS-DSCH introduces a form of selection diversity with important benefits for the spectral efficiency.

The QoS requirements of the interactive and background services are the least demanding of the four UMTS QoS classes, and have been widely regarded as best effort traffic where no service guarantees are provided. As commented in Chapter 2, the UMTS bearers do not set any absolute quality guarantees in terms of data rate for interactive and background traffic classes. However, as also stated in Chapter 2, the interactive users still expect the message within a certain time, which could not be satisfied if any of those users were denied access to the network resources. Moreover, as described in [34], the starvation of NRT users could have negative effects on the performance of higher layer protocols, as such TCP. Hence, the introduction of minimum service guarantees for NRT users is a relevant factor, and it will be taken into consideration in the performance evaluation of the Packet Scheduler in this chapter. The service guarantees interact with the notion of fairness and the level of satisfaction among users. Very unfair scheduling mechanisms can lead to the starvation of the least favourable users in highly loaded networks. These concepts and their effect on the HSDPA performance are to be investigated in this chapter.

There are several publications in the open literature that analyse the performance of Packet Scheduling algorithms that exploit the time-varying channel conditions in time shared channels. In [75], Holtzman presents the so-called Proportional Fair algorithm, which is meant to provide an appealing trade-off between cell throughput and fairness in a time-shared channel for delay tolerant data traffic. In [76], Jalali describes the performance of the Proportional Fair scheduler. This chapter will widely cover this scheduler in the following sections. In [77], Elliot presents a performance comparison of several scheduling algorithms concluding that algorithms that provided the highest throughput per sector also tended to display the largest variations in throughput per user. In [66], Kolding analyses six distinct "prototype" Packet Scheduler algorithms and concludes the inherent trade off among cell capacity and user fairness. In [78], Liu presents an opportunistic transmission-scheduling policy that maximizes the network throughput given a resource sharing constraint. In [79], Hosein illustrates a method to provide user QoS guarantees in terms of minimum data rates, while at the same time achieving some user diversity gain.

The present investigation analyses the performance of various scheduling algorithms with different degrees of fairness among users with and without exploiting the multiuser diversity for NRT users.

The chapter is organized as follows: Section 5.2 describes the input information that is available for the Packet Scheduler functionality to serve the users in the cell. Section 5.3 describes the basic Packet Scheduling strategies to be considered in the present analysis, while section 5.4 assesses their system level performance. Finally section 5.5 draws the main conclusions from the present chapter.

The overall system modelling employed for the network level simulations is described in Appendix B. The chapter uses the simple traffic model outlined in section B.3 of the Appendix B.

5.2 Packet Scheduling Input Parameters

One of the main differences between the Release 99 WCDMA architecture and the Release 5 is the location of the Packet Scheduler for HS-DSCH in the Node B [67]. The scheduler has available diverse input information to serve the users in the cell. The input parameters can be classified in resource allocation, UE feedback measurements, and QoS related parameters. The relevant parameters for this investigation regarding NRT services are described below for each category.

Resource Allocation:

- *HS-PDSCH and HS-SCCH Total Power*: It indicates the maximum power to be used for both HS-PDSCH and HS-SCCH channels. This amount of power is reserved by the RNC to HSDPA. Optionally, the Node B might also add the unused amount of power (up to the maximum base station transmission power). Note that the HS-SCCH represents an overhead power (i.e. it is designated for signalling purposes), which could be non negligible when signalling users with poor radio propagation conditions.

- *HS-PDSCH codes*: It specifies the number of spreading codes reserved by the RNC to be used for HS-PDSCH transmission.

- *Maximum Number of HS-SCCHs*: It identifies the maximum number of HS-SCCH channels to be used in HSDPA. Note that having more than one HS-SCCH enables the Packet Scheduler to code multiplex multiple users in the same TTI, and thus increases the scheduling flexibility, though it also increases the overhead.

UE Channel Quality Measurements:

The UE channel quality measurements aim at gaining knowledge about the user's supportable data rate on a TTI basis. All the methods employed for link adaptation (see section 4.6.1) are equally valid for Packet Scheduling purposes (i.e. CQI reports, power measurements on the associated DPCH, or the Hybrid ARQ acknowledgements).

QoS Parameters:

The Node B has knowledge of the following QoS parameters:

- *Allocation and Retention Priority (ARP)*: The Node B has information of the UMTS QoS attribute ARP, which determines the bearer priority relative to other UMTS bearers. See section 2.2.2 for more information about the ARP parameter.

- *Scheduling Priority Indicator (SPI)*: This parameter is set by the RNC when the flows are to be established or modified. It is used by the Packet Scheduler to prioritise flows relative to other flows [33].

- *Common Transport Channel Priority Indicator (CmCH-PI)*: This indicator allows differentiating the relative priority of the MAC-d PDUs belonging to the same flow [80].

- *Discard Timer*: Is to be employed by the Node B Packet Scheduler to limit the maximum Node B queuing delay to be experienced any MAC-d PDU. This parameter is of particular interest for streaming services and will be widely covered in Chapter 7, but it will be dismissed for NRT flows.

- *Guaranteed Bit Rate*: it indicates the guaranteed number of bits per second that the Node B should deliver over the air interface provided that there is data to deliver [33].

It is relevant to note that the corresponding mapping from UMTS QoS attributes (such as the traffic class, or the THP) to Node B QoS parameters is not specified by the 3GPP (i.e. this is a manufacturer implementation issue), nor it is determined the interpretation of these Node B QoS parameters by the Packet Scheduler.

Miscellaneous:

- *User's Amount of Data Buffered in the Node B*: This information can be of significant relevance for Packet Scheduling to exploit the multi-user diversity and improve the user's QoS.

- *UE Capabilities*: They may limit factors like the maximum number of HS-PDSCH codes to be supported by the terminal, the minimum period between consecutive UE receptions, or the maximum number of soft channel bits the terminal is capable to store.

- *HARQ Manager*: This entity is to indicate to the Packet Scheduler when a certain Hybrid ARQ retransmission is required.

5.3 Packet Scheduling Principles and Strategies

Let consider a set of users $\{i\}$ with a time varying channel quality, whose instantaneous performance is a stochastic process:

$$\{R_i(n)\} \qquad i = 1,..,N \tag{5.1}$$

where $R_i(n)$ is per-TTI maximum supportable data rate for a certain BLER if user i is served by the Packet Scheduler in TTI number n. Note that in equation (5.1) it is assumed that all the power and resources are allocated to a single user in a given TTI.

Under the previous assumptions, let define the operation task of the Packet Scheduler as to select the user i to be served in every TTI. Given the set of users in the cell $\{i\}, i = 1,..,N$ let define the operation goal of the Packet Scheduler as to maximize the cell throughput while satisfying the QoS attributes belonging to the UMTS QoS classes of the cell bearers. Figure 5.1 depicts the procedure by which the Node B Packet Scheduler selects the user to be served.

As described above, the operation of the Packet scheduler is constrained by the satisfaction of the QoS attributes of the users. Users belonging to dissimilar traffic classes require different scheduling treatment. Even within the same traffic class, the ARP attribute differentiates the service priority between the cell bearers. In the interactive class, the THP parameter further

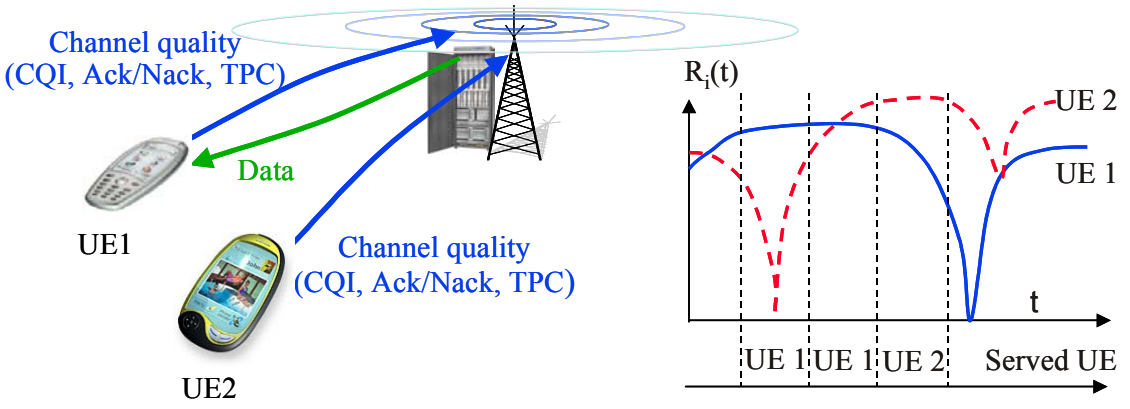


Figure 5.1: Node B Packet Scheduler Operation Procedure.

adds another dimension for bearer prioritisation. In the present assessment of the Packet Scheduler functionality, the scope will be narrowed down to NRT users without any specific QoS prioritisation between them. Chapter 7 will analyse the Packet Scheduler for streaming flows.

The Packet Scheduler rules the distribution of the radio resources among the users in the cell. In [77], Elliot describes that the scheduling algorithms that reach the highest system throughput tend to cause the starvation of the least favourable users (low G Factor users). This behaviour interacts with the fairness in the allocation of the cell resources, which ultimately determine the degree of satisfaction among the users in the cell. For this reason, this investigation concentrates on the analysis of the fairness in the radio resource allocation.

5.3.1 Fairness in the Radio Resource Assignment

When selecting the users to be served in the cell, the Packet Scheduler determines the assignment of the resources to be shared among all those users. The degree of fairness in this resource assignment relates to the level of freedom from favouritism that the users obtain regardless of their inherent conditions (e.g. radio propagation conditions, etc). In this study, three basic degrees of fairness are to be analysed:

- *C/I based*: Scheduling strategies based on a C/I policy favour users with the best radio channel conditions[9] in the resource allocation process. Such strategies maximize the system capacity at expenses of the lack of fairness because the users with poor radio channel quality will only be served when the remaining users with better radio propagation conditions run out of data to transmit. Note that users with very poor channel quality (e.g. at the cell border) experience an emphatically poorer QoS than the rest of the users in the cell, which under heavy cell load conditions can cause the starvation of those users.

- *Fair Resources*: This scheduling strategy equally distributes the cell resources (codes, power and allocation time) to the users in the cell $\{i\}, i = 1,..,N$. This leads to a lower user throughput at the cell border compared to close to the Node B, so the user experienced QoS isn't totally fairly distributed. [PI 1.4] defines the user throughput. The

[9] Either instantaneous or average radio conditions.

cell capacity of this scheme is lower than the one of a cell using C/I based scheduling, since users with a low C/I get the same amount of resources as users with a high C/I.

- *Fair Throughput*: The goal of this kind of scheduling [81] is to give all users the same throughput, regardless of their radio channel quality. Note that this fairness definition aims at the so-called *max-min fairness*. "A set of user throughputs $\{i\}, i = 1,..,N$ is said to be max-min fair if it is feasible[10] and the throughput of each user λ_i can not be increased without maintaining feasibility and decreasing λ_j for some other user j for which $\lambda_j < \lambda_i$ " [82]. This implies that users under more favourable radio conditions get the same throughput as users under less favourable conditions (e.g. users at the cell edge). This kind of scheduling can be seen as a form of inverse C/I scheduling, since users with a low C/I must be allocated larger amount of resources in order to get the same throughput. This scheduling method gives a fair user throughput distribution at the cost of a lower cell throughput.

In order to increase the fairness in the QoS experienced by the end users, users with poorer radio channel quality must be assigned larger amount of time resources to balance for their throughput. Hence, an increase in the fairness distribution of the user throughput in the cell is expected to have a cost in terms of overall spectral efficiency [66].

It is relevant to note that the notion of fairness does not provide any information of the absolute amount of resources a user can obtain, but on the relative distribution of resources among the users. This implies that, under the same degree of fairness in the resource distribution (e.g. fair resources), the user may obtain different amount of absolute resources as the number of users in the system varies with the average load (or even with temporal load variations for a given average load). However, the end user is only concerned about his absolute performance, and not about his relative performance compared to the rest of the users.

5.3.2 Packet Scheduling Algorithms

The present section describes the actual methods that implement the fairness policies described in section 5.3.1. However, the pace of the scheduling process also divides the methods in two main groups: (i) PS methods that base the scheduling decisions on recent UE channel quality measurements (i.e. executed on a TTI basis) that allow to track the instantaneous variations of the user's supportable data rate. These algorithms have to be executed in the Node B in order to acquire the recent channel quality information. These methods can exploit the multiuser selection diversity, which can provide a significant capacity gain when the number of time multiplexed users is sufficient [83]. (ii) PS methods that base their scheduling decisions on the average user's signal quality (or that do not use any user's performance metric at all).

Fast Scheduling Methods

- *Maximum C/I (Max. CI)*: This scheduling algorithm serves in every TTI the user with largest instantaneous supportable data rate. This serving principle has obvious benefits

[10] A set of user throughputs is said to be feasible if the sum of all the users throughputs is lower or equal to the link capacity.

in terms of cell throughput, although it is at the cost of lacking throughput fairness because users under worse average radio conditions are allocated lower amount of radio resources. Nonetheless, since the fast fading dynamics have a larger range than the average radio propagation conditions, users with poor average radio conditions can still access the channel.

- *Proportional Fair (PF)*: This algorithm was firstly described in [75] and further analysed in [84]-[77]. According to [76], the Proportional Fair scheduler serves the user with largest relative channel quality:

$$P_i = \frac{R_i(t)}{\lambda_i(t)} \qquad i = 1,..,N \tag{5.2}$$

where $P_i(t)$ denotes the user priority, $R_i(t)$ is the instantaneous data rate experienced by user i if it is served by the Packet Scheduler, and λ_i is the user throughput. See the averaging procedure for the computation of the user throughput for the Proportional Fair algorithm in 5.3.2.1. This algorithm intends to serve users under very favourable instantaneous radio channel conditions relative to their average ones, thus taking advantage of the temporal variations of the fast fading channel. In [75], Holtzman demonstrated that the Proportional Fair algorithm asymptotically allocates the same amount of power and time resources to all users if their fast fading are iid (identically and independently distributed) and the rate $R_i(t)$ is linear with the instantaneous EsNo. Note that this very last assumption does not hold in HSDPA due to the limitations of the AMC functionality.

The Proportional Fair method has another interesting property. According to [76], the scheduler policy provides the so-called *proportional fairness*. This criterion has been defined in [85]: a set of user throughputs $\{\lambda_i\}, i = 1,..,N$ is proportionally fair if it is feasible and if for any other feasible vector of throughputs $\{\lambda_i^*\}$ the aggregate of proportional changes is zero or negative:

$$\sum_{\forall i} \frac{\lambda_i^* - \lambda_i}{\lambda_i} \le 0 \qquad i = 1,..,N \tag{5.3}$$

This fairness criterion favours poor channel quality flows less emphatically than the max-min fairness criterion, which allows no increase in any λ_i, regardless of how large the increase is, if it is at the expenses of the throughput decrease of any other user j such $\lambda_j < \lambda_i$, no matter how small the decrease is.

- *Fast Fair Throughput (FFTH)*: This method aims at providing a fair throughput distribution among the users in the cell (in a max-min manner), while taking advantage of the short term fading variations of the radio channel. In [86], Barriac proposes an elegant modification of the Proportional Fair algorithm to equalize the user throughput:

$$P_i = \frac{R_i(t)}{\lambda_i(t)} \cdot \left[\frac{\max_j \{\overline{R_j(t)}\}}{\overline{R_i(t)}} \right] \tag{5.4}$$

where P_i describes the priority of user i, $R_i(t)$ represents the supportable data rate of user i at instant t, $\overline{R_i}$ is the average supportable data rate of user i, $\max_j\{\overline{R_j}\}$ is a constant that indicates the maximum average supportable data rate from all the users, $\lambda_i(t)$ represents the throughput of user i up to instant t. Note that the $\overline{R_i}$ term in the denominator of (5.4) compensates the priority of less favourable users, and distributes evenly the cell throughput to all users if their fast fading are iid and the rate $R_i(t)$ is linear with the instantaneous EsNo.

Slow Scheduling Methods

- *Average C/I (Avg. CI)*: This scheduling algorithm serves in every TTI the user with largest average C/I with backlogged data to be transmitted. The default averaging window length employed in this study for the average C/I computation is 100ms. Note that this window length is equivalent to 0.66 and 9.33 λ for UE speeds of respectively 3 and 50 km/h. Note that for the former UE speed the short term fading is not totally averaged out, which increases the potential of this scheduling method compared to other slow ones.

- *Round Robin (RR)*: In this scheme, the users are served in a cyclic order ignoring the channel quality conditions. This method outstands due to its simplicity, and ensures a fair resource distribution among the users in the cell. It is interesting to observe that the Round Robin scheduling method satisfies the proportional fairness criterion described by Kelly in [76]. This property is analysed in Appendix A.

- *Fair Throughput (FTH)*: There are various options to implement a fair throughput scheduler without exploiting any a priori information of the channel quality status [66]. The actual solution considered in this research serves in every TTI the user with lowest average throughput. Note that this implementation satisfies at any given instant the max-min fairness criterion described in section 5.3.1. However, this method is considered as a slow scheduling one because it does not require any instantaneous information of the channel quality.

From an implementation point of view, the slow scheduling methods have a lower degree of complexity than fast scheduling ones, because the latter require the information of the supportable data rate from the UE channel quality measurements for all the users in the cell[11], and later compute their priority on a 2 ms basis. In order to have access to such frequent and recent channel quality measurements, the Packet Scheduler must be located in the Node B. The *Average C/I* scheduling method represents a medium complexity solution because channel quality measurements are still conducted for this algorithm, although at much lower pace (~ 100 ms).

Table 5.1 summarizes the Packet Scheduling methods to be assessed in the present investigation.

[11] Even with the slow scheduling algorithms, the link adaptation algorithm still requires knowledge about the instantaneous channel quality, although this information is only needed for the single user to be served in the next TTI.

PS Method	Scheduling rate	Serve order	Radio Resource Fairness
Round Robin (RR)	Slow (~100 ms)	Round robin in cyclic order	Proportional throughput fairness & Same amount of average radio resources
Fair Throughput (FTH)	Slow (~100 ms)	Served user with lowest average throughput	Max-min throughput fairness
C/I based (CPI)	Slow (~100 ms)	Served according to highest slow-averaged channel quality	Unfair distribution of radio resources in favour of high G Factor users
Proportional Fair (PF)	Fast (~ Per TTI basis)	Served according to highest relative instantaneous channel quality	Proportional throughput fairness & Same amount of average radio resources under certain assumptions
Fast Fair Throughput (FFTH)	Fast (~ Per TTI basis)	Served according to highest equalized relative instantaneous channel quality	Max-min throughput fairness under certain assumptions
Maximum CI (Max CI)	Fast (~ Per TTI basis)	Served according to highest instantaneous channel quality	Unfair distribution of radio resources in favour of high G Factor users

Table 5.1: Summary of Analysed Packet Scheduling Methods.

5.3.2.1 Throughput Averaging for the Proportional Fair Method

The Proportional Fair algorithm intends to exploit the time varying condition of the radio channel by serving a user when the instantaneous relative channel quality outperforms the one from the remaining users in the cell. This algorithm computes the relative channel quality by adding in its priority metric the user throughput. Note that the Fast Fair Throughput also requires the user throughput averaging process for the priority metric computation.

The classical method to average the user throughput assumes an averaging window equal to the lifetime of the considered user:

$$\lambda_i(t) = \frac{\alpha_i(t)}{(t - t_i)} \qquad t \geq t_i \tag{5.5}$$

where $\alpha_i(t)$ describes the amount of successfully transmitted data by user i during the period $\in (t_i,t)$, and t_i indicates the initial instant of the user in the system.

ARMA filters are commonly used in the literature [78] to carry out the averaging process because they can accelerate the convergence of the stochastic approximation [87]:

$$\lambda_i(n) = \left(1 - \frac{1}{T_c}\right) \cdot \lambda_i(n-1) + \frac{1}{T_c} \cdot R_i(n) \tag{5.6}$$

where $\lambda_i(n)$ represents the stochastic approximation to the user throughput, $R_i(n)$ the user data rate in the present TTI t, and T_c is the time constant (in TTIs) of the low pass filter. Equation (5.6) is updated every TTI with a user data rate $R_i(n)$equal to zero if the user is not served, and with a data rate equal to the transmitted bits divided by the TTI duration if the user is scheduled and successfully receives the coded word.

The averaging length should be set long enough to ensure that the process averages out the fast fading variations, but short enough to still reflect medium term conditions such as the shadow fading [88]. In [76], the averaging window length is selected to be 1.6 seconds (≈ 10 λ at 3km/h).

It is interesting to observe that during the periods the user does not have backlogged data to transmit, the the user throughput (as defined by equation (5.5)) decreases. This reduction of the user throughput during the traffic inactivity periods artificially increases the priority computation of the Proportional Fair and the Fast Fair Throughput algorithms (equations (5.2) and (5.4)). In order to avoid that the burstiness influences the priority computations, it is decided that the stochastic approximation of the user throughput (equation (5.6)) should only be updated when the users have data queued for transmission (or when they are served). Note, that this updating technique implies that the stochastic approximation of equation (5.6) does not approximate the user throughput (as defined in equation (5.5)) during the inactivity periods that characterize the NRT traffic.

For simplicity reasons, since the traffic model employed in the present chapter does not include any traffic inactivity, this chapter will use the simple averaging process given in equation (5.5). However, Chapter 6 will employ the ARMA filter for the stochastic approximation of the user throughput (equation (5.6)) due to the burstiness of the traffic model employed in that chapter. In all cases, the metric will be initialised to zero. Alternatively, the stochastic approximation in equation (5.6) could be initialised to a default value to further reduce the initial convergence period of the filtered result.

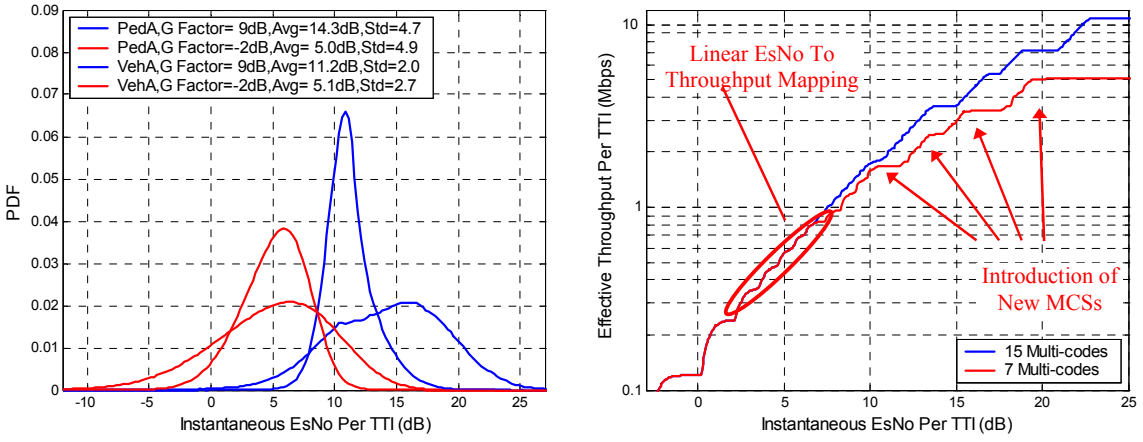
5.3.3 Multiuser Diversity

In HSDPA, the UE channel quality measurements, which enable to track the fast fading process of low speed mobiles, open the possibility of serving users under their most favourable conditions[12]. Due to the time shared nature of the HS-DSCH, the Packet Scheduler can serve at any instant the user with favourable channel quality characteristics, thereby introducing a degree of selection (multi-user) diversity, with the subsequent benefit for the user and/or system throughput.

The potential benefit of the selection diversity scheme completely depends on the dynamics of the short term fading, and the ability of the Node B to obtain recent and accurate channel quality measurements of the fast fading process. Figure 5.2 a) shows the pdf of the instantaneous symbol energy to noise ratio (EsNo) per TTI for Pedestrian A and Vehicular A at 3 km/h for 9 and –2 dB G Factors obtained under the model of the quasi-dynamic network simulator described in Appendix B (around 50% of average Node B transmission power is allocated for HS-DSCH). Note that the per TTI EsNo computation in Figure 5.2 a) is a geometrical average over three slots. See the EsNo definition in [PI 4.1].

The picture shows the strong similarities between the distributions of the instantaneous EsNo for both G Factors in Pedestrian A. This is because the thermal noise and the other cell interference are dominant over the intracell interference, which leads to the similarity in the distribution with the only main difference of the average instantaneous EsNo. It is also interesting to note that the dynamic range of the received symbol energy over the noise is

[12] Note that very low speed or static mobiles hardly experience any Doppler signal quality variation, and can hardly benefit from the selection diversity process.



a) *b)*

Figure 5.2: (a) PDF of the Instantaneous EsNo Per TTI for Pedestrian A and Vehicular at 3 km/h for two Different G Factors (b) 1st Transmission Effective Throughput for Various Multi-codes in Pedestrian A 3km/h from [88].

around 20-25 dB for Pedestrian A (and for a given G Factor). For Vehicular A environment, the average EsNo is not degraded compared to Pedestrian A for low G Factor users due to the dominance of the other cell interference over the intra cell interference, however there is a significant difference in the dynamic range due to the two dominant paths of the power delay profile of the Vehicular A channel model. For higher G Factor users, this multipath delay profile causes not only the reduction of the dynamic range, but also the decrease of the average instantaneous EsNo compared to the Pedestrian A channel model.

Figure 5.2 b) plots the mapping from instantaneous EsNo per TTI to effective throughput per TTI in the first transmission in Pedestrian A for 7 and 15 multi-codes [88]. See the definition of the first transmission effective throughput in [PI 5.1]. The mapping presents a linear EsNo to throughput mapping (e.g. a 3 dB instantaneous EsNo increase leads to a double effective throughput) in the region where the dynamic range of the EsNo is covered with the most spectral efficient MCS and the full range of available multi-codes. Note how this linear region reaches up to around 10 and 7 dB for respectively 15 and 7 available multi-codes. Beyond this region, the instantaneous EsNo dynamic range is coped with the introduction of the higher MCSs, but the introduction of these MCSs is at the expenses of less spectrally efficient EsNo to throughput mapping (i.e. non linear). The link adaptation dynamic range is upper limited by the last MCS and the maximum number of codes, and any EsNo increase over this saturation point does not produce any throughput benefit. See in Figure 5.2 b) how the 15 codes extend the total dynamic range around 3 dB over the 7 codes case.

The process of multiuser selection diversity obtains its benefits from serving the users on the top of their fast fading characteristics. However, the dynamic range covered by the link adaptation may limit the diversity gain. From Figure 5.2, it can be seen that the distribution of the instantaneous EsNo of a low G Factor user in Pedestrian A mostly falls within the region of linear EsNo to throughput mapping. This implies that the EsNo improvement obtained by the user while being served under his favourable conditions is linearly translated into a user throughput improvement. However, a 9 dB G Factor user in Pedestrian A will employ the less spectrally efficient MCSs to cope with the dynamic range of the signal. This implies that, although the selection diversity can provide a certain EsNo gain, this improvement is not

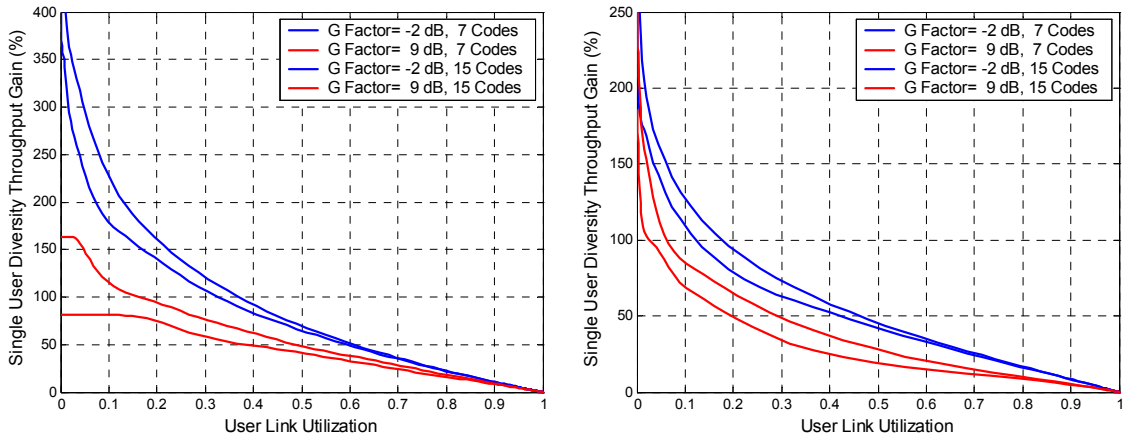


Figure 5.3: Single User Diversity Gain versus the User Link Utilization Assuming the User is Served Under his Most Favourable Channel Quality Conditions for a) Pedestrian A and b) Vehicular A.

linearly translated into user throughput. Moreover with a certain probability the user will overcome the throughput saturation point and the EsNo gain will not provide any throughput benefit for such a high G Factor user.

In Vehicular A channels, the lower dynamic range of the instantaneous EsNo and the lower average EsNo for high G Factor users in such an environment, decreases the risk of the throughput gain being limited in the link adaptation process.

Let consider the set of HSDPA users $\{i\}$ coexisting in the cell. Let define the link utilization factor of user i, denoted as ρ_i, as the proportion of time the HS-DSCH link is busy transmitting packets to this user. Then, the overall HS-DSCH utilization can be defined as the sum of the utilization of all the users:

$$\rho = \sum_i \rho_i \qquad i = 1,..,N \tag{5.7}$$

where ρ is the overall HS-DSCH utilization factor.

Let assume that the Packet Scheduler can ideally serve the user under his most favourable per-TTI EsNo conditions (i.e. ideal channel measurements, negligible delay in signalling the UE measurements, negligible Node B processing delay, and negligible delay for the Node B to signal the token to the served user). The throughput achieved by a user by serving him a certain proportion of time (e.g. 50%) can be obtained by collecting the EsNo statistics corresponding to that proportion of time and mapping them into throughput (see Figure 5.2). Hence, it can be computed the potential throughput gain achieved by serving the user under his most favourable channel quality conditions relative to scheduling the user randomly (i.e. without information of the instantaneous channel variations). Figure 5.3 plots that throughput gain versus the user link utilization factor for ITU Pedestrian A and Vehicular A at 3km/h.

The first conclusion from Figure 5.3 is that the throughput gain of serving the user on his fading tops increases as the link utilization of the user decreases. However, as the user link utilization decreases, the absolute user throughput decreases because the user holds the link

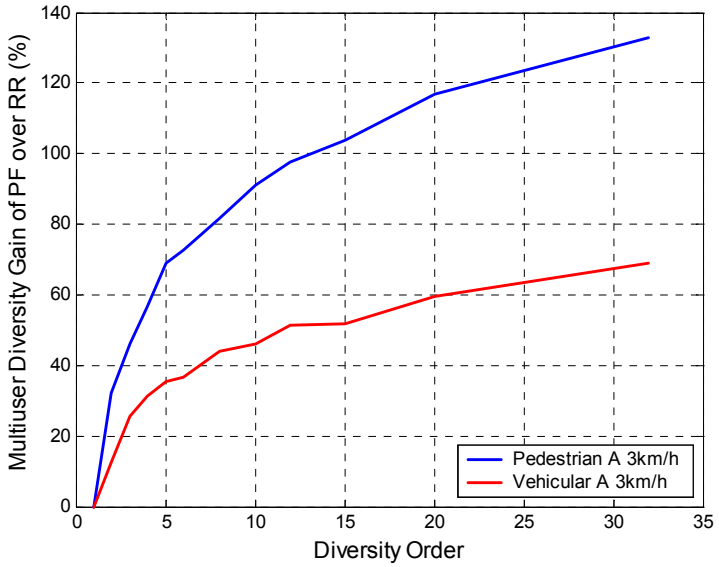


Figure 5.4: Multiuser Diversity Gain of PF over RR versus the Number of Users to Be Mapped into HS-DSCH in Overload for Pedestrian A and Vehicular A at 3km/h.

for shorter time. Note that the lower the link utilization of the users, the larger the number of users can be accommodated into the HS-DSCH. Hence, the multiuser diversity gain should be considered as a cell throughput benefit rather than a user throughput improvement because the larger gains are obtained for lower user's throughput. Therefore, the final cell capacity gain of the multiuser diversity process (relative to schedule the users ignoring the radio channel conditions) depends on the link utilization requirements of the most needy users (low G Factor users). This relates to the minimum throughput requirements of the end users.

Another conclusion from Figure 5.3 is that for low G Factor users in Pedestrian A, the gain is hardly affected by the dynamic range of the link adaptation process (for the number of multi-codes considered in this example), whereas high G Factors users are much more limited by the non linearity of the instantaneous EsNo to throughput mapping and by the throughput saturation point. Due to the smaller dynamic range of the instantaneous EsNo of Vehicular A users than Pedestrian A ones, their potential throughput gain is significantly reduced.

Previous conclusions were drawn assuming that the scheduled user is always served under his most favourable short term conditions. When various users share the HS-DSCH link, there is a certain probability that, simultaneously, the signal quality of none of them is under favourable conditions. Hence, the gain figures plotted in the previous pictures represent a potential user throughput improvement. In order to compute the multiuser diversity gain including this phenomenon, a network level simulation is done where the cell throughput is compared between a fast and a slow scheduler. With the purpose of having a similar distribution of resources between both scheduler types, the Round Robin and the Proportional Fair methods are selected for the comparison. The simulation parameters are described in sections B.1 to B.3 of the Appendix B. The simulation has been carried out in full overload. The multiuser diversity order is controlled with the number of users in the cell.

Figure 5.4 represents the multiuser diversity gain of the Proportional Fair over the Round Robin method as a function of the diversity order (number of users) in overload for Pedestrian

A and Vehicular A channels at 3km/h and for 7 multi-codes. Assuming a fair distribution of resources among the users in the cell (as ideally achieved by Round Robin and Proportional Fair methods), the diversity order is inversely proportional to the user link utilization.

The results show that a multiuser diversity gain of around 100% for Pedestrian and 50% for Vehicular is already achieved with a diversity order of 10-15. However, the gain does not fully saturate for the Pedestrian A channel within the considered range of diversity order. This could be already expected because the diversity gain to be achieved by a single low G Factor user (larger than 200% as shown in Figure 5.3) is still significantly higher than the maximum system diversity gain described in Figure 5.4. Note as well the logarithmic-like increase of the multiuser diversity gain with the diversity order. For the Vehicular A channel the gain becomes smoother beyond a diversity order of 15-20. It is also relevant to note that the gain of Pedestrian A nearly doubles the one of Vehicular A for 32 users.

However, the burstiness of packet data transmission causes the diversity order to get reduced for a given number of users mapped into the HS-DSCH. The effect of the packet data burstiness into the HS-DSCH performance is to be analysed in section Chapter 6.

5.4 Performance Analysis of Scheduling Methods

The present section analyses the HSDPA performance of the scheduling methods described in section 5.3.2. This analysis has been conducted by means of quasi-dynamic network simulations whose overall modelling is detailed in sections B.1 to B.3 of the Appendix B.

5.4.1 Distribution of the User Throughput

Besides the multiuser diversity introduced by the fast scheduling methods, the main difference of the scheduling principles to be analysed in this study is the degree of fairness in the allocation of the available radio resources among the users.

Figure 5.5 shows the cdf of the user throughput for the scheduling methods described in section 5.3.2 in Pedestrian A with a mobile speed of 3km/h. The simulations have been carried out under heavy cell input load (2 Mbps for slow algorithms, whereas fast ones have been offered an input load of 3 Mbps due to their higher system capacity) because this permits to appreciate more clearly the difference in prioritisation between the different schedulers independently of the temporal load variations (see section 5.3.1).

The results show that the fairness determines the slope of the cdf of the user throughput, and while fair throughput strategies have a very steep distribution of the user throughput, fair resources strategies have a less steep cdf because users with better average radio conditions better utilize the same amount of resources. Moreover, it can be seen that, for either slow or fast scheduling, the fair throughput curve outperforms the fair resources one from 0% to the 10% percentile, whereas fair resources outperforms for the rest of the cdf. This implies that the little user throughput improvement for the 10% poorer users provided by fair throughput is at expenses of the remaining 90% of the users. Furthermore, C/I methods have a much smoother cdf due to the unfairness in the resource distribution, and the users on the high percentile (e.g. 90%) of the cdf achieve very large throughputs (e.g. 2-3 Mbps) at expenses of the throughput obtained by the users in the tail of the cdf (e.g. percentiles below 10%).

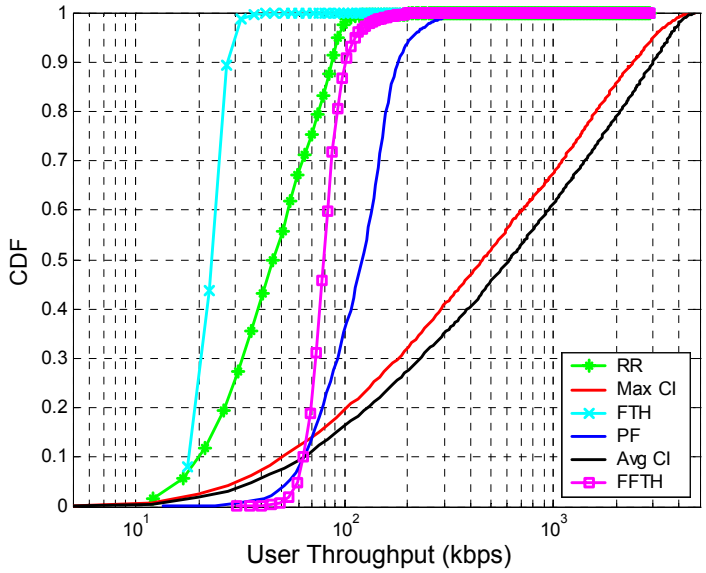


Figure 5.5: CDF of the User Throughput for the Scheduling Methods described in Section 5.3.2 in Pedestrian A at 3km/h. Cell Load for Slow Scheduling Methods equals 2 Mbps, and for Fast Scheduling Methods equals 3 Mbps.

Another relevant conclusion from Figure 5.5 is that the distribution of the user throughput of fast scheduling methods approximates a shifted version of the slow ones due to the user throughput improvement introduced by the multiuser diversity. It is interesting to observe that the user throughput improvement is significantly higher for fair throughput than for fair resources. This is because the low G factor users achieve the major benefit from the multiuser diversity (see section 5.3.3) and these users receive the larger amount of resources in the fair throughput scheduling strategies. Regarding the CI scheduling methods, the Avg CI scheduler actually outperforms the one of the Max CI in Figure 5.5, however the input load for the latter was 50% higher than for the former. For the same offered input load, Max CI is still expected to surpass Avg CI (see in Figure 5.8 b) the performance of MaxCI with an offered input load of 2Mbps).

The comparison between the Avg. CI and the rest of the slow mechanisms show a significant user throughput improvement because there is still a capacity gain in utilizing the average signal quality in the serving process. However, in the case of the Max CI scheduler, the significant throughput benefit of the users in the upper part of the cdf (beyond 15% outage) is at expenses of the very poor performance of the users in the lower part of the cdf (e.g. 5% of the users achieve a throughput below 30 kbps).

Figure 5.6 depicts the user throughput under the same simulation conditions as Figure 5.5 as a function of the G Factor. The user throughput is averaged in bins of 1 dB G Factor. The results confirm that the tails of the cdf of the user throughput in the Figure 5.5 comprise the low G Factor users (i.e. users under poor average path loss and shadow fading conditions).

Figure 5.7 plots the distribution of the number of HS-PDSCH codes and the MCS in Pedestrian at 3 km/h for Round Robin, Fair Throughput, Proportional Fair, and Fast Fair Throughput. The description of the MCSs is given in Chapter 4. The results show how, under the same degree of fairness, the fast scheduling algorithms utilize more frequently the higher

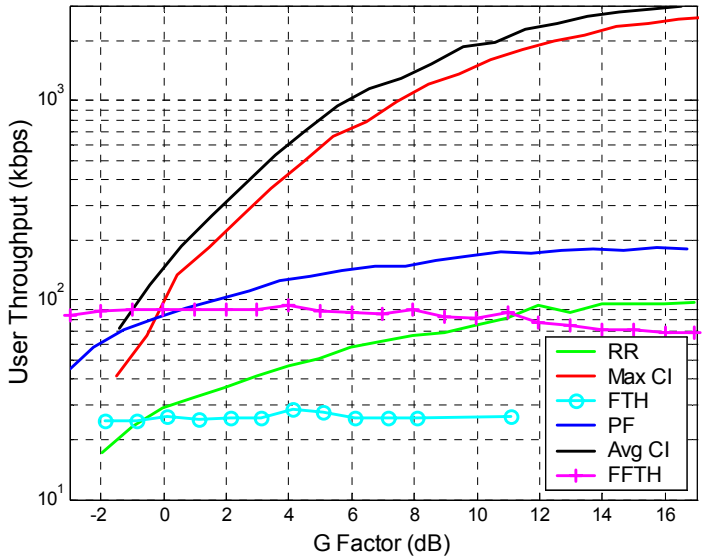
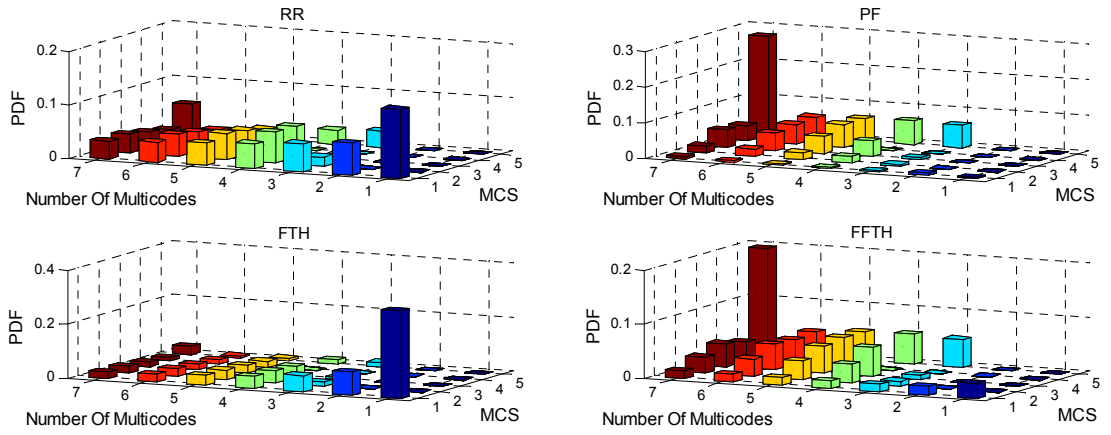


Figure 5.6: Average User Throughput versus the User's G Factor for the Scheduling Methods described in Section 5.3.2 in Pedestrian A at 3km/h. Cell Load for Slow Scheduling Methods equals 2 Mbps, and for Fast Scheduling Methods equals 3 Mbps.

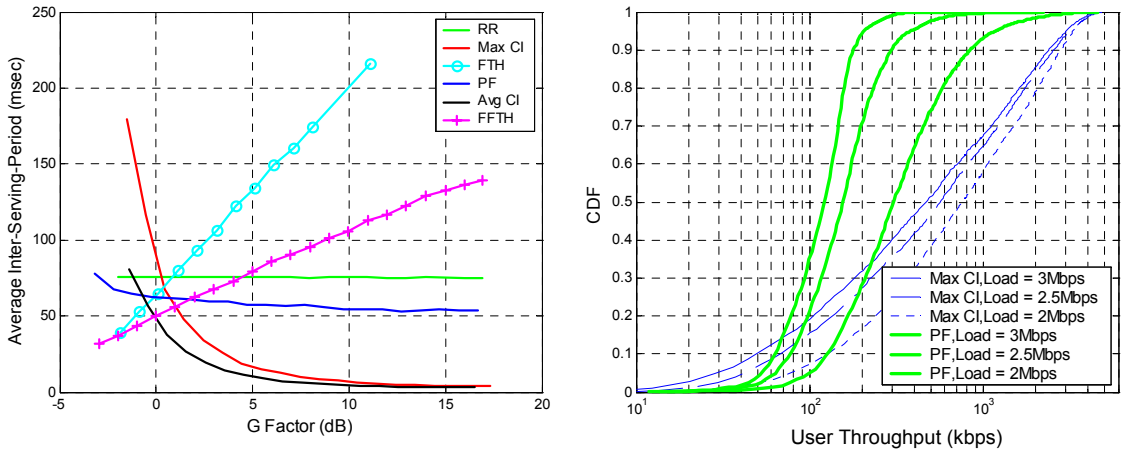
MCSs because these methods can serve the users on the top of their fading characteristics. Moreover, the large usage of the largest MCS (MCS 5) and the largest number of multi-codes (7 codes) suggests that the dynamic range of the link adaptation may not be sufficient to cope with the fading tops of certain users. It is also noticeable in Figure 5.7 the high usage of the first MCS with a single spreading code for the Fair Throughput algorithm, which gives an indication of the low efficiency of this method.

Another clear trend to be extracted from Figure 5.7 is that fair resources policies make use more often of higher order MCSs and larger number of multi-codes than fair throughput



a) *b)*

Figure 5.7: Distribution of the Number of Multi-codes and MCS in Pedestrian A at 3km/h for a) Round Robin & Fair Throughput and b) Proportional Fair & Fast Fair Throughput.



a) *b)*

Figure 5.8: a) Average Inter-Serving Period for the Scheduling Methods described in Section 5.3.2 in Pedestrian A at 3km/h. Cell Load for Slow Scheduling Methods equals 2 Mbps, and for Fast Scheduling Methods equals 3 Mbps. H-ARQ Retransmissions not Included in the Computation. and b) Average User Throughput Variations of Max CI & Proportional Fair with Load in Pedestrian A at 3km/h.

policies because the latter serve more frequently users under worse average radio quality conditions.

Let define the average Inter-Serving Period of user *i*, denoted as $\overline{ST}_i$, as the mean period elapsed between consecutives instants when the scheduler serves the user *i*. Note that the user link utilization factor can be approximated as inversely proportional to the average Inter-Serving Period if the number of times the user is served is sufficiently large:

$$\frac{\overline{ST}_i}{TTI} \approx {}^{1}\!/_{\rho_i} \qquad i = 1,..,N \tag{5.8}$$

where *TTI* represents the TTI duration (i.e. 2 ms).

Figure 5.8 a) plots the average Inter-Serving Period versus the G Factor in order to evaluate how the different schedulers distribute the cell resources among the users. Observe that H-ARQ retransmissions are not included in the computation of the Inter-Serving Period. Considering the slow scheduling algorithms, as expected, the Round Robin algorithm represents an even share of the HS-DSCH resources among users, while the Avg. CI method serves more sporadically the users under worse average radio conditions. On the other hand, the average Inter-Serving Period shows a linear behaviour with the G Factor for the Fair Throughput algorithm. This linear behaviour inherits from the quasi-linear increase of the average user supportable data rate with the G Factor for this algorithm:

$$\lambda_i = \rho_i \cdot R_i = \frac{TTI}{\overline{ST}_i} \cdot R_i \qquad i = 1,..,N \tag{5.9}$$

where λ_i is the user throughput, which is the same throughput for all the users under this scheduler. Working out the average Inter-Serving Period, it is extracted the linear relation

between this metric and the supportable data rate, and hence the linear relation between the Inter-Serving Period and the G Factor. For the Fair Throughput algorithm:

$$\overline{ST}_i = \frac{TTI}{\lambda} \cdot R_i \qquad i = 1,..,N \tag{5.10}$$

The fast scheduling methods show similar trends to slow ones. One of the main differences is that the Proportional Fair algorithm allocates a larger (but not much larger) fraction of time to users with higher G Factor, so that the resource share is not utterly even among the users. This is mainly caused by the non linearity in the EsNo to throughput mapping (see Figure 5.2), mainly suffered by high G Factor users. This agrees with the conclusions about the Proportional Fair scheduler in [84] that state "the users with greater (signal quality) variability use a lower fraction of time transmitting". The results for the Fast Fair scheduler show another linear relation between the average Inter-Serving Period and the G Factor.

Besides the inherent trade-off between the cell throughput and the user achievable QoS (e.g. the user throughput) [89], the effect of the load on the distribution of the throughput among users in the cell is very different depending on the degree of fairness applied by the scheduling policy. Figure 5.8 b) depicts the cdf of the average user throughput of Max CI and Proportional Fair schedulers in Pedestrian A at 3km/h for various input loads. While in Max CI especially the poor users suffer from the load increase, in Proportional Fair a load increase causes a shift in the average user throughput of the users but still maintaining the relation in the resources distribution among the users in the cell (i.e. the slope of the cdf hardly varies with the load).

5.4.2 HSDPA Cell Throughput with Minimum User Throughput Guarantees

The present section aims at evaluating the HSDPA cell throughput provided by the different scheduling policies. Recall from section 5.3 that the operation goal of the Packet Scheduler is to maximize the cell throughput under certain constraints.

If no constraints were imposed on the users' QoS, the CI schedulers would undoubtedly provide maximum cell capacity because they serve the users under most favourable conditions (and hence supportable data rate). More specifically, the Max CI algorithm, with the benefit of the multiuser diversity, is the scheduler reaching the highest system capacity. This can already be concluded from Figure 5.5 where the average user throughput of the Max CI scheduler outperforms the rest of the fast scheduling methods in the cdf from the 15% outage onwards[13]. This trend was already found in [66], [77]. However, for highly loaded networks, the users having poor average radio channel conditions are allocated very little time resources, and thus starve for throughput, and will ultimately be dropped without the network satisfying their service. Therefore, this scheduler does not fully meet the QoS needs of interactive users, which, according to the description of the interactive UMTS QoS class [27], expect the message during a certain time. Note that this is not the case of background users.

[13] Note that the cdf of the average user throughput of the Avg CI scheduler actually outperforms the one of the Max CI in Figure 5.5, however the input load for the latter was 50% higher than for the former.

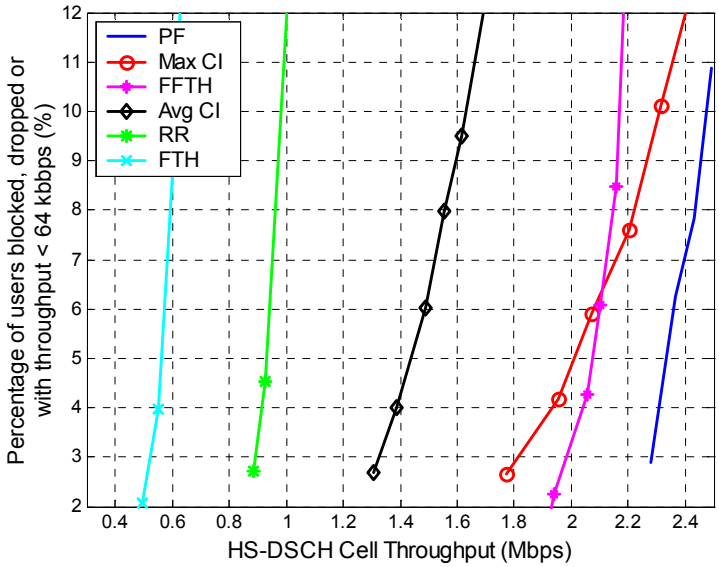


Figure 5.9: Cell Throughput Performance for NRT Bearers in Pedestrian A at 3km/h with User's Minimum Data Rate of 64 kbps. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

The cell throughput of the proposed scheduling strategies will be assessed under minimum user throughput guarantees, which ensures that the network renders the service to the users within a given time. With this constraint, the goal will be to search for a scheduling policy that maximizes the cell throughput. In order to test different QoS requirements, three minimum user throughput guarantees are considered: 64, 128 and 384 kbps. Hence, users not satisfying this minimum throughput will be counted as an outage. See section B.3 of the Appendix B for the outage definition for NRT services. The outage level of interest is considered to be 5%.

Figure 5.9 shows the HS-DSCH cell throughput for NRT bearers in Pedestrian A at 3 km/h for all the scheduling policies described in section 5.3.2. The minimum user throughput considered in the outage computation has been 64 kbps. See the definition of the HS-DSCH cell throughput in [PI 5.2].

Obviously, all the schedulers show the aforementioned inherent trade-off between the cell throughput and the user throughput. This trade-off implies that the higher the load in the cell is, the higher throughput the system can achieve, but the less resources are available for every single user, and, therefore, the lower throughput can be achieved by the users. One of the first conclusions from Figure 5.9 is that the Proportional Fair algorithm is the method performing best for the outage level of interest. This algorithm provides an interesting trade-off between cell throughput and user fairness because the proportional fairness property ensures that poor users get a certain share of the cell resources that enables them to strive for acceptable QoS, without still consuming the overall capacity, and letting the remaining share of those resources to be exploited by better performing users (medium or high G Factor ones) who can raise the cell throughput.

The very unfair distribution of resources of Max CI causes that the poor users incur outage unless the cell load is low, which turns to yield a low cell capacity. However, the slope of the cell throughput versus the outage of Max CI is smoother than for Fast Fair Throughput or

Scheduler	FTH	RR	Avg. CI	FFTH	PF	Max CI
Avg. HS-DSCH Cell Throughput (Mbps)	0.56	0.93	1.44	2.08	2.33	2.01
Blocking (%)	0.0	0.1	0.0	3.1	1.5	0.0
Dropping (%)	0.0	0.1	1.6	0.0	0.4	1.5
Users Achieving < 64 kbps (%)	5.0	4.8	3.4	1.9	3.1	3.5

Table 5.2: Cell Throughput Performance for NRT Bearers in Pedestrian A at 3km/h with User's Minimum Data Rate of 64 kbps at 5% outage level.

Proportional Fair, which implies that for higher outage levels the cell throughput of Max CI tends to outperform the rest of the schedulers (see in Figure 5.9 how the capacity of Max CI crosses the one of Fast Fair Throughput at around 6.5% outage, and approximates close to the capacity of Proportional Fair at 11% outage). On the other hand, the Fast Fair Throughput method provides the steepest slope of all the fast methods, and, although not depicted in Figure 5.9, it is expected that at very low outages it can outperform the Proportional Fair algorithm.

Regarding the slow scheduling methods, the Average CI algorithm obtains significant cell throughput gain from the average signal quality information and outperforms Fair Throughput by 157% and Round Robin by 55% (at 5% outage).

It is also relevant to note that the gain of fast scheduling methods varies with the degree of fairness provided by the methods. While Max CI provides only 40% cell throughput gain at 5% outage, Proportional Fair improves the cell throughput by 150% over Round Robin. Moreover, the Fast Fair Throughput method achieves a cell throughput gain of 270% over its slow version. The mechanisms that allocate larger amount of resources to low G Factor users (i.e. fair throughput) achieve the higher benefit from multiuser diversity because the user gain is not limited by the non linearity and the saturation of the EsNo to throughput mapping in the link adaptation process.

Table 5.2 describes the degree of contribution of the blocking and the dropping in the outage level depicted in Figure 5.9. It can clearly be seen that the dropping is a very significant contributor to the outage for CI algorithms, whereas the blocking is the main reason causing the outage for Proportional Fair and Fast Fair Throughput.

It is also interesting to note that the cell throughput gain of Proportional Fair over Round Robin at 5% outage (≈150%) overcomes the one achieved in overload (≈ 133%, see Figure 5.4). The reason behind this effect is linked to the definition of the outage criterion (see B.3) and its relation to the blocking. On the one hand, the Proportional Fair algorithm can still provide a user throughput of 64kbps at high loads levels that already start to cause blocking. On the other hand, in order to support a user throughput of 64kbps, the Round Robin scheduler has to decrease the load below the operating points that produce blocking.

Figure 5.10 draws the HSDPA cell throughput for NRT bearers in Vehicular A at 3 km/h with a minimum user throughput of 64 kbps. The most outstanding peculiarity of the results is the significant cell capacity degradation compared to the Pedestrian A power delay profile due to the interpath interference of time dispersive environments. The HSDPA cell throughput is reduced by approximately 50% for the fast scheduling methods. The cell throughput gain of Proportional Fair over Round Robin has been reduced to around 70%, which matches the gain

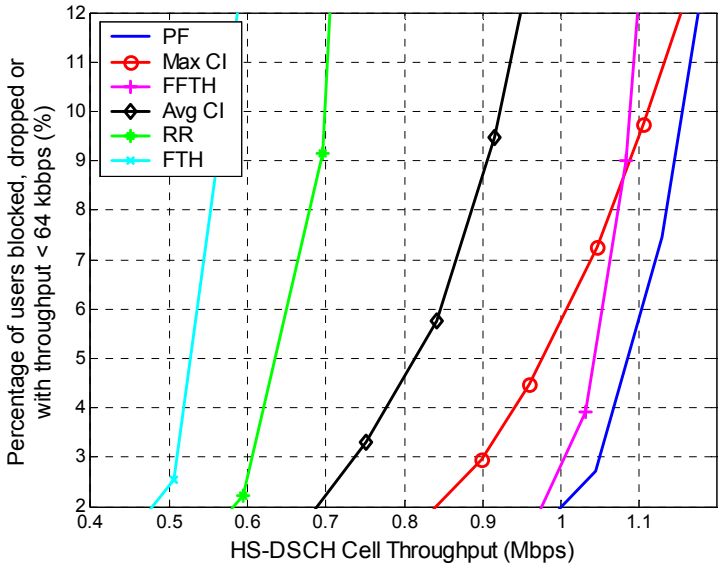
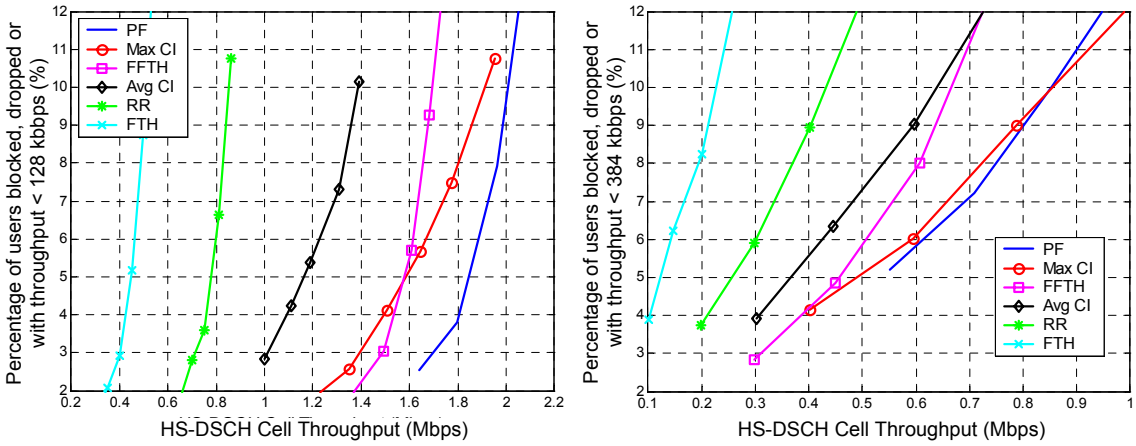


Figure 5.10: Cell Throughput Performance for NRT Bearers in Vehicular A at 3km/h with User's Minimum Data Rate of 64 kbps. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

in overload (see Figure 5.4). In any case, it is also relevant to note that Proportional Fair is still the scheduler providing the largest cell throughput in the outage range of interest.

Figure 5.11 plots the cell throughput for NRT bearers in Pedestrian A at 3 km/h but with a minimum user throughput for the outage computation equal to 128 and 384 kbps. The results with the minimum user throughput of 128 kbps show an acceptable cell throughput reduction compared to the results of the 64 kbps case (from 2.33 down to 1.85 Mbps at 5% outage with Proportional Fair, i.e. a cell throughput reduction of 20%), however the relative performance between the different scheduling strategies is very similar between both cases.



a) *b)*

Figure 5.11: Cell Throughput Performance for NRT Bearers in Pedestrian A at 3km/h with User's Minimum Data Rate of a) 128 kbps and b) 384 kbps. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

Scheduler		FTH	RR	Avg. CI	FFTH	PF	Max CI
Min. User Throughput 64kbps	HS-DSCH Cell Throughput (Mbps)	0.56	0.93	1.44	2.08	2.33	2.01
	Avg. Num. Users	2.3	4.5	5.2	17.5	15.9	7.8
	Avg. User Thr. at 90% CDF (Mbps)	2.38	1.54	3.1	0.27	0.39	2.76
Min. User Throughput 128kbps	HS-DSCH Cell Throughput (Mbps)	0.45	0.78	1.16	1.58	1.85	1.58
	Avg. Num. Users	1.0	2.0	2.6	5.7	6.2	3.8
	Avg. User Thr. at 90% CDF (Mbps)	2.75	2.06	3.12	0.93	0.99	2.96
Min. User Throughput 384kbps	HS-DSCH Cell Throughput (Mbps)	0.12	0.26	0.37	0.46	0.54	0.49
	Avg. Num. Users	0.13	0.27	0.37	0.52	0.58	0.5
	Avg. User Thr. at 90% CDF (Mbps)	3.19	3.04	3.26	2.91	2.81	3.21

Table 5.3: Cell Throughput Performance for NRT Bearers in Pedestrian A at 3km/h with User's Minimum Data Rate of 64, 128, and 384 kbps at 5% outage level. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

Providing minimum user throughput guarantees of 384 kbps is still feasible, though it has high cost in terms of spectral efficiency, and the cell throughput with Proportional Fair is reduced from 1.85 Mbps down to 540 kbps when changing the minimum throughput from 128 kbps to 384 kbps, which represents a cell throughput reduction of 71%. Note as well the significant difference in the slope of the load curve with minimum throughput of 384 kbps compared to the 128 and 64 kbps cases.

Table 5.3 summarizes the cell throughput results at 5% outage level with the three minimum user throughputs considered in this study. The table adds the average number of users mapped into HS-DSCH at the same outage level. The table also includes the user throughput at the 90% of the cdf, which gives an indication of the data rate achieved by the best performing users. The results indicate that for high minimum user throughput (384 kbps) to be provided by the network, large amount of time resources must be allocated per user, and therefore the less number of users can be supported in the cell. Note however, that for minimum user throughput of 128kbps around 10% of the users reach a throughput higher than 1Mbps (see Table 5.3) and 45% achieve a throughput larger than 384kbps (not depicted in Table 5.3).

5.5 Conclusions

The present chapter has analysed the performance of various Packet Scheduling strategies for NRT traffic in HSDPA. The investigation has specially concentrated on assessing different scheduling methods with different degrees of fairness because the fairness in the distribution of the radio resources ultimately determines the QoS perceived by the users. In addition, the Packet Scheduling strategies have also aimed at exploiting the multiuser diversity provided by the time-shared nature of the HS-DSCH transport channel.

As described in section 5.3.3, the capacity gain by multiuser diversity increases as the population of users in cell enlarges, which implies that the larger diversity gains are achieved for lower absolute user's throughputs. With 10-15 active users, the multiuser diversity gain of Proportional Fair over Round Robin is around 100% and 50% respectively in Pedestrian A

and Vehicular A at 3km/h. With 32 active users the gain increases up to 130% and 70% in previous environments.

From the results it has been concluded that Max CI is the scheduling method providing highest HSDPA cell capacity for best effort NRT traffic, although the unfairness of this policy causes the starvation of users under poor average radio propagation conditions. This starvation has very negative effects on the performance of the TCP protocol, and moreover, it can cause the dropping of the poor users if the starvation is prolonged.

When minimum user throughput guarantees are to be provided to the overall user community, the Proportional Fair scheduler clearly outperforms the remaining analysed schedulers for all the tested cases. Under the overall system model described in Appendix B (HSDPA power around 50% of Node B transmission power and seven HSDPA multi-codes), the Proportional Fair scheduling principle provides a cell throughput of around 2.3 Mbps and 1.1 Mbps at 5% outage with a minimum user throughput of 64 kbps at 3km/h in Pedestrian A and in Vehicular A respectively. The Proportional Fair policy, by favouring poor users less emphatically than the fair throughput strategies but still ensuring them a certain share of the cell resources that enables them to strive for acceptable QoS, provides an attractive trade-off between cell capacity and user fairness by conveniently exploiting the multiuser diversity. Unfortunately, the time dispersion of the channel has a very harmful influence on the cell throughput that is degraded by around 55% from the Pedestrian A to the Vehicular A environment.

Proportional Fair provides a cell throughput gain of around 150% and 70% over Round Robin at 5% outage with a minimum user throughput of 64 kbps respectively in Pedestrian A and Vehicular A. The Fast Fair Throughput scheduling method also appears as an attractive principle because it provides max-min throughput fairness while its capacity versus that of Proportional Fair is only reduced by 10% in Pedestrian A at 3km/h and by 4% in Vehicular A at 3km/h at an outage of 5% with a minimum user throughput of 64 kbps.

The provision of higher user throughputs such as 128 kbps shows an acceptable cell throughput reduction compared to the results with minimum user throughput of 64 kbps (from 2.33 down to 1.85 Mbps at 5% outage in Pedestrian A at 3km/h with Proportional Fair scheduler, i.e. a cell throughput reduction of 20%), whereas the provision of higher user throughput such as 384 kbps is still feasible though it has a high cost in terms of spectral efficiency (that is reduced 71% from supporting a minimum user throughput of 128 kbps to 384 kbps in Pedestrian A at 3km/h with Proportional Fair scheduler).

Let compare previous results with the WCDMA cell capacity figures provided in the available literature. The maximum capacity achieved by the Proportional Fair algorithm in macro cell (Vehicular A) at 3 km/h with an HSDPA power around 50% of Node B transmission power and seven multi-codes is around 1.45 Mbps, while the corresponding capacity for Round Robin equals around 850 kbps. A rough estimate of the cell capacity assuming 80 % of HSDPA (the remaining cell power is supposed to be overhead) could be obtained by simply rescaling previous results. Comparing[14] these absolute cell capacity figures with the WCDMA capacity, reported in [10] to be 915 kbps for macrocell environment (and fair resources distribution among users in the cell), yields an absolute HSDPA capacity gain of 150% of with Proportional Fair, and a gain of around 50% with Round Robin over WCDMA.

[14] Note that the comparison of absolute cell capacity figures is difficult when the underlying assumptions to provide those figures are not exactly the same. Therefore, this capacity comparison must be considered as a first order approximation.

Chapter 6

System Performance of TCP Traffic on HSDPA

6.1 Introduction

The present chapter evaluates the HSDPA system performance of TCP traffic. As described in Chapter 3, the reliability provided by the TCP transport protocol is of particular interest for interactive and background applications due to the error free message delivery. As also stated in Chapter 3, the majority of Internet traffic is based on a TCP/IP platform. Then, there exist strong reasons to focus on TCP traffic when evaluating the behaviour of interactive and background applications on wireless networks, including HSDPA. The effects of the TCP flow control have strong impact on both the end user and the wireless network performance, and hence must be seriously considered in the analysis of the behaviour of NRT services.

For simplicity reasons, the present chapter will only concentrate on one fast and one slow scheduling strategy. The scheduling principles considered are Proportional Fair due to its exceptional performance, and Round Robin because it has the same fairness properties as Proportional Fair. The traffic modelling employed for the system level simulations conducted for this section includes a simplified version of the TCP protocol (described in B.4). The remaining system model and default parameters is kept as in previous chapter.

The chapter is organized as follows: section 6.2 gives a general overview of the main aspects of the performance of TCP traffic on HSDPA. Section 6.3 analyses the interaction between the system throughput, the Node B queuing delay, and the user throughput, while section 6.4 investigates the effect of the TCP traffic activity (burstiness) on the HSDPA cell throughput. Section 6.5 draws the main conclusions of the chapter.

6.2 TCP Packet Traffic on HSDPA

The characteristics of the HS-DSCH transport channel have been tailored to increase the packet data performance and deliver high user throughput while reaching significant cell throughput gain over the DCH based system. The time-shared nature of HS-DSCH is particularly beneficial to cope with the sources of channelization code inefficiency such as the inherent burstiness of the TCP's slow start or the DCH inactivity timer.

The advantages of the statistical multiplexing on the code tree usage have already been shown for DSCH channels [90], [91]. However, DSCH channels suffer from a higher cost in terms energy per data bit than DCH channels due to the lack of soft handover gain of the former channels [12]. Moreover, the combination of fast power control and the considerable amount of power allocation in DSCH channels causes significant NodeB transmission power variations with negative effects on the overall system stability.

On the other hand, HSDPA has been designed to provide high user throughputs achieved by the combination of high order modulations and fast packet scheduling. Note that for TCP traffic, the increase of the available data rate of the connection raises the bandwidth-delay product and therefore reduces the user activity. However, the statistical multiplexing of the users onto the HS-DSCH allows to cope with the traffic inactivity periods. Moreover, as already commented in Chapter 4, HSDPA incorporates other spectral efficiency enhancements over DCH or DSCH based schemes.

From a TCP point of view, one of the most relevant effects of HSDPA is that the introduction of H-ARQ adds another layer performing retransmissions to cope with the unsuccessful transmissions over the radio channel. Figure 6.1 plots the retransmission mechanisms available for TCP traffic in HSDPA. Note that the radio link transmissions errors can still be recovered at RLC and TCP level. However, since the H-ARQ retransmissions are executed at physical layer level (i.e. from the NodeB) the overall retransmission delay is lower than the RLC retransmissions (i.e. performed from the RNC). Moreover, the usage of soft combining and incremental redundancy further improves the retransmission performance over conventional ARQ mechanisms. For example, in [74], with a simple link adaptation algorithm, it is described that the uncorrected BLER after 3 transmissions is only 0.8% with a BLER for first transmission of 66% for Vehicular A at 3 km/h and a G Factor of 0 dB. Such efficient ARQ mechanism contributes to hide the wireless unreliability to TCP, and thus, decreases the probability of triggering TCP flow control mechanisms such as the fast retransmission and recovery or spurious TCP time outs [64] that degrade the TCP performance.

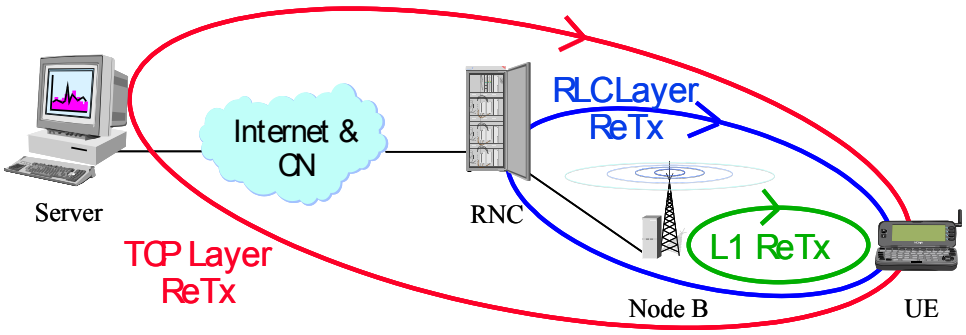


Figure 6.1: Retransmission Recovery Mechanisms from 3 Different Layers for TCP Traffic in HSDPA.

6.2.1 The Node B Queueing and the Maximum Number of RLC Retransmissions

The present section aims at giving a general overview of the interaction between the RLC layer and Node B queuing in HSDPA. Note that it is not the target of this section to assess this interaction in quantitative terms, but rather to extract some conclusions for further analysis of the TCP traffic on HSDPA.

In HSDPA, the RLC layer functionality remains intact from the Release'99 [26], and it is also located in the RNC of UTRAN. However, in HSDPA the MAC-hs layer of the Node B keeps the main buffering of the data to be forwarded to every UE in the cell.

On the one hand, too little amount of buffered user data may have a negative impact when exploiting the multiuser diversity of the radio channel because the user under favourable radio conditions could run out of data. The time required by the flow control functionality of the Iub interface to refill the data buffer of the user is not expected to be shorter than the coherence time of the channel (~ 27 ms at 3km/h according to [92]).

On the other hand, very large amount of Node B buffering raises the TCP segment queuing delay at the Node B. In TCP, as the slow start or the congestion avoidance algorithms increase the congestion window over the bandwidth delay product, the data excess is cumulated in the buffer of the limiting link of the connection (i.e. in the case of HSDPA it corresponds to the Node B). Assuming that the radio interface is the unique link limiting the bandwidth of the TCP connection, the average Node B buffering associated to the i-th user can be expressed in steady state as:

$$B_i = cwnd_i - \lambda_i \cdot RTT_i \qquad i = 1,..,N \tag{6.1}$$

where $cwnd_i$ denotes the congestion window of the TCP connection of user i, λ_i is the throughput of user i, and RTT_i is the overall round trip time of a TCP segment (except the Node B queuing). See in [PI 6.1] the definition of B_i the average Node B buffering for user i. Then, the average queuing delay suffered by the TCP segments of user i in the Node B can be described as:

$$D_i = \frac{B_i}{\lambda_i} = {cwnd_i}\Big/{\lambda_i} - RTT_i \qquad i = 1,..,N \tag{6.2}$$

where D_i represents the average Node B queuing delay per user of the TCP segments. See definition of the average Node B queueing delay per user in [PI 6.2].

Table 6.1 evaluates the average Node B queuing delay per user for various user throughputs, round trip times, and TCP receiver buffers. In the computation it is assumed that the congestion window has grown up to the upper limit imposed by the receiver buffer size. Note that some combinations of the user throughput and the round trip times of the table may not reach the bandwidth delay product (empty cells of the table). The results from Table 6.1 indicate that the average Node B queuing may be significant (in the order of seconds) when the congestion window reaches the receiver window size.

User Thr.	**RTT**	**TCP Receiver Buffer**			
		8 kbytes	**16 kbytes**	**32 kbytes**	**64 kbytes**
564kbps	150 ms	0.85	1.85	3.85	7.85
	300 ms	0.7	1.7	3.7	7.7
128kbps	150 ms	0.35	0.85	1.85	3.85
	300 ms	0.2	0.7	1.7	3.7
256kbps	150 ms	0.1	0.35	0.85	1.85
	300 ms	-	0.2	0.7	1.7
512kbps	150 ms	-	0.1	0.35	0.85
	300 ms	-	-	0.2	0.7
1024kbps	150 ms	-	-	0.1	0.35
	300 ms	-	-	-	0.2

Table 6.1: Average Node B Queuing Delay (sec) per User for Various User Throughputs and Round Trip Times.

The TCP segments to be retransmitted at RLC level experience an increased overall round trip time due to the RLC retransmission process. Every RLC PDU retransmission sums up a delay equal to the layer 2 RTT, which in HSDPA includes the Node B queuing, which as shown in Table 6.1 can be significantly high. See in Figure 6.2 the RLC Retransmission procedure and Node B queuing delay addition to the layer 2 RTT in HSDPA.

Striving to retransmit an unsuccessful RLC PDU lifts the overall round trip time and will ultimately trigger a TCP time out if the RLC retransmission process persists. Note that the RLC acknowledged mode usually applies in sequence delivery, which implies that not only the unsuccessfully decoded PDU is not forwarded to the higher layer, but also any other RLC PDU with lower sequence number. Hence, it can be concluded that there exists a trade off between the amount of data to be queued in the Node B, and the number of RLC retransmissions if the probability of triggering a TCP time out is to be kept below a certain margin. Note that the conclusion of Chapter 3, stating that the RLC level correction mechanisms are preferable than the TCP ones, might still hold in HSDPA (with significantly higher layer 2 overall round trip time).

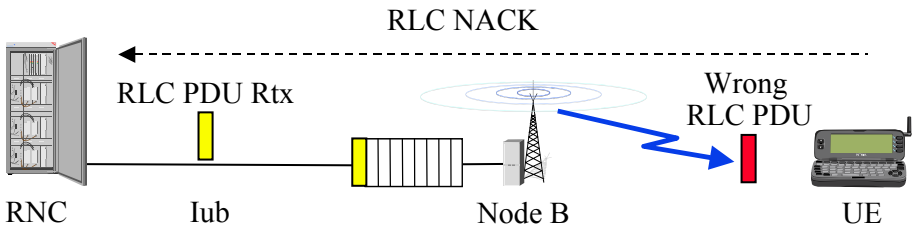


Figure 6.2: RLC Retransmission Procedure And Node B Queuing Delay Addition to the Link Layer RTT in HSDPA.

Besides the TCP protocol, the flow control of the Iub interface can regulate the amount of data buffered in the Node B, and therefore reduce it to minimize the impact of the Node B queuing in the link layer round trip time. However, a minimum level of data buffering is required for the Node B to be able to fully exploit the multiuser diversity of the time-shared channel. On the other hand, as already explained in section 6.2, the introduction of the HARQ protocol at physical layer, which provides an efficient retransmission mechanism with a low SDU error rate, suggests that the need for error correction from the RLC layer are considerably lower than for a DCH.

6.3 Interaction Between System Throughput, User Throughput and Average Packet Queuing Delay

The overall performance of TCP traffic over HSDPA networks completely depends on various factors such as the round trip time of the connection, the segment loss probability, the TCP flavour (Reno, Tahoe, …), etc. However, one of the most relevant factors of the performance is directly related to the performance of the HSDPA network itself, i.e. the data bit rate offered to the TCP connection, whose main determining factors are the Packet Scheduling strategy, the resources allocated to HSDPA, and the cell load.

The flow control of TCP aims at offering the input user load to allow full-speed transmission while avoiding large delays and congestion at the network nodes crossed by the TCP traffic. As explained in Chapter 3, the transmitting side controls the offered load that can be injected into the network by means of the congestion window whose size is determined by the slow start and congestion avoidance algorithms. Additionally, the TCP receiver buffer limits the amount of data the receiver is able to accept without saturating its memory. Since the server does not have an a priori knowledge of the bandwidth delay product, the selection of the congestion window size represents a basic trade-off between the user throughput and the packet delay at the limiting link of the connection.

During the periods when the congestion window is lower than the bandwidth delay product $cwnd < BW \cdot RTT$, the user throughput can not grow above:

$$\lambda_i \leq \frac{cwnd_i}{RTT_i} \qquad i = 1,..,N \tag{6.3}$$

which imposes an upper limit on the activity and thus on link utilization of the user. Then, the average user link utilization can not increase above:

$$\rho_i = \lambda_i \Big/ \overline{R_i} \leq \frac{cwnd_i}{rtt_i \cdot \overline{R_i}} \qquad i = 1,..,N \tag{6.4}$$

where $\overline{R_i}$ represents the average supportable data rate of user *i*, and ρ_i denotes the user link utilization. Note that equation (6.4) has assumed a blind scheduling algorithm that ignores the instantaneous channel quality of the user, and hence the user throughput is linearly proportional to the average proportion of time the user holds the HS-DSCH channel ρ_i.

Table 6.2 computes the upper limit of the average user link utilization for various average user data rates, round trip times, and TCP receiver buffers. The computation has assumed that the congestion window has grown up to the upper limit imposed by the receiver buffer size. The results clearly show that as the data rate of the user increases, the upper limit of the average user link utilization decreases. For example for an average supportable data rate of 2 Mbps, a receiver buffer of 8 kbytes, and a round trip time of 300 ms, the offered load by the TCP protocol is not large enough for the user to hold the channel for more 10% of the time. At this stage, the average Node B queuing delay experienced by the user is expected to be minor. A low link utilization could be the result of a low traffic activity, which could potentially impact the overall HS-DSCH cell throughput.

$\overline{R}$	RTT	TCP Receiver Buffer			
		8 kbytes	16 kbytes	32 kbytes	64 kbytes
512kbps	150 ms	83	-	-	-
	300 ms	42	83	-	-
1024kbps	150 ms	42	83	-	-
	300 ms	21	42	83	-
2048kbps	150 ms	21	42	83	-
	300 ms	10	21	42	83

Table 6.2: Upper Limit of the Average Link Utilization (%) per User for Various User Throughputs and Round Trip Times Assuming that the Congestion is Lower than Bandwidth Delay Product.

On the other hand, during the periods the congestion window is larger than the bandwidth delay product $cwnd > BW \cdot RTT$, the user throughput is simply limited by the amount of time resources ρ_i assigned to user *i*, which depends on the scheduling strategy and the load of the system:

$$\lambda_i = \rho_i \cdot \overline{R_i} \qquad i = 1,..,N \tag{6.5}$$

where again a blind scheduler has been assumed.

The congestion window growth over the bandwidth delay product does not increase the user throughput but on the contrary causes the TCP segments to queue in the buffer of the limiting link (i.e. Node B), thereby increasing the queuing delay experienced by the data flow, which has the very negative effects described in section 6.2.1.

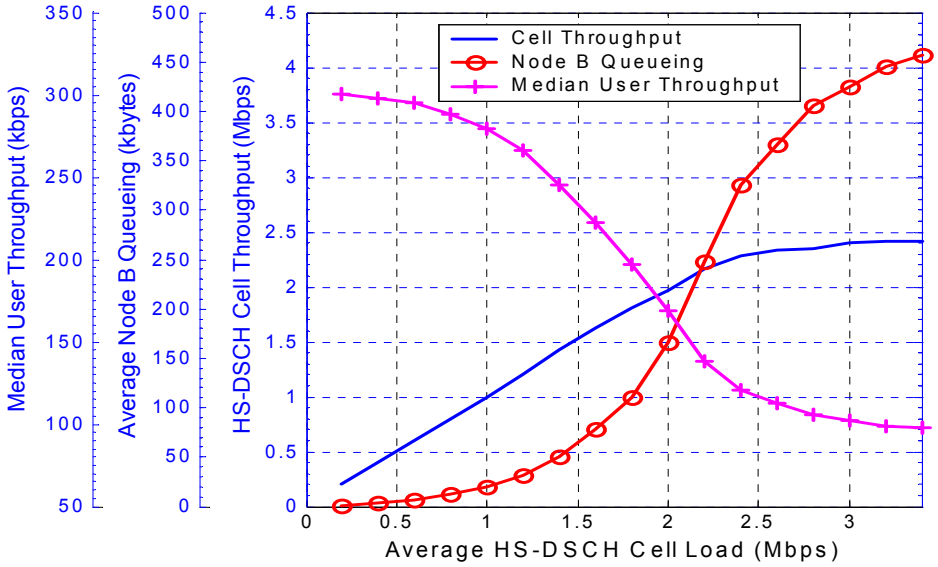


Figure 6.3: Median User Throughput, Average Node B Queuing and HS-DSCH Cell Capacity as a function of the Offered HS-DSCH Cell Load for Proportional Fair Scheduler in Pedestrian A, 3km/h. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

The user throughput, the average Node B queuing, and the system throughput are factors very closely interrelated. Figure 6.3 plots the three metrics of interest (HS-DSCH cell throughput, average Node B queuing, and median user throughput) as a function of the offered cell load for Proportional Fair scheduler in Pedestrian A at 3km/h. See in [PI 6.3] and [PI 6.4] the definitions of the average Node B queueing and the median user throughput.

For low offered loads, the system underachieves its full capacity potential. The few users in the system have available a large amount of resources, and hence their bandwidth delay product is so high that the time required by the slow start algorithm to adapt the congestion window to it causes significant penalty in the user throughput. Moreover, the maximum user throughput is clipped by the receiver's window size and the round trip time (32kbytes and 300 ms respectively for the simulated case), yielding a instantaneous user throughput upper bound of Rx Window/RTT ≈ 855 kbps (note that the slow start further limits the user throughput). Regarding the average Node B queuing, the high user throughputs cause that the users do not store large amounts of data in the buffer (see equation (6.1)), which combined with the low average number of users in the system yields a very low Node B queuing.

As the load increases in the range ∈ [1-2.3] Mbps, the system throughput follows linearly the load increase. The number of users in the system raises and the user throughput starts to be limited by the maximum share of the channel the scheduler assigns to every user. It is interesting to observe that at this offered load range, the median user throughput decreases quasi-linearly. However, the Node B queuing has an exponential like growth in the aforementioned load region due to the build up of data in the user's buffers during the periods the users are waiting to be served. Unfortunately, this exponential like behaviour is directly reflected in the average queuing delay per TCP segment in the Node B through the Little's theorem:

$$D = \frac{B}{\lambda} \tag{6.6}$$

where D represents the average queuing delay per TCP segment, B denotes the average Node B queuing and λ is the average HS-DSCH cell throughput. See in [PI 6.5] the definition of the average queuing delay D.

The results of Table 6.3 present the average Node B queuing delay per TCP segment in Pedestrian A at 3km/h. The results for the Proportional Fair scheduler clearly show the explosion of the average queuing delay per TCP packet when the load comes closer to the maximum capacity. A load increase from 1.4 to 1.8 Mbps yields a cell throughput growth of 29% but an average packet delay increase of 72% (from 0.286 to 0.493 seconds). Further increase of the offered load from 1.8 to 2.2 Mbps yields a cell throughput benefit of 20% (from 1.8 to 2.16 Mbps) but an average TCP segment delay of 85% (from 0.493 to 0.912 seconds).

Round Robin	Offered Cell Load (Mbps)	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
	Cell Capacity (Mbps)	0.2	0.4	0.6	0.8	0.97	1.12	1.15	1.16
	Avg. Packet Delay (ms)	65	110	202	424	1067	2860	3522	3853
Proportional Fair	Offered Cell Load (Mbps)	0.2	0.6	1.0	1.4	1.8	2.2	2.6	3.0
	Cell Capacity (Mbps)	0.2	0.6	1.0	1.4	1.8	2.17	2.34	2.40
	Avg. Packet Delay (ms)	54	91	154	286	493	912	1251	1413

Table 6.3: Average Node B Queuing Delay (sec) per Packet and Cell Throughput for Round Robin and Proportional Fair in Pedestrian A, 3km/h. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

The cell throughput saturates at around 2.4 Mbps, and the increase of the offered load beyond this load operating point hardly raises the cell throughput and only produces the undesired call blocking in the system. Since the average number of users in the cell hardly grows beyond this point, the average Node B queuing also shows a saturation trend reaching around 460 kbytes for the highest load simulated point. Note that maximum Node B buffering capabilities are set to 1.25 Mbytes. In this overload situation, the user throughput reaches a lower bound whose median is around 100 kbps, and the cdf at 10% and 90% outages are 50 and 140 kbps respectively.

Figure 6.4 depicts the HS-DSCH cell capacity, the average Node B queuing and the median user throughput for the Round Robin scheduler. The results show a similar general trend as for the Proportional Fair scheduler, though the scheduling efficiency of the fast scheduling algorithm obviously extends its dynamic range of cell capacity much further than that of Round Robin, whose capacity saturates at around 1.15 Mbps. Regarding the median user throughput, it can be seen that for very low offered loads (up to 0.6 Mbps), there is hardly any difference with the results obtained by the Proportional Fair scheduler because the slow start algorithm is the one imposing the limit on the maximum achievable user throughput. However, in overload (offered loads larger than 1.4 Mbps), the median user throughput reduces to around 50 kbps, which is about half of what the Proportional Fair algorithm provides under such an overload situation.

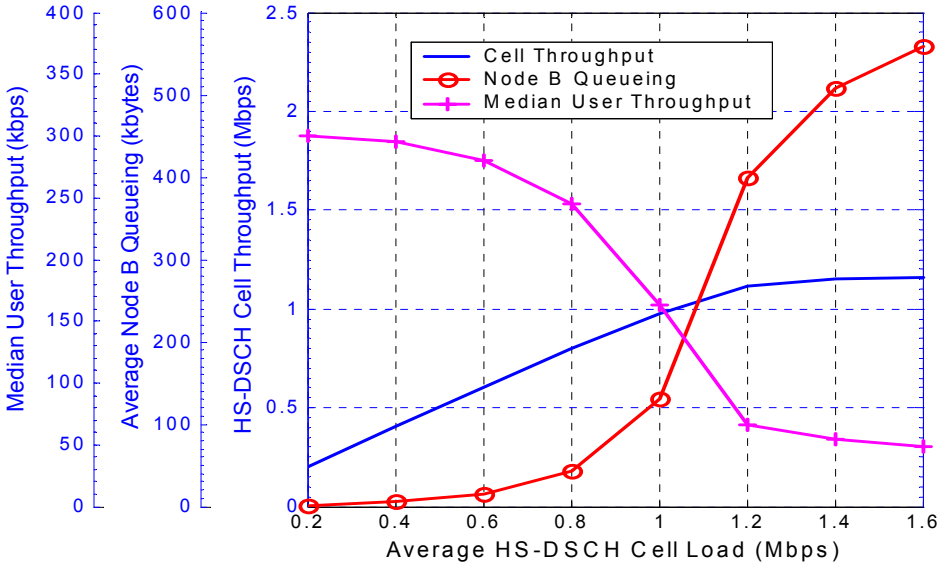


Figure 6.4: Median User Throughput, Average Node B Queuing and HS-DSCH Cell Capacity as a function of the Offered HS-DSCH Cell Load for Round Robin Scheduler in Pedestrian A, 3km/h. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

Regarding the system queuing the system also presents an exponential like behaviour, but the final average Node B queuing in overload is larger than for the Proportional Fair case reaching up to around 560 kbytes due to the lower user throughput. Moreover, again due to the lower user throughput, the queued data more harmful causing significant average packet queuing delay. A cell capacity increase from from 0.8 to 0.97Mbps (21%) causes an average TCP segment delay increase of 152% (from 424 to 1067msec), and for a cell throughput growth from 0.97 to 1.12Mbps (15%) raises the average packet delay increases 168% (from 1067 to 2860 ms). See Table 6.3 for more information about the average packet queuing delays for Round Robin.

Figure 6.5 draws the mean number of alive users in the cell versus the offered HS-DSCH cell load. A user is defined to be alive when it occupies one associated DPCH channel, i.e. when the user has been admitted in the system regardless of whether he has or not data queued in the Node B to transmit. See definition of the mean number of alive users in [PI 6.6].

The resulting curve for the Round Robin scheduler resembles an exponential build up of the average number of alive users. This type of performance is present in typical queuing systems when the overall system utilization factor ρ (i.e. the long-term proportion of time the time-shared channel is busy) approaches one (e.g. the M/M/1 queuing system [82]). In HSDPA, this exponential like growth is obviously limited by the maximum number of associated DPCHs available in the system (i.e. blocking).

For very low offered loads, with such a few users in the system, there is hardly any difference between both scheduling methods. However, the performance of the Proportional Fair scheduler has significant differences with the Round Robin case when load rises. For the same offered cell load, the scheduling efficiency of the fast scheduling method causes that users transmit their packet call swiftly, and hence disappear from the system faster, which results in a decrease of the average number of users compared to Round Robin. But perhaps the most

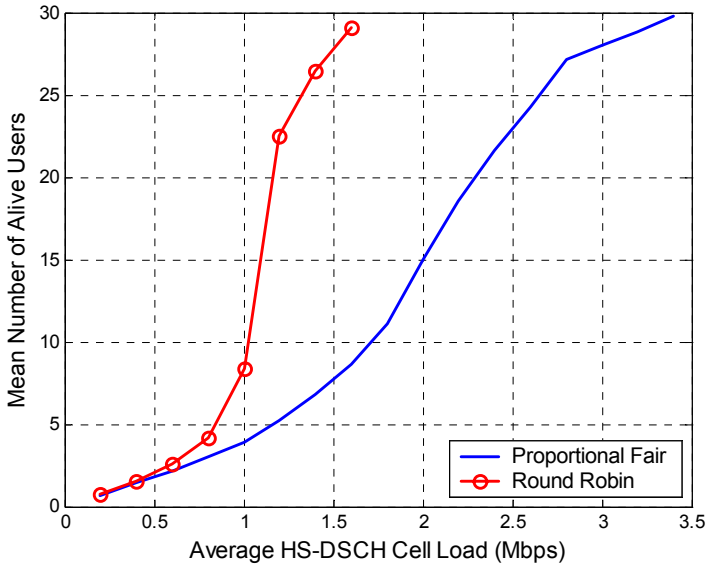


Figure 6.5: Mean Number of Alive Users in the Cell as a function of the Offered HS-DSCH Cell Load for Round Robin and Proportional Fair Schedulers in Pedestrian A, 3km/h. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

remarkable difference is that the growth of the average number of users does not match the exponential like behaviour. The main reason behind it is the non linear mapping between the amount of time resources a user is assigned and his achievable throughput (see Figure 5.3). In any case, the blocking also limits the performance of the Proportional Fair scheduler when the overall system utilization factor tends to one.

6.3.1 Distribution of the Queuing Delay in the Cell

Section 6.3 has given a general description of the interaction between the system throughput and the average Node B queuing, and its implications in terms of average packet queuing delay. It has been concluded that when the system is loaded close to its maximum capacity the average Node B queuing delay explodes while obtaining minor benefits in terms of system throughput.

With the Proportional Fair and Round Robin schedulers the Node B queuing is not uniformly distributed among all the users in the system. Figure 6.6 plots the average Node B queuing plus transmission delay per user as a function of the G Factor for both scheduling algorithms in Pedestrian A at 3km/h. The metric depicted in Figure 6.6 has been measured from the time instant a TCP segment reaches the Node B and until the last bit of the packet is correctly received in the UE. The simulations have been conducted under such an offered load that the blocking equals 0.2% and 0.4% for the Round Robin and Proportional Fair schedulers respectively. The average TCP segment queuing delay D_i is expected to dominate over the TCP segment transmission delay when the user link utilization ρ_i is low (as occurs for the simulated offered cell loads in Figure 6.6).

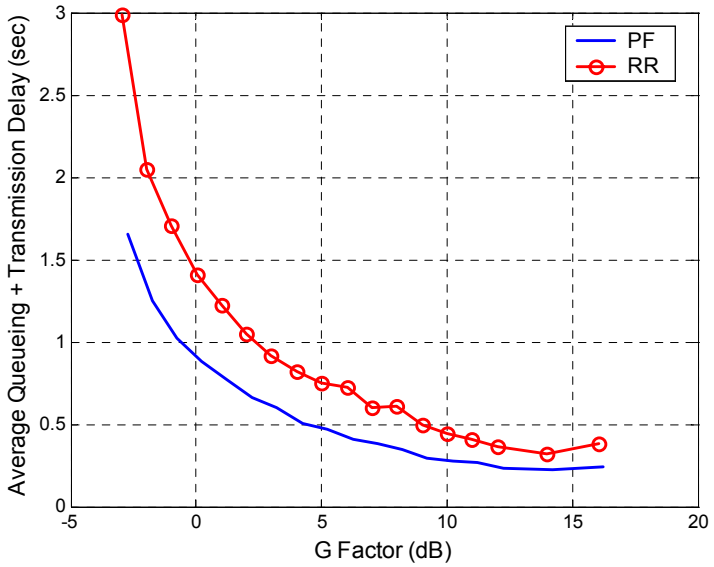


Figure 6.6: Average Queuing Plus Transmission Delay per User as a function of the G Factor for Round Robin and Proportional Fair Schedulers in Pedestrian A, 3km/h. Offered Cell Load equals 1.0 and 2.0 Mbps for the Fast and Slow Scheduler Respectively. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

The results show the growth of the average packet queuing delay per user as the G Factor decreases, which is determined by the inversely proportional relation between the queuing delay and the average user throughput as presented by equation (6.2).

Additionally, the non uniform distribution of the cell throughput (unless a fair throughput scheduling algorithm is executed), causes that not all users weight evenly in the overall Node B queuing. It can be easily demonstrated that the user's contribution to any system level metric (such as the temporal average Node B queuing B) depends on the percentage of time they exist in the system:

$$B = \frac{1}{T}\int_0^T B(t)\cdot dt = \sum_{\forall i \in S} B_i \cdot \frac{T_i}{T} \qquad (6.7)$$

where T denotes the duration of the observation period (e.g. simulation length), T_i is the period user i was alive, B_i is the temporal average queuing of user i, and S represents the set of all users that were alive during the period $(0,T)$. Then, users with lower average throughput, or users with larger packet call sizes (note that the simulations have been conducted with constant packet call size between all the users), who remain in the system for longer tend to contribute more heavily to the overall Node B queuing.

It can be concluded that with the Proportional Fair and Round Robin scheduling strategies, the users with worse average radio conditions are the ones suffering from a larger queuing delay and contributing most to the average Node B queuing due to their longer existence in the system.

6.4 Impact of the TCP Flow Control Mechanisms on the System Performance

As explained in Chapter 3, the main purpose of the flow control of the TCP transport protocol is to offer as much traffic as to fully utilize the resources of the limiting link of the connection (and therefore achieve the maximum possible transmission rate) while avoiding congestion in the nodes of the network that would cause segment losses due to buffer overflow. The ability of the TCP transmitting entity to operate within this trade off depends on numerous factors such as the round trip time, the probability of segment losses, the degree of congestion in the network, the size of the file to be transmitted (note that longer files permit a better convergence of the flow control algorithms), the variations of the limiting link capacity, transmission disruptions during cell hand over procedure, etc. It is of particular interest how these factors impact not only the user throughput, but also the HSDPA system throughput. As described in Chapter 5, the multiuser diversity gain of the HS-DSCH is very sensible to the diversity order (especially in Pedestrian A). This section focuses on evaluating how the TCP traffic inactivity (burstiness) influences the multiuser diversity order and ultimately the HSDPA cell capacity.

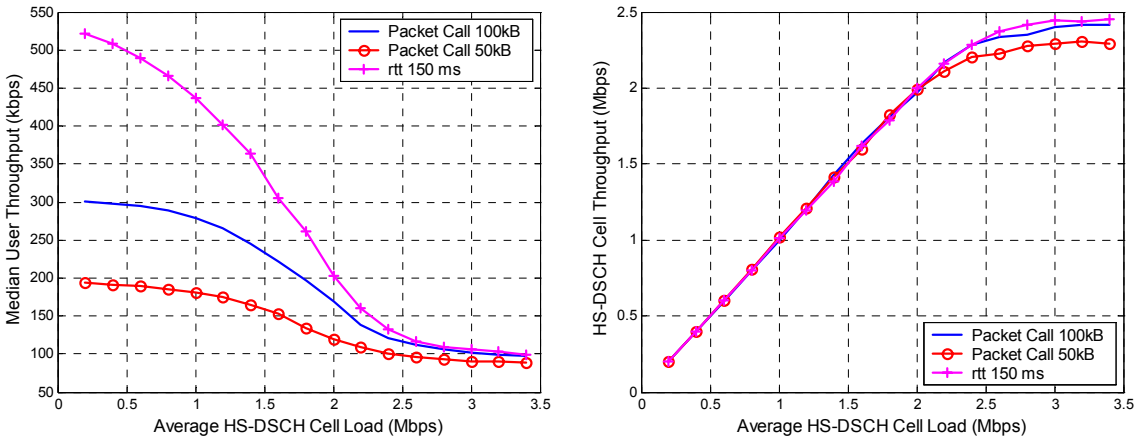
For simplicity reasons, here it will only be analysed the impact of the round trip time and the packet call size on the system throughput and the user throughput. The simulations for section 6.3 have been conducted with a packet call size of 100 kbytes and a round trip time of 300 milliseconds. Two new set ups considered here are described in Table 6.4. The factor *rtt* represents the overall round trip time of a TCP segment except the Node B queuing and radio transmission delay.

Notation	Rtt	Packet Call Size
Packet Call 100 kB (*)	300 ms	100 kbytes
Packet Call 50 kB	300 ms	50 kbytes
rtt 150 ms	150 ms	100 kbytes

Table 6.4: Simulation Set-Ups Considered in the Analysis of the Effect of the Round Trip Time and Packet Call Size on the User Throughput and the HSDPA Cell Capacity. The Set Up (*) is the Same Used in Section 6.3.

Figure 6.7 depicts the user throughput and the HS-DSCH cell throughput for the three simulation set-ups described in Table 6.4 for the Proportional Fair scheduler. The results show how the flow control of TCP hardly impacts the system throughput. For a packet call of 50 kbytes the HS-DSCH cell throughput is reduced by around 5% compared to the case with packet call size of 100kbytes. At high load, the large number of users in the cell reduces the throughput allocated to every user, and thereby the bandwidth delay product of their TCP connections. The resulting traffic inactivity is low and hardly affects the multiuser diversity of the HS-DSCH. However, the application performance strongly depends on the round trip time or packet call size when the system operates at low or medium loads as it can be seen in Figure 6.7 a).

Table 6.5 summarizes the cell throughput results and average packet delay for the two new set-ups described in Table 6.4 for Round Robin and Proportional Fair schedulers. See how the



a) *b)*

Figure 6.7: a) Median User Throughput, and b) HS-DSCH Cell Capacity as a function of the Offered HS-DSCH Cell Load for Proportional Fair Scheduler in Pedestrian A, 3km/h. HSDPA Power ≈ 50% of Total Node B Power & 7Multi-codes.

Round Robin scheduler has a marginal degradation of the cell capacity (from 1.16 to 1.14 Mbps) in overload when the packet call size is reduced to from 100 kbytes down to 50 kbytes.

The time-shared nature of the HS-DSCH impacts the variance of the overall round trip time experienced by the TCP segments, which might have negative impacts on the protocol performance (spurious time outs) [93]. This effect has not been analysed in the present study, and it is for further investigation.

Round Robin	Offered Load (Mbps)		0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
	Packet Call 50 kbytes	Cell Capacity (Mbps)	0.2	0.4	0.6	0.8	1.0	1.11	1.13	1.14
		Avg. Packet Delay (ms)	37	60	105	231	723	1605	1944	2085
	rtt 150 ms	Cell Capacity (Mbps)	0.2	0.4	0.6	0.8	1.0	1.12	1.15	1.16
		Avg. Packet Delay (ms)	85	136	218	411	1171	3083	3727	4043
Proportional Fair	Offered Load (Mbps)		0.2	0.6	1.0	1.4	1.8	2.2	2.6	3.0
	Packet Call 50 kbytes	Cell Capacity (Mbps)	0.2	0.6	1.0	1.4	1.8	2.11	2.23	2.29
		Avg. Packet Delay (ms)	33	53	94	168	330	494	620	676
	rtt 150 ms	Cell Capacity (Mbps)	0.2	0.6	1.0	1.4	1.8	2.15	2.36	2.44
		Avg. Packet Delay (msec)	72	115	184	295	529	983	1454	1626

Table 6.5: Average Node B Queuing Delay (sec) per Packet and Cell Throughput for Round Robin and Proportional Fair Scheduler in Pedestrian A, 3km/h for the Two Set-Ups Described in Table 6.4.

6.5 Conclusions

The present chapter has analysed the system level performance of TCP based NRT traffic on HSDPA. It has been shown that there exists a strong interaction between end user thoughput, system throughput, and packet queuing delay. The flow control functionality of the TCP

protocol heavily participates in the aforementioned interaction and could easily influence the optimum operation point.

If high user throughputs are to be provided (by lowering the offered cell load operating point), special attention is to be paid to the potential sources of end user throughput limitation introduced by TCP (such as the slow start, potential TCP segment losses, receiver buffer capabilities, etc). For the Proportional Fair scheduler, under the traffic model assumptions described in the section B.4 of the Appendix B, it has been shown that the decrease of the offered load below 1Mbps yields a median user throughput no larger than 420 kbps (in Pedestrian A at 3km/h, for the Proportional Fair scheduler, with a packet call of 100 kbytes and a round trip time *rtt* of 300 milliseconds).

For heavy system loads, the cell throughput increases accordingly up to the maximum cell capacity, although the increased number of users limits the proportion of time every user holds the time-shared channel, and hence, restricts the achievable user throughput beyond the limits imposed by TCP. Under such conditions, the Node B queuing exhibits an exponential like growth with the load causing significant packet queuing delay at the Node B. For example, under the previous assumptions, an offered load increase from 1.4 to 1.8 Mbps leads to a cell throughput benefit of 29%, but an average TCP packet queuing delay of 72%, and an offered load increase from 1.8 to 2.2 Mbps yields a cell throughput growth of 20% but an average segment queuing delay of 85%. Furthermore, for similar blocking level, the Proportional Fair scheduler shows a significant reduction of the queuing delay compared to Round Robin due to its higher data rates. At around 0.2% blocking Round Robin has an average queuing delay of around 1.07 seconds, while for 0.4% blocking Proportional Fair only causes an average segment queuing delay of around 625 milliseconds.

It has also been shown that the average segment queuing delay is not distributed uniformly among all the users in the cell for the Proportional Fair and Round Robin schedulers. As depicted in Figure 6.6, for an offered load of 1 Mbps with the Round Robin scheduler, a –3 dB G Factor user suffers an average segment queuing delay of 3 seconds, while a 15 dB G Factor user only experiences around 350 milliseconds of average queuing delay.

High levels of Node B queuing are undesirable because they raise the layer 2 round trip time, which may cause spurious TCP time outs during the RLC layer retransmission process. It has been concluded that there exists a trade off between the amount of data queued in the Node B and the number of RLC retransmissions if the probability of triggering spurious TCP time outs is to be kept below a certain margin.

Finally, it has been described that the time-shared nature of the HS-DSCH channel is suitable to cope with the burstiness of TCP traffic. At high loads, the large number of users in the cell reduces the throughput allocated to every user. The consequence is that the number of active users (multiuser diversity order) is hardly affected, and therefore the cell capacity reduction is minor (5% for a packet call size reduction from 100 to 50 kbytes in Pedestrian A at 3km/h).

Chapter 7

Streaming Services On HSDPA

7.1 Introduction

Multimedia streaming services are becoming increasingly popular among Internet users and are expected to play an important role in the evolution of packet data services in Third Generation (3G) mobile networks. These services impose stringent constraints on the conveying networks in order to satisfy their QoS requirements. In UMTS, dedicated channels can provide a feasible solution to convey streaming bearers and fulfil their QoS demands as long as the channel bandwidth allocation is sufficient. On the other hand, the increased spectral efficiency of High Speed Downlink Packet Access (HSDPA) over dedicated channels [94] motivates the mapping of streaming services on to this technology. The time-shared nature of the transport channel of HSDPA is very well suited for Non Real Time (NRT) bursty traffic such as interactive and background, but the transport of streaming traffic over HSDPA faces important challenges in order guarantee its QoS requirements. Among those challenges, it must be addressed a packet scheduling policy capable of exploiting the multiuser diversity while guaranteeing the exigent delay jitter requirements. Similarly, an admission control algorithm is required to maintain the system load within the margin where the QoS can be guaranteed. The satisfaction of the QoS requirements has to be ensured during the handover process, etc. Moreover, it is unclear if the impact of those QoS demands is too costly in terms of spectral efficiency.

The literature regarding the conveyance of streaming services on 3G networks is not as extensive as the related to interactive and background traffic. In [30], Montes evaluated the performance of streaming services over EGPRS networks. In [95], Schieder also investigated the performance of streaming bearers for GERAN comparing to the NRT traffic bearers. However, little attention has been paid to the support of streaming services on UMTS with dedicated channels. Regarding the HSDPA related literature, most of it concentrates on best effort traffic too. In [96], Furuskär analysed the performance of HSDPA for both TCP and streaming traffic. For the streaming case Round Robin and Max CI schedulers were considered, and the author claimed that the latter outperforms Round Robin by 40%. The author also remarked that a relaxation of the delay jitter constraints can be directly translated on a cell capacity increase. In [97], Moulsley studied the performance of various scheduling algorithms for data streaming applications on HSDPA, and already stated that the scheduling decision should be based on both the packet age and the instantaneous channel quality.

The present chapter aims at addressing some of the previously mentioned challenges that have not been solved by the currently available literature. The chapter starts debating the RLC operating mode for streaming services in HSDPA, and describing the discard timer parameter. Then, assuming that the radio interface is the limiting link in the satisfaction of the QoS attributes, different scheduling policies for streaming services in HSDPA are discussed, and the performance of the most relevant ones is assessed for constant bit rate streaming flows by means of system level simulations. The impact of the delay jitter constrains (through the discard timer) on the system performance is extensively treated. The spectral efficiency results for streaming services on HSDPA are compared to the NRT traffic ones obtained in the previous chapter[15].

The outline is as follows: section 7.2 discusses the RLC operating mode for streaming services in HSDPA. Section 7.3 introduces the discard timer parameter, and the section 7.4 discusses the Packet Scheduling strategies for streaming flows in HSDPA. Section 7.5 assesses the HSDPA cell capacity for CBR flows, and section 7.6 compares those results to the NRT traffic capacity. Finally, section 7.7 draws the concluding remarks.

7.2 RLC Mode of Operation for Streaming Services in HSDPA

According to [26], HSDPA functional split keeps the PDCP and RLC layers intact from the Release'99. That means that RLC can operate in either acknowledged or unacknowledged mode, but not in transparent mode due to ciphering [26].

The RLC acknowledged mode strives to maintain a low SDU Error Ratio at expenses of a variable delay by retransmitting every unsuccessfully received RLC PDU (up to a maximum number of transmissions determined by the MaxDAT parameter). Every RLC PDU retransmission sums up a delay equal to the layer 2 RTT, which in HSDPA includes the Node B queuing (note that this queuing could reach up to a few seconds, see section 7.5). Hence, the RLC retransmissions raise the transfer delay (note that the larger the delay jitter introduced by the network, the larger the buffering capabilities for jitter removal at the end

[15] Note that the present chapter does not consider the effect of hand over in the performance of streaming services over HSDPA.

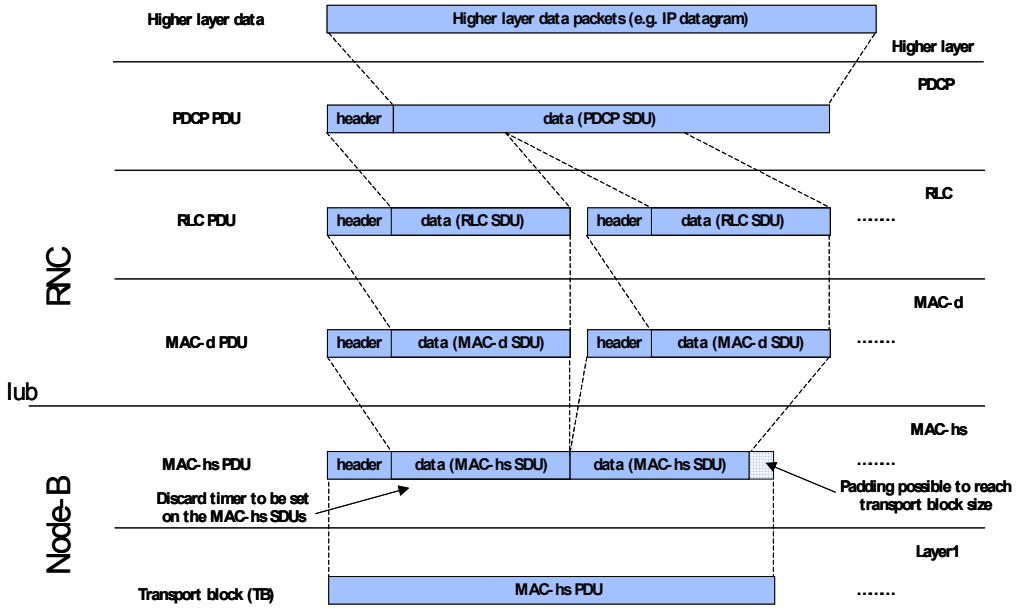


Figure 7.1: Multimedia Data Flow in HSDPA.

terminal). Moreover, the in sequence delivery applied to the Dedicated Traffic Channel (DTCH) further exacerbates the IP datagram delay jitter. Thereby, it is concluded that, if the residual error rate of the H-ARQ process is low enough to reach a tolerable bearer SDU error ratio, the RLC retransmissions of this operating mode are not convenient for streaming services in HSDPA.

Thus, in the rest of this thesis, under the assumption that the HSDPA link adaptation algorithm and the H-ARQ protocol provide a sufficiently low residual error rate at the physical layer, it will be further assumed that the RLC operating mode for streaming services on HSDPA is unacknowledged.

7.3 Data Flow And Discard Timer In HSDPA

Section 2.3.2 gave a general overview of the end-to-end transmission of multimedia data services. The present section concentrates on the flow of such multimedia streams in HSDPA, and describes the discard timer parameter.

Figure 7.1 plots the multimedia data flow in HSDPA. As mentioned in section 7.2, the HSDPA functional split keeps the PDCP and RLC layers intact from the Release'99 [26]. Once in UTRAN, and after the header compression of the PDCP layer and the segmentation (or concatenation) of the RLC layer, the MAC-d layer becomes responsible for handling the data flow. The flow control function of the Iub interface will rule the transfer of MAC-d PDUs from the RNC to the Node B. According to [33], the Node B rules this flow control of the Iub interface by sending capacity allocation messages to the RNC.

After being transferred through the Iub interface, the MAC-d PDUs reach the MAC-hs layer at the Node B. In order to limit the maximum queuing of the MAC-d PDUs, a QoS related

timer is set on those PDUs (see section 5.2). This timer is denominated *discard timer*[16]. A MAC-d PDU is queued in the Node B until the Packet Scheduler serves it, or the discard timer expires (which implies the discarding of that PDU). As the discard timer sets the Node B queuing limit, it must be included in the UMTS delay jitter budget in order to meet the transfer delay QoS attribute. Note that the RNC does not have any related queuing timer that limits the maximum RNC queuing delay. Hence, in order to keep this source of delay under control, the flow control must forward the MAC-d PDUs to the Node B rapidly (as long as the Node B buffering capabilities allow so).

7.4 Packet Scheduling for Streaming Services in HSDPA

Let define *elastic traffic* as the one that is able to modify its data transfer rate according to the available bandwidth within the network (such as TCP traffic) [98]. Strictly speaking, streaming traffic should be considered as elastic traffic since streaming servers have nowadays the possibility to probe the network and adapt the encoding process so that the bit rate of the resulting encoded data stream is adjusted to the network capacity. However, the margin of bit rate variability of streaming traffic is significantly reduced compared to that of interactive or background traffic (due to the TCP protocol [34]). The degree of elasticity of the traffic directly relates to the degree of fairness to be provided to the users in the cell, and therefore must be seriously considered in the Packet Scheduling policy. Due to the significantly lower elasticity margin of the streaming traffic (compared to TCP traffic), henceforth, the present study will treat streaming traffic as fully non-elastic.

As described in Chapter 5, the Proportional Fair algorithm is a very suitable scheduling option in HSDPA for such elastic traffic. However, a more advanced scheduling algorithm is needed to be able to guarantee Quality of Service over a shared wireless link. A scheduling algorithm will successfully provide service in HSDPA to streaming flows only if it is able to meet the transfer delay and the GBR attributes while keeping the SDU error ratio below a certain threshold manageable for the application (e.g. 5%, 10%, 1%).

Provision of QoS guarantees over a wireless shared link with continuously varying capacity and frequent channel degradations is not an easy task. Probably due to the non-elasticity property of real time traffic, most authors in the currently available bibliography address the challenge by targeting at providing throughput fairness (in a max-min manner) to all the users in the cell [99]-[86]. This way, users under worse average radio conditions (low G Factor), with lower potential throughput capabilities, could experience an improved transmission delay that contributes to guarantee their QoS.

An extensive part of the related literature attempts to modify the existing scheduling algorithms from wired networks to wireless ones. This section presents an overview of the Packet Scheduling algorithms that could successfully provide service in HSDPA.

[16] Note that here it is not treated the on going discussion within the 3GPP between the discard timer and the T1 reordering timer.

In [99], Lu proposes an interesting adaptation of the Fluid Fair Queuing (FFQ) algorithm to wireless shared links. The FFQ algorithm is a wired link scheduling policy that treats each packet flow as a fluid flow and serves every non-empty queue with a rate equal to [100]:

$$\frac{\phi_i}{\sum_{\forall j}\phi_j}\cdot C \tag{7.1}$$

where ϕ_i represents the weight applied to the flow i, and C denotes the total link capacity. The above definition is applicable for both channels with constant and time-varying capacity [99].

In [99], Lu models the radio link as an error prone channel, and identifies that a flow may be unable to transmit due to channel error. To accommodate those channel errors, the channel allocation is swapped between users over short time windows. However, the channel allocation swapping cause that some flows have a queue length shorter than its corresponding FFQ service (denominated *leading flows*), while other flows cumulate a queue length longer than their corresponding FFQ service (denominated *lagging flows*). The key feature of Lu's wireless FFQ algorithm is to allow lagging flows to be compensated at expenses of leading flows.

Lu further deals with the implementation aspects of his wireless FFQ algorithm in [99]. However, besides the fact that the radio link is simply treated as an error prone channel with single channel rate, the algorithm suffers from an inherent tight coupling between the throughput and delay guarantees, which is not interesting for HSDPA[17] since that tight coupling does not fully exploit the multi-user diversity that provides significant capacity improvement in HSDPA.

In [81], Rhee takes into account the dynamic rate transmission of 3G evolution systems, and combines the Proportional Fair algorithm with the Weighted Fair Queuing (WFQ) algorithm. The WFQ is another wired link algorithm that emulates the FFQ policy but treats the user data not as a fluid (infinitesimally divisible) but rather as packets with finite granularity. As described in [100], the WFQ algorithm, among all the backlogged flows at a given instant, it serves the flow whose packet would complete service first in the corresponding FFQ system if no additional packets were to arrive after that instant. Indeed, the algorithm in [81] has strong similarities to [99], since it monitors the order of backlogged packets for the WFQ algorithm and serves the lagging flow whose WFQ priority is highest. However, the main difference with [99] is that a lagging flow is only served if it belongs to a so-called Transmission Candidate Set (TCS). The Transmission Candidate Set comprises all the users whose relative channel quality is higher than a given threshold.

$$TCS = \left\{ i : \frac{R_i(t)}{\lambda_i(t)} \geq \gamma \right\} \tag{7.2}$$

where $R_i(t)$ represents the Channel Quality Indicator (CQI) or supportable data rate of user i at instant t, $\lambda_i(t)$ denotes the average throughput of user I, and γ is a certain threshold.

[17] Note that in HSDPA the MAC-d PDUs can be queued in the Node B for a certain period (discard timer), which enables the decoupling between throughput and delay guarantees.

Although this algorithm can nicely provide a high degree of fairness between users under different average radio channel quality conditions, it does not seem interesting for streaming services for two reasons: (i) although the multiuser diversity is partially taken into account through the TCS, it also suffers from the coupling of throughput and delay guarantees as in [99]. (ii) it disregards the user's packet delay since it may schedule a certain data flow to compensate for its average throughput while another flow may keep packets close to its due delay.

In [86], Barriac and Holtzman propose to give a video data flow a delay dependent priority. For users with a delay below a certain threshold (e.g. 60% of the deadline) the priority is computed according to the Proportional Fair algorithm, while for users with a delay larger than the threshold, the priority of the Proportional Fair algorithm is elegantly modified to equalizing the users' throughput:

$$P_i = \begin{cases} \dfrac{R_i(t)}{\lambda_i(t)} & , D_i < \tau \\ \dfrac{R_i(t)}{\lambda_i(t)} \cdot \left[\dfrac{\max_j \left\{ \overline{R_j(t)} \right\}}{\overline{R_i(t)}} \right] & , D_i \geq \tau \end{cases} \tag{7.3}$$

where P_i describes the priority of user i, $R_i(t)$ represents the Channel Quality Indicator (CQI) or supportable data rate of user i at instant t, $\overline{R_i}$ is the average CQI report of user i, $\max_j \left\{ \overline{R_j} \right\}$ indicates the maximum average CQI report from all the users, $\lambda_i(t)$ represents the throughput of user i up to instant t, D_i indicates the head of line packet delay of user i, and τ represents the aforementioned threshold (e.g. $\tau = 0.6 \cdot T$ being T the packet due time). The head of line packet delay is the queuing delay experienced by the packet that currently occupies the front of the queue.

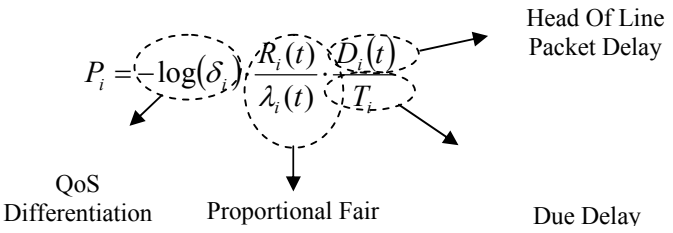
This algorithm is efficient at on average increasing the priority of low G factor users with delays larger than the described threshold because the $\overline{R_i}$ term in the denominator of (7.3) compensates their priority (initially reduced in the Proportional Fair algorithm due to the lower average supportable data rate). However, (7.3) does not emphatically prioritise flows with delays close to their due time, which seems to be interesting to avoid (or reduce) the packet discarding in the Node B.

In [101], Andrews proposes an algorithm to keep the probability of the packet delay exceeding the due time below the SDU error ratio:

$$\Pr\{D_i > T_i\} < \delta_i \tag{7.4}$$

where D_i indicates the head of line packet delay of user i, T_i represents the due time, and δ_i is the percentage of discarded packets that exceed their due time (SDU error ratio).

The so-called Modified Largest Weighted Delay First (M-LWDF) computes the priority of a given user as [101]:



(7.5)

Note that the term $D_i(t)/T_i$ ranges from zero to one, becoming closer to the latter when the head of line packet delay approaches the due delay. This algorithm does not only take advantage of the multiuser diversity available in the shared channel through the Proportional Fair algorithm, but also increases the priority of flows with head of line packets close to their deadline violation. This way, the algorithm can control the flow delays to provide QoS for real time traffic. The algorithm also adds a QoS differentiation term to imbalance the priority between users with different application demands on error rate. Henceforward, it will be assumed that all the users in the cell have the same application requirements, and therefore, this QoS term will be disregarded. In [101], the author further states that the M-LWDF algorithm is *throughput optimal*, meaning that it is able to keep all queues stable if this is at all feasible to do with any other algorithm.

However, this solution does not control the user throughput, which, without any further rate control, is fully determined by the input traffic load of the user and the SDU error ratio[18]. This implies that data flows, for example VBR video streams, temporarily raising their bit rate over the GBR would subsequently increase their cell resources consumption without any restriction, which could cause the starvation of the remaining flows in the cell. Andrews further proposes in [101] another method to provide minimum throughput guarantees for real time flows, although no solution is described to combine the packet delay control with the minimum throughput guarantees. Nevertheless, a combination of the packet delay control given by equation (7.5) and a token bucket algorithm that guarantees the scheduling to traffic up to the GBR may possibly solve the paradigm. In HSDPA, the rate control is done by means of flow control through the Iub interface, which is the one to ensure that all bearers are guaranteed service up to the GBR, and therefore should be in charge of implementing the rate control algorithm.

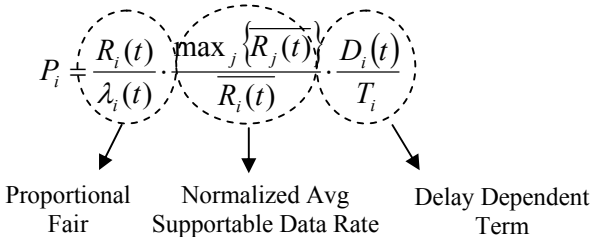
As previously commented, most of the literature that address the paradigm of providing QoS for real time users in wireless time shared links, base the solution on distributing the cell throughput evenly between the users in the cell [99]- [86]. Due to the non-elasticity property of the real time flows, the M-LWDF algorithm can manage to provide approximately the same average throughput for all the users in the cell. However, from equation (7.5) it can be concluded that the M-LWDF algorithm is rather unfair, because two data flows with the same head of line packet delay will be assigned different priorities if their achievable data rates are different. As it will be shown in Figure 7.4 and Figure 7.6, this leads to a lack of queuing delay fairness in favour of users with high transmission rates at expenses of the low G Factor users, whose packets will be delayed in the Node B and ultimately discarded if they exceed the discard timer.

In order to attempt to reduce the overall probability of RTP packet discarding in the Node B, the large queuing of low G Factor users could be compensated with the low delay of more

[18] Note that Barriac's algorithm suffers from similar lack of rate control.

favourable users. Here it is proposed to combine the modification of the Proportional Fair algorithm proposed by Barriac [86], and the priority term depending on the head of line packet delay of the M-LWDF algorithm. Note that, in equilibrium, Barriac's algorithm tends to equalize the average user throughput regardless of the average achievable data rate of the user. The resulting priority would be computed as follows:

$$P_i = \frac{R_i(t)}{\lambda_i(t)} \cdot \frac{\max_j \{\overline{R_j(t)}\}}{\overline{R_i(t)}} \cdot \frac{D_i(t)}{T_i} \qquad (7.6)$$



where P_i describes the priority of user i, $R_i(t)$ represents the CQI of user i at instant t, $\overline{R_i}$ is the average CQI report of user i or supportable data rate, $\max_j \{\overline{R_j}\}$ indicates the maximum average CQI report from all the users, $\lambda_i(t)$ represents the average throughput of user i up to instant t, D_i indicates the head of line packet delay of user i, and T_i represents the packet's due time.

In the section 7.5, the performance of the M-LWDF and its fair version proposed here (whose priority is given by equation (7.6)) will be analysed for CBR flows.

7.4.1 Average User Bit Rate for the M-LWDF Algorithm

In order to compute the user throughput for the denominator of equation (7.5) an averaging mechanism is required. The ARMA filtering described in section 5.3.2.1 will be used as the averaging procedure. Moreover, for consistency with Chapter 6, those users that do not have data queued for transmission, do not update their ARMA filter.

Note that the time constant T_c of the ARMA filter (equation (5.6)) is related to the period an individual data flow can starve without being served by the Packet Scheduler. Starvation periods longer than T_c will decrease the average user data rate and will therefore increase the priority in equation (7.5). Thus, high T_c values will improve the spectral efficiency because the scheduler can wait for longer until the varying channel reaches a constructive fade, but also increase the packet delay because the starvation period is longer. Although no results are given to back, here it is considered that the last term of equation (7.5) is in control of the packet delay, and therefore T_c should be selected long enough to improve the system performance (e.g. 1.5 seconds ≈ 10 λ at 3km/h).

Since the ARMA filter is not updated when the user's queue is empty, the result of equation (5.6) may not be necessarily equal to the user throughput. In order not to confuse them both, from here onwards, the result of equation (5.6) will be referred to as the Average M-LWDF Bit Rate.

7.5 System Level Performance of CBR Video over HSDPA

The present section evaluates the HSDPA system performance for video streaming services.

In section 2.4 and 2.5 it was introduced the two available techniques (CBR and VBR) to encode a video clip stream, the server possibilities to transmit the encoded video sequence (CBRP or VBRP transmission) and their implications in terms of either Predecoder buffer requirements or bearer capabilities to temporarily increase the allocated bandwidth.

As explained in section 2.4.2, the Dejitter buffer is to remove the variable delay introduced by various sources along the end-to-end connection. For a UMTS bearer mapped into HSDPA, the dominant source of delay in the total transfer delay is introduced in the MAC-hs layer of the Node B due to the queuing delay the RTP packets experience before being transmitted to the UE. In order to assess the contribution of the Node B queuing delay to the total transfer delay (and thus, the Dejitter buffer requirements at the UE) and its interaction with the HSDPA system capacity, the focus concentrates on the evaluation the Node B and MAC-hs layer performance by means of quasi-dynamic network level simulations.

In order to simplify the analysis, only CBR encoded video flows are considered. With this simplification, the influence of the temporary bit rate peaks is avoided, which will allow to decouple the impact of the flow delay guarantees from the minimum throughput guarantees. Hence, no rate control mechanism that guarantees the GBR attribute over the Iub interface is required in the simulation model. Moreover, this simplification allows to concentrate on the Dejitter buffer since the Predecoder buffer is not required at the UE for this type of encoding. Note, however, that the HSDPA performance results that will be presented here could still be applied to CBRP transmission flows if the UE supported the corresponding Predecoder buffer.

The investigated Packet Scheduling policy is the M-LWDF algorithm.

The overall modelling and default parameters for the network level assessment are described in the Appendix B. The propagation model, mobility, transmission power and receiver modelling is common with the one in previous two chapters (specified in section B.2), while the streaming traffic model, the architecture modelling, and the outage definition are detailed in section B.5.

7.5.1 Analysis of the Factors that Determine the Performance

7.5.1.1 Influence of the Dynamics on the Delay Jitter

In HSDPA, the guarantee of the packet delay constraint as expressed in equation (7.4), is influenced by the temporal variations of the time-shared channel capacity, whose transitory reduction will lead either to a packet discard in the MAC-hs if the discard timer is exceeded, or to a delay jitter increase otherwise. In the latter case, if the initial Dejitter buffering is large enough to accommodate the jitter increase, then the buffer will temporarily decrease its amount of stored data until the channel capacity recovers from its transitory reduction. But if the Dejitter buffer cannot accommodate the jitter increase, then the RTP packet will arrive to

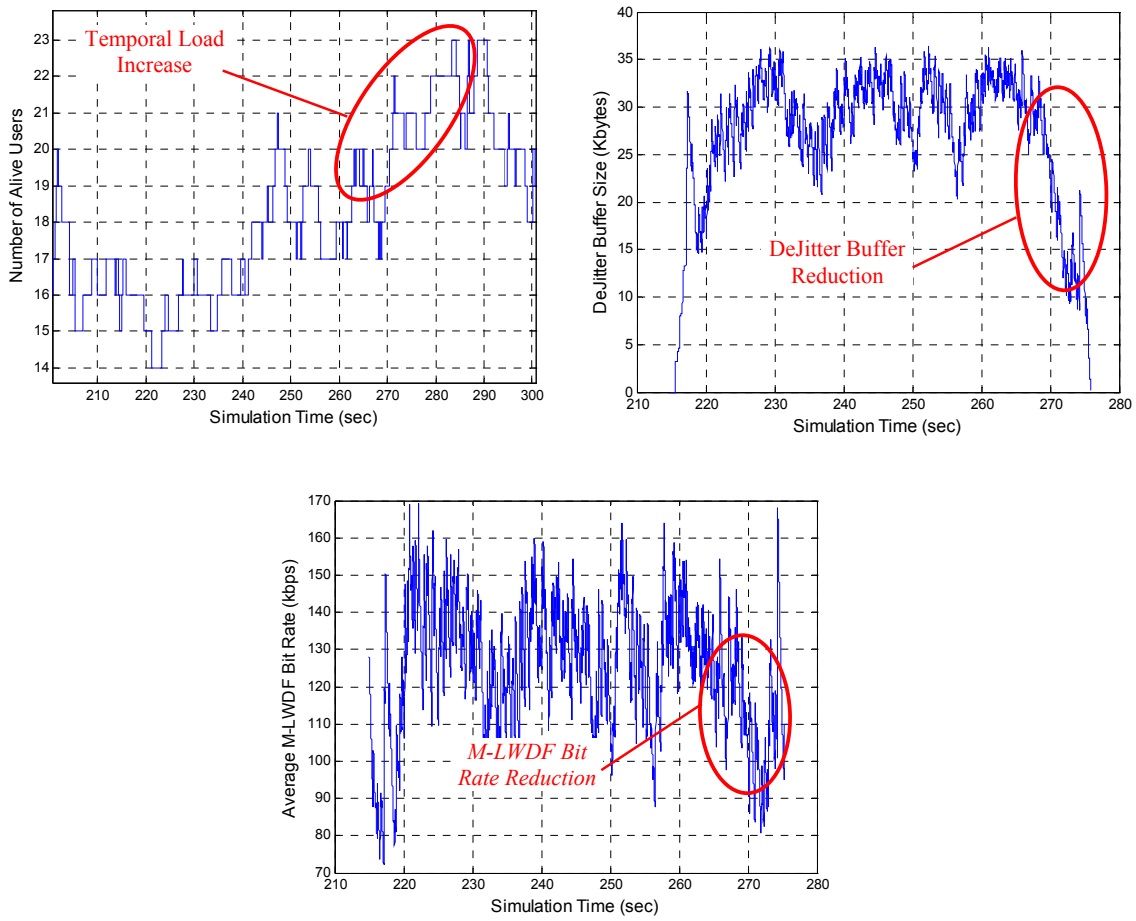


Figure 7.2: Instantaneous Number of Alive Users, Dejitter Buffer Size and Average M-LWDF Bit Rate for a UE with G Factor below -2dB, Average Source Bit Rate 128 kbps, Discard Timer 2 sec, and Initial Dejitter Buffering 1.8 sec.

the UE too late to be played out in time. Note that this very last case should be avoided by a proper dimensioning of the total delay budget, the discard timer as a part of the total budget, and the Dejitter buffer, since this packet transmission simply leads to a radio resource waste.

Short term capacity variations caused by fast fading of the channel quality of a certain UE only provokes small variations of its RTP packet queuing delay, which is subsequently translated into small variations of the amount of data stored in the Dejitter buffer compared to realistic amounts of initial Dejitter buffering (see Table B.3). However, either propagation loss increases (e.g. shadowing or path loss), or transitory cell load increases may cause capacity reductions that last for long periods, and therefore introduce long delay jitters which could ultimately lead to a massive discard of RTP packets in the Node B, and its subsequent Dejitter buffer starvation in the UE. Against this buffer underflow situation, the client will trigger a UE Dejitter re-buffering, although in the present system level evaluation this situation will be accounted for in the outage.

The impact of the short and long term capacity variations on a low G Factor user can be seen in Figure 7.2, that depicts the instantaneous cell load, the Dejitter buffer size and the Average M-LWDF bit rate for a user with G Factor below –2dB, with a CBR source bit rate of 128

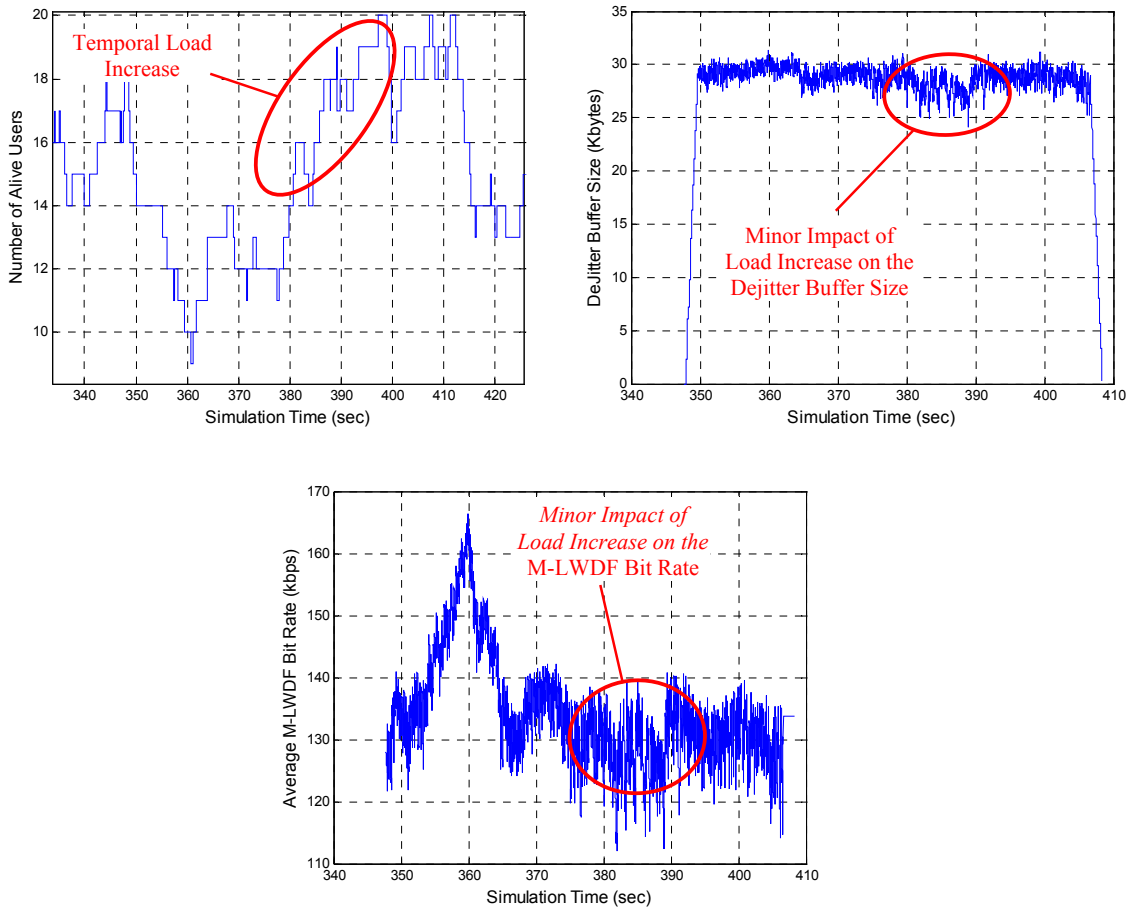
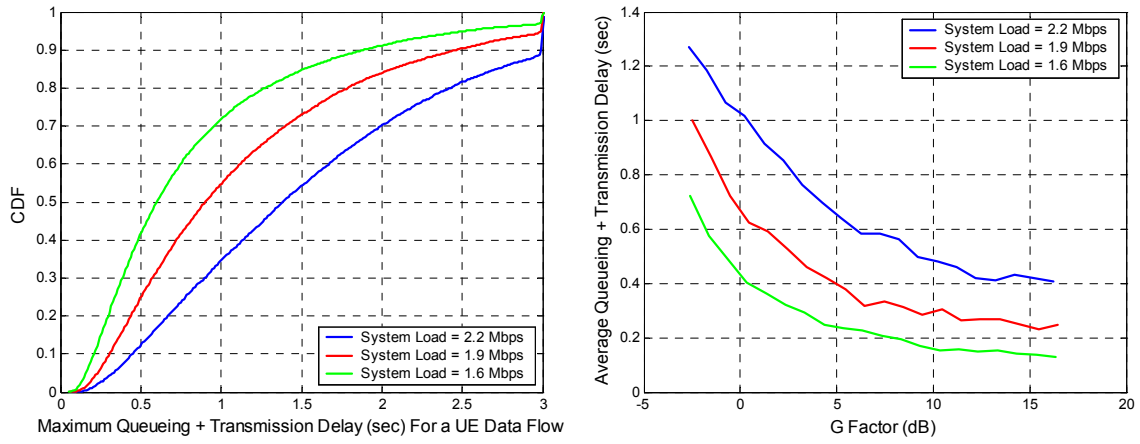


Figure 7.3: Cell Load, Dejitter Buffer Size and Average M-LWDF Bit Rate for a UE with G Factor over 12dB, Average Source Bit Rate 128 kbps, Discard Timer 2 sec, and Initial Dejitter Buffering 1.8 sec.

kbps, a discard timer of 2 sec, and initial Dejitter buffering of 1.8 sec. The cell load is measured in terms of instantaneous number of *alive users*. The definition of an alive user is given in [PI 6.6]. As shown in the pictures, short term effects, such as the fast fading, do not have a significant impact on the Dejitter buffer, however long term load increases, that reduce the resources available for each user, create a user throughput decrease, which significantly reduces the amount of data stored in the Dejitter buffer.

Due to the Proportional Fair term in the priority computation of the M-LWDF algorithm (equation (7.5)), under the same head of line packet delay, users with higher transmission rate possibilities (high G Factor) enjoy higher priority than users with worse average radio conditions (low G Factor users), and thus, similar long term capacity variations produce a minor impact on the delay jitter and the Dejitter buffer for those users. This minor impact can be seen in Figure 7.3, that draws the instantaneous cell load, the Dejitter buffer size and the Average M-LWDF bit rate for user with G Factor over 12dB, with a CBR source bit rate of 128 kbps, a discard timer of 2 sec, and initial Dejitter buffering of 1.8 sec. Besides the minor impact of a temporal load increase, Figure 7.3 also shows the behaviour of the Average M-LWDF Bit Rate from the ARMA filter (equation (5.6)), which significantly increases its value beyond the user throughput (128 kbps) during periods of low cell load because the filter is only updated for users with data queued for transmission.



a) *b)*

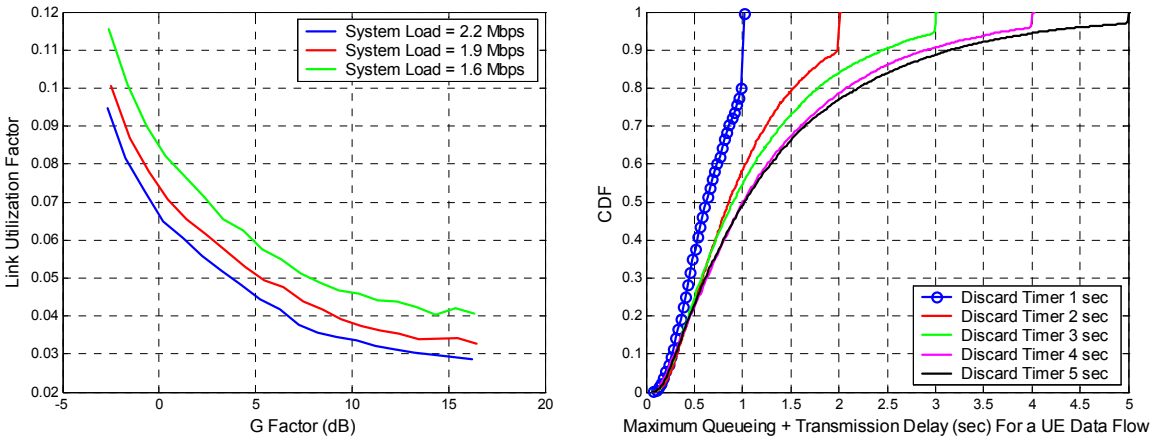
Figure 7.4: a) Maximum and b) Average Queuing Plus Transmission Delay For a UE Data Flow for a Discard Timer of 3 sec, and Average Source Bit Rate 128 kbps.

7.5.1.2 The Queueing Delay and the Discard Timer

As explained in section 7.3, once in the Node B, the MAC PDUs conveying the streaming data belonging to a video flow are queued in the MAC-hs layer waiting to be served by the Packet Scheduler. As known from data networks theory [82], the average input load, the system capacity, as well as the applied scheduling algorithm rule the average queuing delay of those PDUs, but, the statistical probability to delay them beyond a certain bound (i.e. the discard timer) strongly depends on the involved dynamics too (such as described in previous section).

Figure 7.4 considers the maximum queuing in the Node B plus the transmission delay of a RTP packet belonging to a UE data flow[19] (note that under the current system modelling the RTP packets are not further segmented into lower layer PDUs)[20]. Figure 7.4 a) represents its CDF for all the users in the central cell of the layout for different average loads. See in [PI 7.1] the definition of the maximum queuing plus transmission delay for a UE data flow. The figure shows that the maximum queuing delay ranges from a very low delay (close to zero) to the maximum discard timer delay, and the percentage of users whose maximum queuing delay reaches the discard timer increases with the average load. Figure 7.4 b) depicts the average queuing plus transmission delay versus the user's G Factor. See the definition of the average queuing plus transmission delay for a UE data flow in [PI 7.2]. The result has been averaged in bins of 1dB G factor. Figure 7.4 b) shows how the average queuing of the low G Factor users significantly overcomes the average queuing of high G Factor users. This is due to the lack of fairness introduced by the M-LWDF algorithm in favour of users with higher transmission rate probability. Thus, for high loads, and under this scheduling algorithm, the low G Factor users are the ones whose packet delay will reach the maximum delay (i.e. the discard timer) and determining the tolerable outage within the cell.

[19] RTP packets discarded are not accounted for in the queuing delay computations.
[20] Note as well that the queuing delay factor determines the metric depicted in Figure 7.4 since it dominates over the transmission delay.



a) *b)*

Figure 7.5: a) Link Utilization For a UE Data Flow for a Discard Timer of 3 sec, and Average Source Bit Rate 128 kbps. b) CDF of the Maximum and Average Queuing Plus Transmission Delay For a UE Data Flow, Input Load of 1.9 Mbps, and Average Source Bit Rate 128 kbps.

Although under the same head of line packet delay, users with higher G Factor enjoy higher priority than users with worse average radio conditions, low G Factor users still consume most of the cell resources to keep up the same average throughput as high G Factor users. This can be corroborated with the picture Figure 7.5 a). Note as well, how an average load increase reduces the amount of time resources to be shared among all the users, and thereby will impose tougher limitations for low G Factor users to maintain the required throughput. In Figure 7.5 b) it can be seen the CDF of the maximum queuing plus transmission delay for different discard timer values. The results show that, for a given cell load, the discard timer cuts the tails of the distribution of the maximum Node B queuing, and therefore the percentage of packets to be discarded in the Node B. Note that for the specified load (1.9 Mbps), when increasing the discard timer from one to two seconds, there is a significant reduction of the percentage of users that reach the discard timer (around 10%), whereas the reduction is minor when the discard timer is increased from four to five seconds.

Figure 7.6 shows a scatter plot with the percentage of RTP packets that are successfully played out in time by every user in a simulation with a discard timer of 3 seconds, an initial Dejitter buffering of 5 sec, an average cell load of 1.9 Mbps and an average source bit rate of 128 kbps. Since the initial Dejitter buffering is larger than the maximum jitter to be introduced by the Node B, the Node B buffering capabilities are sufficiently large, and the CN & Internet do not introduce any delay jitter in the current model, then it can be concluded that the UE initial buffering can remove all potential sources of delay variations to be introduced along the streaming connection. Hence, under this configuration, it can be further stated that the RTP packets not played successfully in time by the application are discarded in the Node B due to a packet delay larger than the discard timer. Then, from the Figure 7.6 it can be seen how, for the considered load, the percentage of discarded packets dramatically increases for low G Factors, whereas high G Factor users experience a good quality.

According to the outage criterion described in section B.5, all those users that do not play 95% of their RTP packets successfully in time are considered as not perceiving a good QoS, and will be included for the outage computation. Moreover, if the Node B discards more than

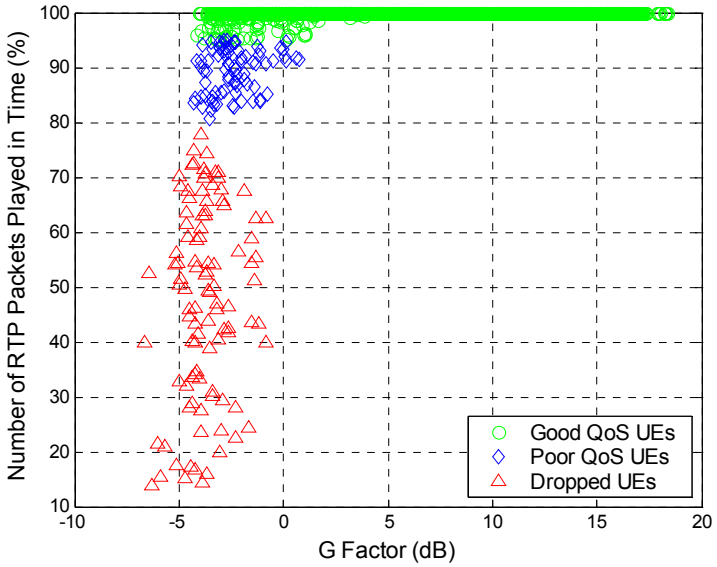


Figure 7.6: Percentage of RTP Packets Played Out Correctly In a Simulation with a Discard Timer of 3 sec, a Initial Dejitter Buffering of 5 sec, an Average Cell Load of 1.9 Mbps, and Average Source Bit Rate 128 kbps.

20% of the RTP packets, then the streaming connection is assumed to be so poor that is to be dropped by the network (and accounted for in the outage too).

7.5.2 Performance of 64 kbps CBR Video

In today's streaming applications, during the data contents transmission, the server probes the network capacity and adapts the encoded sequence bit rate (and therefore the encoded image quality) to the available bandwidth. In HSDPA, as explained in section 7.3, the flow control functionality is in charge of the rate control for the different streaming bearers mapped into HSDPA. The present study, will consider two CBR bit rate bearers for HSDPA: 64 and 128 kbps. In order to assess their suitability for the HSDPA technology, a network simulation with 64 kbps CBR bit rate bearers has been conducted with the results depicted in Table 7.1.

Avg. Cell Load (Mbps)	1.54	1.73	1.92	2.12	2.31
Avg. HS-DSCH Cell Throughput (Mbps)	1.46	1.59	1.65	1.70	1.75
Outage (%)	6.1	11.3	15.6	21.3	28.1
Blocking (%)	4.0	8.7	13.2	18.6	25.3
Dropping (%)	0.4	0.4	0.4	0.4	0.5

Table 7.1: Cell Load Performance for 64 kbps CBR Bearers in Ped_A 3km/h, Total Node B Buffering of 10 Mbits[21], Initial UE Buffering of 1.8 sec, Discard Timer 1 sec.

[21] Note that this Node B buffering is half than the default specified in Table B.3, but, nonetheless, it does not impose any buffering limitation for 64 kbps bearers.

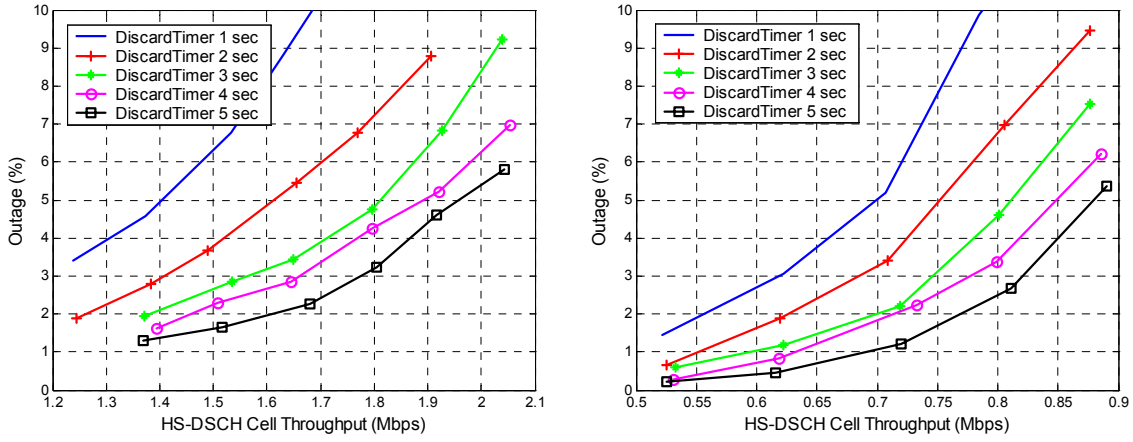


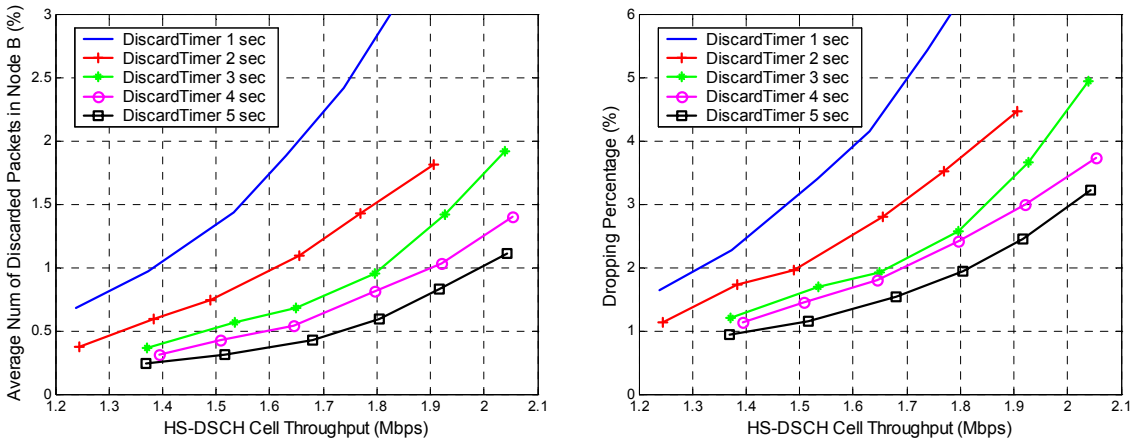
Figure 7.7: a) Cell Throughput Performance for 128 kbps CBR Bearers in Ped_A 3km/h and b) Veh_A 3km/h, Total Node B Buffering of 20 Mbits, Initial UE Buffering of 5 sec.

Table 7.1 depicts the HSDPA cell throughput and the associated outage for different input cell loads. As it can be seen, the cell load increase yields a cell throughput increase but at expenses of a QoS degradation. For the simulated cases, the obtained outage (indication of the user's experienced QoS) is dominated by the associated blocking caused by the limitation of the maximum number of associated DPCHs (i.e. 32), while the dropping and the percentage of users with a poor QoS represents a minor fraction of the outage. Note that the upper bound of the achievable cell throughput is 2.05 Mbps (i.e. the bearer bit rate multiplied by the maximum number of users to be mapped in the HS-DSCH). Due to the non-elasticity property of all the streaming bearers in the cell, which implies a fair throughput distribution among the users (in a max-min manner), users that experience a good average channel quality can not exploit their favourable radio conditions and raise their throughput beyond the 64 kbps, leading to a situation where the cell throughput is limited (imbalance between associated DPCH code requirements and HS-DSCH power allocation).

There are two potential actions that could be taken to avoid the blocking in the cell. On the one hand, the maximum number of associated DPCHs could be increased (e.g. to 48 or 64) allowing a higher input load with a lower blocking. Alternatively, the amount of power resources assigned to HS-DSCH could be decreased, allocating those resources to other purposes (e.g. DCHs) and therefore avoiding wasting them. In either case, the solution incurs a code and power overhead from the associated DPCHs, and the SCCH larger than for 128 kbps bearers.

7.5.3 Performance of 128 kbps CBR Video in PedA and VehA

In order to assess the performance of 128 kbps CBR video bearers on HSDPA in Pedestrian A and Vehicular A environments, extensive quasi-dynamic network level simulations have been carried out for different cell input loads and Node B discard timers. The main performance metrics are, for a given input load, the HSDPA cell throughput and the associated outage, which are the ones to determine the cell performance for the aforementioned bearers. Note that the lower the operating outage level, the larger percentage of users must fulfil the QoS criterion defined in section B.5, and thus, the further is the operating point from the maximum



a) *b)*

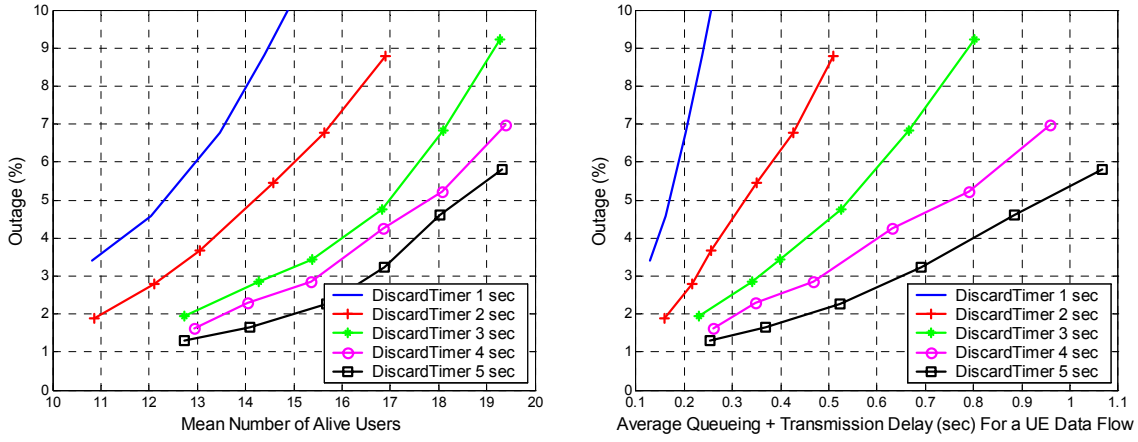
Figure 7.8: a) Percentage of RTP Packets Discarded in the Node B and b)Dropping Rate for 128 kbps CBR Bearers in Ped_A 3km/h , Total Node B Buffering of 20 Mbits, Initial UE Buffering of 5 sec.

capacity. The outage level of interest is considered to be 5%. See section B.4 for discussion about the outage criterion.

Figure 7.7 shows the exponential-like behaviour of the outage versus the cell throughput, so that the higher system throughput, and therefore the more users in the system, the less time resources are available for every user, and therefore less users are able to fulfil the QoS criterion. Note how the discard timer changes the slope of the outage exponential increase. At the considered outage level (5%), there is a cell throughput gain of 35% by increasing the discard timer from 1 to 4 seconds in Pedestrian A environment, while the same increase of the discard timer yields a capacity gain of 21% in Vehicular A environment. Note how there is a meaningful cell capacity gain by raising the discard timer from 1 to 2 seconds, while a discard timer rise from 4 to 5 seconds does not yield a significant cell capacity improvement. This was already expected according to Figure 7.5.

In terms of the power delay profile, the time dispersion of the Vehicular A model yields a cell throughput reduction of 55% versus the Pedestrian A model (for a discard timer of 3 seconds), which clearly indicates the capacity degradation due to the interpath interference that severely impairs the achievable data rate of both high and low G Factor users.

Let analyse the reasons that cause the outage in the present case. Unlike the case of 64 kbps video bearers, the current case (128 kbps bearers) incurs a negligible percentage of blocking due to the reduction on the average number of users. The initial Dejitter buffering is sufficient to remove all the delay introduced by the Node B queuing. Moreover during the system level simulations, the Dejitter buffer did never overflow because the initial buffering of 60% of the total UE buffering capabilities was a pretty conservative one. Since the current model assumes a constant delay in the Internet and CN, then, the sources of the QoS degradation come from the discarding of RTP packets in the Node B due to the lack of cell resources, which in extreme case turns into the dropping of the user. This can be corroborated in Figure 7.8 that depicts the percentage of total RTP packets discarded in the Node B and the dropping rate. See the definition of the definition of the average number of discarded packets in [PI 7.3]. The similarity between both pictures is due to fact that the dropping criterion is based on the percentage of RTP packets discarded in the Node B.



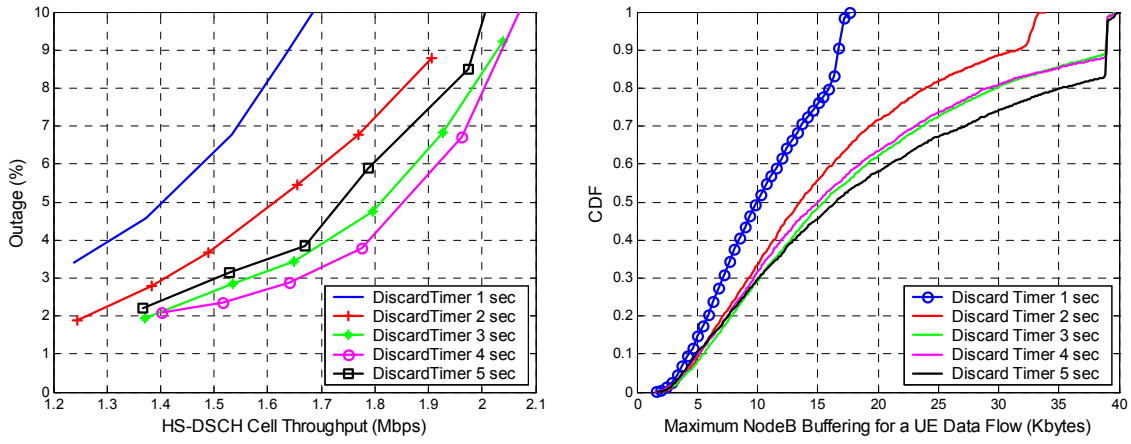
a) *b)*

Figure 7.9: a) Mean Number of Alive Users and b) Overall Average Queuing + Transmission Delay for 128 kbps CBR Bearers in Ped_A 3km/h, Total Node B Buffering of 20 Mbits, Initial UE Buffering of 5 sec.

Hence, in macrocell Pedestrian A environments, it can be concluded that the rise of the bearer bit rate from 64 to 128 kbps avoids the blocking (for a maximum number of DPCHs of 32), and balances the power and code resource requirements in the cell, which turns to provide a higher spectral efficiency for the HSDPA system (for the given maximum number of DPCHs). However, this conclusion does not to hold for Vehicular A environments, where the time dispersion of the channel leads the system to be power (and not code) limited even for 64 kbps bearers.

Figure 7.9 shows the average number of alive users mapped into the HS-DSCH. The comparison with Figure 7.7 indicates the linear relation between this metric and the system throughput, where the scaling factor is the average user throughput. Note that this average user throughput is slightly lower than the encoded video bit rate (128 kbps) because the discard timer gives a few additional seconds to the full retrieval.

Figure 7.9 also plots the overall average Node B queuing and transmission delay for a UE data flow for the simulated discard timers. See the definition of this performance indicator in [PI 7.4]. It is worthy to note that the average queuing delays are significantly lower than the maximum delay (determined by the discard timer). For example, for a discard timer of 1 second, the outage rapidly raises when the average queuing and transmission delay of the data flow exceeds 0.2 seconds. It can be concluded that the average need of Node B buffering capabilities for every data flow are much lower than the maximum ones. This would suggest that the hard split of the Node B memory among the UE data flows could be substituted by a dynamic Node B buffering management where the flows that do not utilise their allocated memory would release it in favour of the remaining flows. This dynamic memory management would lead to a more efficient usage of the Node B buffering capabilities and would reduce the overall Node B memory requirements. Nonetheless, for simplicity purposes of the system model, the following sections will still assume that the Node B buffering capabilities are hard split among the HSDPA users.



a) *b)*

Figure 7.10: a)Cell Throughput Performance and b)Maximum Node B Buffering for Input Load of 1.9 Mbps for 128 kbps CBR Bearers in Ped_A 3km/h, Total Node B Buffering of 10 Mbits, Initial UE Buffering of 5 sec.

7.5.4 Performance of 128 kbps CBR Video With Node B Buffering and UE Buffering Limitations

The present section aims at understanding the effects introduced by potential Node B or UE buffering limitations.

Firstly, a simulation was carried out halving the overall Node B buffering capabilities to 10 Mbits, which, under the current assumption of hard split among all the HSDPA users, yields a 39 kbytes of Node B buffering capabilities for every data flow. Note that for 128 kbps constant bit rate flows, the 39 kbytes of memory permit to queue the data in the Node B for 2.44 seconds without exceeding the maximum buffering size. However, if the data is queued in the Node B for longer (i.e. the discard timer is larger than the aforementioned 2.44 seconds), the flow control of the Iub interface will not always be able to transfer the data from the RNC to the Node B, and the data will be stored in the RNC until the Node B partially empties the data flow queue. The resulting effect can be seen on Figure 7.10 b) that plots the maximum Node B queuing for a UE data flow when the overall Node B memory is limited to 10 Mbits. See the definition of the maximum Node B buffering for a UE data flow in [PI 7.5]. As it can be seen in the figure, for discard timer values of 3, 4, and 5 seconds, a certain percentage of data flows hit the maximum buffering capabilities.

This RNC queuing introduces further uncontrolled delay variations and may raise the overall transfer delay beyond the maximum jitter that can be removed by the Dejitter buffering. However, as shown in Figure 7.10 a), the impact on the user's perceived quality (i.e. on the outage level) is negligible for discard timers below 3 or 4 seconds, and can only be significant for larger discard timers (such as 5 seconds).

Secondly, a simulation was carried out to analyse the effect of limited terminal memory capabilities. More specifically, the total Dejitter buffering capabilities are reduced to 5 seconds and the initial Dejitter buffering equals 60% of the total value (i.e. 3 seconds or 48 kbytes for a 128 kbps bearer). These buffering capabilities reduction impose tougher

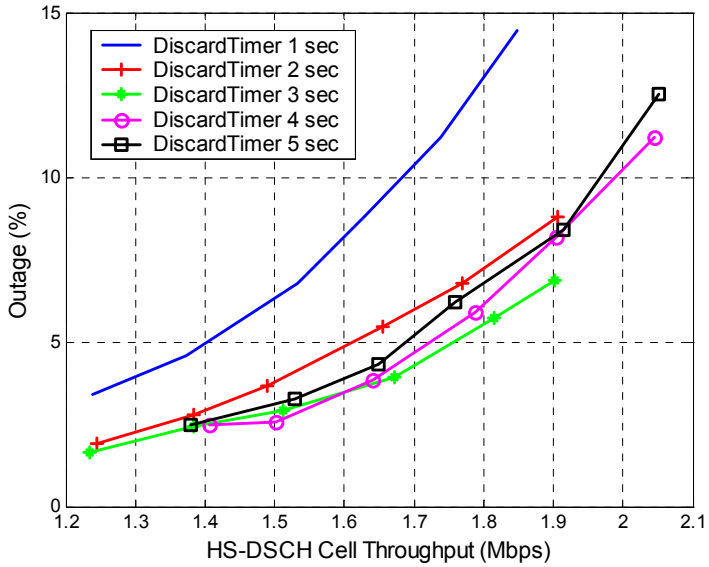


Figure 7.11: Cell Throughput Performance For 128 kbps CBR Bearers in Ped_A 3km/h, Total Node B Buffering of 20 Mbits, Total Dejitter Buffering of 5 sec, Initial UE Buffering of 3 sec.

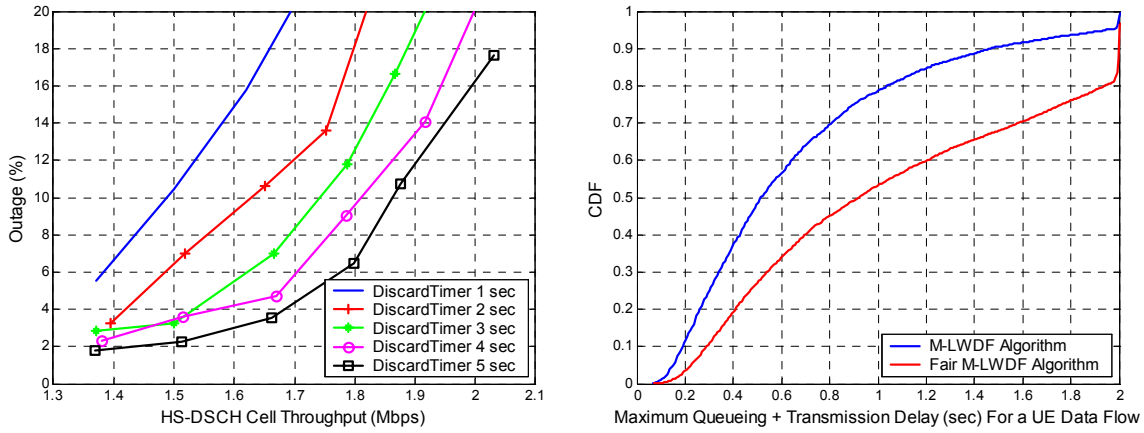
limitations on the maximum delay jitter the conveying networks can introduce. In the present system model, the limitation is directly imposed on the radio access network (and more specifically on the dominant source of jitter, i.e. the Node B queuing) because the Internet and the CN are assumed not to introduce any delay jitter at all.

The cell throughput results are depicted in Figure 7.11. It can clearly be seen that the results for discard timers up to 3 seconds do not significantly differ from previous simulation cases without any type of buffering limitations. However, the increase of the discard timer beyond 3 seconds does not yield any system throughput gain but rather a loss, which could be expected because the initial UE Dejitter buffering was reduced to exactly 3 seconds, although the performance loss is minor even for a discard timer of 5 seconds. This is a positive result because it implies that the user's perceived quality would not be very degraded by a minor error in the planning of the total delay jitter budget.

7.5.5 Introducing Fairness Into the M-LWDF Algorithm

Figure 7.12 plots the system throughput and the associated outage for the Packet Scheduling algorithm whose priority is computed according to the equation (7.6). Comparing to Figure 7.7, it can be concluded that the cell capacity at 5% outage level of the M-LWDF algorithm is higher than its fair version regardless of the discard timer. For example, for a discard timer of 3 seconds, the capacity reduction with the fair version of the M-LWDF equals 13%. The reason can be highlighted with the picture on the right hand side of Figure 7.12. It depicts the maximum Node B queuing plus transmission delay for a UE data flow for both types of schedulers.

The fair version of the M-LWDF algorithm, by raising the priority of less favourable users, increases the overall queuing delay causing that a larger number of data flows reach the due delay. Since all the streaming users in the simulation achieve approximately the same



a) *b)*

Figure 7.12: Cell Throughput Performance (a) and Maximum Node B Queuing + Transmission Delay for a Data Flow in Case of Input Load of 1.9 Mbps and Discard Timer 2 sec (b). 128 kbps CBR Bearers in Pedestrian A 3km/h, Total Node B Buffering of 20 Mbits, Initial UE Buffering of 5 sec.

throughput, it is less convenient to give higher priority to low G Factor users because they consume larger amount of resources. Thereby, as the QoS criterion (outage) does not evaluate the nature of the poor performing users and it only counts how many flows do not experience an acceptable QoS, it is more efficient to degrade the quality of the most resource demanding users (low G Factor).

Hence, it can be concluded that, for non-elastic traffic, the provision of user fairness (in a max-min manner) in a wireless shared link does not necessarily yield a cell throughput improvement, and in this case it rather gives a cell capacity loss.

7.6 Comparison of System Level Performance between CBR Video Streaming and NRT Traffic

The attractive trade-off between cell throughput and degree of user fairness achieved by the Proportional Fair algorithm provides a promising solution to convey NRT traffic on HSDPA while reaching a high spectral efficiency. In the previous section, it has been shown that the M-LWDF algorithm represents a viable solution for satisfying the demanding real time QoS requirements while still providing a high system capacity in HSDPA. However, in order to assess the feasibility of mapping streaming services on to HSDPA, it is relevant to evaluate what is the potential degradation of conveying such services compared to NRT services.

For the system level performance comparison, quasi-dynamic system level simulations of NRT traffic have been carried for the Proportional Fair and the Max CI schedulers. For these simulations, the mobility, cell layout and propagation model, the transmission power and the receiver modelling, and their associated default parameters are kept as in Chapter 5 and Chapter 6. A simple traffic model is employed for the NRT traffic (i.e. the one described in the section B.3 of the Appendix B).

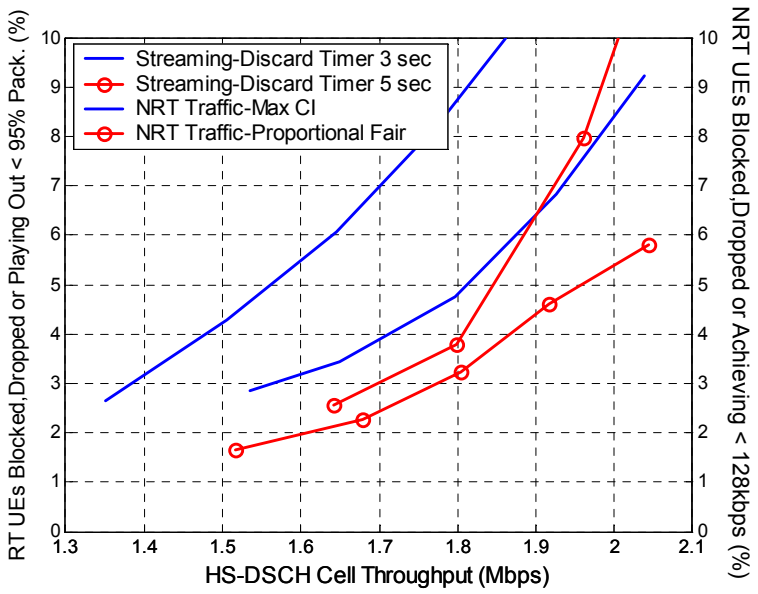


Figure 7.13: Cell Throughput Performance for NRT and Streaming traffic in Ped_A 3km/h. For the Streaming Case, the Bearers Convey 128 kbps CBR video, the Total Node B Buffering is 20 Mbits, Initial UE Buffering is 5 sec. For the NRT Case the Minimum Data Rate is 128 kbps, and the Packet Call Size is 100 kbytes.

Figure 7.13 plots the HSDPA cell throughput versus the real time outage level (defined in the section B.5, and expressed on the left y-axis of the figure) for streaming services served with the M-LWDF algorithm and with a discard timer of 3 and 5 seconds. The same picture depicts the HSDPA cell throughput versus the non real time outage level with minimum user throughput of 128kbps (defined in section B.3 of the Appendix B, and expressed on the right y-axis of the figure) for NRT traffic served with the Proportional Fair and the Max CI algorithm.

It is relevant to note that the outage definition has strong implications on determining the cell throughput. Nonetheless, the performance comparison is surprising because the real time flows with the M-LWDF algorithm (e.g. with a discard timer of 3 seconds) achieve a 20% higher cell capacity than the NRT flows with the Max CI scheduler and similar capacity to the Proportional Fair scheduler at 5% outage level. This could be initially unexpected because the provision of very demanding QoS constrains is assumed to have a cost in terms of spectral efficiency.

The main reason behind this result is the difference on the dropping criterion applied for both types of traffics. The dropping criterion is very exigent for streaming services (i.e. all flows not capable of transmitting at least 80% of the total amount of RTP packets are dropped) because users not fulfilling this criterion are expected to perceive an unacceptable video image quality. On the other hand, the dropping criterion is rather loose for NRT services, since bearers are dropped only if they achieve lower throughput than 9.6 kbps. This is because NRT users reaching a throughput lower than 128 kbps (e.g. obtaining 64 kbps) are not satisfied because do not obtain the minimum data rate, but still they may not perceive this situation as unacceptable. This difference on the dropping criterion permits the NRT users with poor average radio propagation conditions to maintain their connection alive, whereas the same users with poor radio channel quality will be dropped if they have a streaming

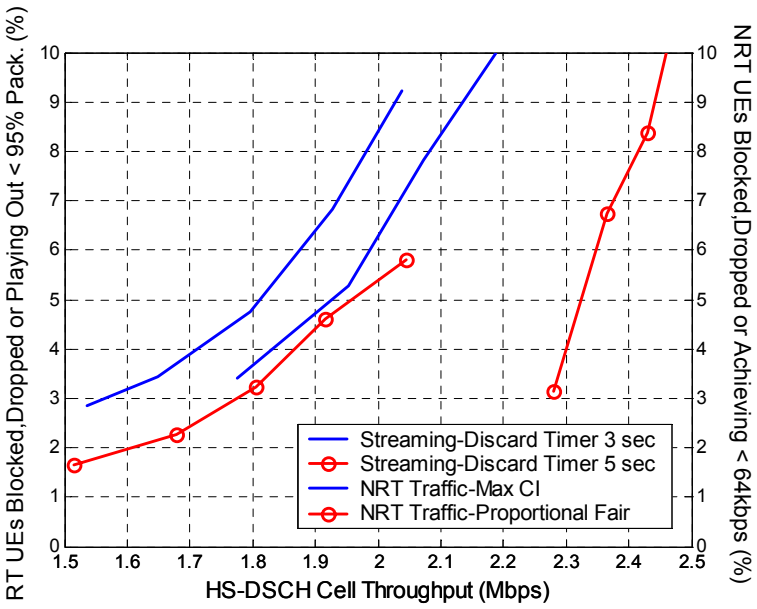


Figure 7.14: Cell Throughput Performance for NRT and Streaming traffic in Ped_A 3km/h. For the Streaming Case, the Bearers Convey 128 kbps CBR video, the Total Node B Buffering is 20 Mbits, Initial UE Buffering is 5 sec. For the NRT Case the Minimum Data Rate is 64 kbps, and the Packet Call Size is 100 kbytes.

connection. Recall from previous sections, that, in the case of streaming services, the low G Factor users are the ones consuming most of the cell resources.

Another interesting observation of the Figure 7.13 is that the slope of the cell throughput achieved by streaming flows with the M-LWDF algorithm is similar to the slope of the cell throughput obtained by NRT bearers with the Max CI algorithm. The reason behind is that, as pointed out in the previous section, the M-LWDF algorithm is rather unfair, while the unfairness is one of the main properties of the Max CI scheduler.

Figure 7.14 plots the outage level versus the HSDPA cell throughput for streaming and NRT services as done in Figure 7.13, with the difference that the minimum throughput for the NRT bearers is 64 kbps in Figure 20. In the case of a less exigent QoS criterion for the NRT traffic (i.e. a minimum data rate of 64 kbps), its cell throughput gets closer to the maximum capacity, which unbalances the system performance in favour of NRT traffic. See how the Proportional Fair algorithm outperforms by 29% the cell throughput of the M-LWDF algorithm with a discard timer of 3 seconds at 5% outage level.

7.7 Conclusions

The present chapter has analysed the performance of multimedia streaming services on HSDPA focusing on the video component of the multimedia service due to its larger bandwidth requirements.

For Packet Scheduling of CBR encoded streams on HSDPA, the M-LWDF algorithm described by Andrews, that aims at satisfying the packet delay while utilizing the fast channel quality information, represents a well suited algorithm for providing high efficiency streaming

services on HSDPA. The results have shown that the Proportional Fair algorithm provides 30% cell throughput gain for NRT traffic (with a minimum data rate of 64kbps) relative to the cell throughput achieved by the M-LWDF for CBR streaming flows (for 128 kbps bearers with a discard timer of 3 seconds and in macro cell Pedestrian A environment). That performance gain disappears when the cell throughput of those CBR streaming flows is compared to that of NRT traffic with minimum user throughput of 128kbps. As it has been commented, the M-LWDF algorithm can not be directly applied to the case of VBR encoded streams, although the paradigm may be possibly solved by the combination of this scheduling algorithm with a token bucket algorithm that control the data flows rates and guarantees the GBR.

Focusing on CBR encoded video streams, from a system level performance point of view, it has been shown that the M-LWDF algorithm is a rather unfair scheduling principle where the users with poor average radio propagation conditions, which suffer from higher delays than the remaining users in the cell, are not able to fulfil the QoS criteria during high load situations. As shown in Figure 7.6, for 128 kbps bearers and a cell load of 1.9 Mbps, all the users whose percentage of packet discarding at the Node B is larger than 5% have a G Factor below 1dB, while users with a G Factor higher than 5 dB obtain a percentage of packet discarding at the Node B equal to zero. Motivated by the fact that the available literature addresses the challenge of guaranteeing QoS for streaming flows by targeting at providing throughput fairness, here it has been proposed a fair scheduling algorithm that combines the M-LWDF policy with Barriac's algorithm [86]. However, this fair version of the M-LWDF has been shown to yield a cell throughput reduction of 13% for 128 kbps bearers in Pedestrian A and with a discard timer of 3 seconds. Hence, it must be concluded that, for non-elastic traffic, the provision of user fairness (in a max-min manner) in a wireless shared link does not necessarily yield a cell throughput improvement, and in this case it rather gives a cell performance loss.

Regarding the sensitivity to the application delay jitter constraints, it has been shown that its impact on the system throughput performance is not as large as might be initially expected. A discard timer increase from 1 to 4 seconds in a Pedestrian A and 3 km/h environment produces a cell capacity gain of around 35% gain, while the same discard timer increase in a Vehicular A and 3 km/h environment yields a 20% gain. It is concluded that a discard timer of 3 seconds represents a good trade-off between HSDPA cell throughput and Node B & UE buffering requirements. In terms of the power delay profile, the time dispersion of the Vehicular A model yields a cell throughput reduction of 55% versus the Pedestrian A model (for a discard timer of 3 seconds).

Regarding the average bit rate of the source, it has been described that 64 kbps video streaming incurs slightly higher power and code overhead than 128 kbps bearers. Streaming connections with 128 kbps bit rates seem a suitable choice for HSDPA, and only represent a minor cell capacity reduction compared to best effort NRT traffic using the Proportional Fair algorithm with a minimum user throughput of 128 kbps.

Chapter 8

Conclusions

8.1 Preliminaries

This dissertation has investigated the paradigm of providing packet switched services in the evolution of Third Generation cellular systems. The goal of Third Generation systems is to provide users not only with the traditional circuit switched services, but also with new multimedia services with high quality images and video for person-to-person communication, and with access to services and information in private and public networks. The provision of IP-based services has created expectations of a major traffic growth in mobile communication systems.

During the evolution of the UMTS, the WCDMA technology faces the challenge of supporting an extensive variety of services with ever increasing demands for higher data rates, and with a wide diversity of QoS needs. Moreover, the expected increase for traffic demand drives operators to require larger system capacity to sustain their business growth.

Under that scenario, this Ph.D. thesis has assessed HSDPA as a potential radio access technology that could facilitate the evolution of the wireless communication market by satisfying the demands of the end users and the operators. Regarding Non Real Time (NRT) services, the focus has been on TCP based traffic, because a large share of NRT applications are expected to employ a TCP/IP platform due to its enormous success on the Internet. Regarding Real Time services, the focus has been on packet switched streaming applications because they also are a relevant traffic type to be conveyed by HSDPA.

In this concluding chapter, a summary of this dissertation is given. Every section of this chapter draws the main conclusions of the investigations of each chapter of the thesis.

8.2 Packet Switched Services in Third Generation Systems

Chapter 2 concentrated on building understanding on the services to be conveyed by HSDPA.

Regarding the analysis of a multimedia session, the main attention has been on the video component of the session due to its variable rate nature and its bandwidth demands (higher than other components such as the audio). Two different types of video encoding and have been presented. On the one side Constant Bit Rate (CBR) video encoding generates a constant bit rate data flow at expenses of the degradation of the final image quality. UMTS can easily convey this type of video encoding data flow by setting the guaranteed bit rate (GBR) and the maximum bit rate (MBR) equal to average data stream bit rate.

On the other side, Variable Bit Rate (VBR) video encoding provides a constant image quality at the cost of generating an encoded data stream with variable bit rate. The long bit rate peaks caused by this type of encoding can be supported by UMTS by setting the GBR and MBR equal to the maximum encoder bit rate, although this solution represents a significant waste of the resources because the encoded sequence seldom requires a bearer capabilities equal to the maximum encoder bit rate. VBR flows can still be transmitted at near constant bit rate through the Constant Bit Rate Packet (CBRP) transmission technique, but it seems to require too large Predecoder buffering at the UE to be a viable solution. A potential solution to convey VBR video sequences consists of setting the GBR equal to the MBR and relatively higher (e.g. 15-60%) than the encoder average bit rate while still using the UE Predecoder buffering. The limited increase of the GBR is directly translated into a reduction of the Predecoder buffer size (e.g. for the video trace case considered in Chapter 2, a 15% increase of the GBR implies around 50% reduction of the Predecoder buffer dynamic range).

Regarding interactive services, the focus has turn to Web browsing applications. Every browsing session is composed by several page downloads separated by periods when the end user (human) analyses the retrieved information, where both, Web pages and reading times, follow a heavy tail distribution. With HTTP 1.1, multiple objects within a Web page (usually no larger than seven objects on average) can be downloaded in the same TCP connection. However, if the reading time outlasts the server's connection holding time (e.g. 15 seconds for a common Internet server), then new TCP connections will be needed for subsequent pages with the negative effect the TCP connection set-up and slow start.

8.3 TCP Performance over WCDMA

Chapter 3 investigated the performance of the TCP protocol over WCDMA/DCH. On a general perspective, it has been concluded that there exists a significant performance difference between low and high bit rate channels for realistic round trip times (150-450 ms). The higher bit rates allocated to the user, the bigger is the potential risk of the TCP congestion window not reaching the bandwidth delay product. This risk can lead to the user throughput

being degraded below the allocated channel bit rate, as well as potential code shortage (especially in microcell environments).

For small file sizes, such as web pages, the initial slow start can harm the user throughput of users allocated high channel bit rates. For example, a user that is allocated a 256 kbps channel experiences 25% RLC throughput degradation with a 50 kbytes file and a round trip time of 140 ms. Users assigned lower bit rates channels, such as a 64 kbps, experience a higher file download delay, but hardly suffer any degradation from the initial slow start. Users assigned bit rates up to 256kbps can efficiently utilize the allocated resources if their terminals are equipped with TCP receiver's buffer capabilities of at least 16 kB. For higher bit rates, such as 384 kbps, they require at least 32 kB.

The introduction of random segment losses significantly changes the behaviour of the protocol. Low bit rate channels, such as 64 kbps, hardly suffer any throughput degradation for 1% loss probability, whereas for a 256 kbps channel and a round trip time of 300 ms the same segment loss rate already degrades the RLC throughput by 25%. Regarding the interaction between the RLC and the TCP layer, it has been shown that the more reliable is the RLC layer, the more benefits are obtained in terms of end user throughput. This conclusion has been evaluated for RLC retransmission delays up to 180 ms.

8.4 Packet Scheduling for NRT Services on HSDPA

Chapter 5 has assessed the performance of HSDPA for elastic non delay sensitive traffic. The maximum capacity achieved by the Proportional Fair algorithm in macro cell (Vehicular A) at 3 km/h with around 50% HSDPA power and seven codes is around 1.44 Mbps, while the maximum capacity for Round Robin under the same conditions equals around 850 kbps. A rough estimate of the cell capacity assuming 80 % of HSDPA (the remaining cell power is supposed to be overhead) could be obtained by simply rescaling previous results. Comparing[22] these absolute cell capacity figures with the WCDMA capacity, reported in [10] to be 915 kbps for macrocell environment (and fair resources distribution among users in the cell), yields an absolute capacity gain of 150% of HSDPA with Proportional Fair, and a gain of around 50% of HSDPA with Round Robin over WCDMA. Therefore, HSDPA (with fast packet scheduling) is indeed a radio technology capable of providing a significant gain that could contribute to cope with the increase of traffic demands in the evolution of WCDMA.

The investigation has specially concentrated on assessing different scheduling methods. Max C/I is the scheduling method providing highest HSDPA cell capacity for best effort NRT traffic, although it causes the starvation of the poor G Factor users (reduces the cell coverage). However, when minimum user throughput is to be guaranteed, the Proportional Fair scheduling algorithm outperforms all other tested algorithms, because it provides an attractive trade-off between cell throughput and user fairness while it exploits the multi-user diversity. With this algorithm, in Pedestrian at 3km/h, providing a minimum user throughput of 64 kbps (at 5% outage) only incurs 11% cell throughput reduction compared to the maximum the cell capacity. Supporting at least 128 kbps (at 5% outage) shows an acceptable cell throughput reduction of around 30% compared to the maximum cell capacity. In this case around 45% of

[22] Note that the comparison of absolute cell capacity figures is difficult when the underlying assumptions to provide those figures are not exactly the same. Therefore, this capacity comparison must be considered as a first order approximation.

the users obtain at least 384kbps and 10% of the users reach more than 1Mbps. HSDPA can still guarantee higher data rates such as 384 kbps (at 5% outage), though it notably reduces the cell throughput (80% reduction in Pedestrian A at 3km/h).

8.5 System Performance of TCP Traffic on HSDPA

Chapter 6 has studied the performance of TCP based NRT traffic over HSDPA. It has been concluded that the load operating point should be selected to avoid large Node B queuing delays because they increase the risk of triggering TCP time outs when RLC retransmissions occur. The Node B queuing exhibits an exponential like growth with the load causing significant packet queuing delay at the Node B for high system loads. It has been shown that under the model assumptions given in Appendix B, an offered load increase from 1.4 to 1.8 Mbps leads to a cell throughput benefit of 29%, but an average TCP packet queuing delay of 72%, and an offered load increase from 1.8 to 2.2 Mbps yields a cell throughput growth of 20% but an average segment queuing delay of 85%. Furthermore, for similar blocking level, the Proportional Fair scheduler shows a significant reduction of the queuing delay compared to Round Robin due to its higher data rates. At around 0.2% blocking Round Robin has an average queuing delay of around 1.07 seconds, while for 0.4% blocking Proportional Fair only causes an average segment queuing delay of around 625 milliseconds.

It has been described that the time-shared nature of the HS-DSCH channel is suitable to cope with the burstiness of TCP traffic. At high loads, the large number of users in the cell reduces the throughput allocated to every user. The consequence is that the number of active users (multiuser diversity order) is hardly affected, and therefore the cell capacity reduction is minor (for example 5% capacity decrease for a packet call size reduction from 100 to 50 kbytes in Pedestrian A at 3km/h).

8.6 Streaming Services on HSDPA

Chapter 7 has analysed the performance of multimedia streaming services on HSDPA focusing on the Constant Bit Rate encoded stream from the video component of a multimedia service. The most relevant conclusion of the investigation is that HSDPA can actually provide service to delay sensitive applications without major capacity loss as long as the discard timer is long enough (2-3 seconds, which should be manageable for future terminals).

For Packet Scheduling of CBR encoded streams on HSDPA, the M-LWDF algorithm described by Andrews, that aims at satisfying the packet delay while utilizing the fast channel quality information, represents a well suited algorithm for providing high efficiency streaming services on HSDPA. The results have shown that the Proportional Fair algorithm provides 30% cell capacity gain for NRT traffic (with a minimum data rate of 64kbps) relative to the cell throughput achieved by the M-LWDF for CBR streaming flows (for 128 kbps bearers with a discard timer of 3 seconds and in Pedestrian A environment). That capacity gain disappears when the cell throughput of those CBR streaming flows is compared to that of NRT traffic with minimum data rate of 128kbps.

Motivated by the fact that the available literature addresses the challenge of guaranteeing QoS for streaming flows by targeting at providing throughput fairness, here it has been proposed a

fair scheduling algorithm that combines the M-LWDF policy with Barriac's algorithm [86]. However, this fair version of the M-LWDF has been shown to yield a cell capacity reduction of 13% for 128 kbps bearers in Pedestrian A and with a discard timer of 3 seconds. Hence, it must be concluded that, for non-elastic traffic, the provision of user fairness (in a max-min manner) in a wireless shared link does not necessarily yield a cell throughput improvement, and in this case it rather gives a cell capacity loss.

8.7 Future Research

There exist multiple lines of investigation to continue the research carried out in this Ph.D. thesis.

One of the areas requiring most immediate investigation is the influence of the flow control functionality of the Iub interface on the HSDPA performance. As it has been shown before, for TCP based traffic, the large Node B queuing in HSDPA has the potential risk of triggering spurious TCP time outs when RLC retransmissions occur. A significant fraction of all the queued data could be buffered in the RNC instead. Then, the Iub flow control functionality should track the Node B queuing associated to each MAC-d flow and refill them from the RNC when needed. This is not an easy task because the flow control functionality must be fast enough to avoid that any MAC-d flow queue in the Node B becomes empty. Moreover, the very different user data rates provided by HSDPA (some users only obtain 64kbps while other reach more than 1Mbps) further complicate the task.

The flow control of the Iub interface does not only require investigation for NRT traffic, but also for streaming traffic. The scheduling algorithms described in Chapter 7 do not have any control on the actual rate of the streaming flow. The Iub flow control must ensure that the Guaranteed Bit Rate of the streaming bearers is satisfied. Moreover, if the HSDPA capacity allows it, the Iub flow control should enable data rates larger than the Guaranteed Bit Rate (up to the maximum specified by the Maximum Bit Rate attribute). This extra bit rate should be able to accommodate the temporary bit rate peaks of Variable Bit Rate encoded video sequences.

Another relevant topic that requires further investigation is the HSDPA coverage. In this study it has been assumed that the HS-DSCH provides service to all the area in the cell. But the soft handover feature of the Dedicated Channels might be a more power efficient solution for users located close to the cell edge. This analysis might be more difficult than initially expected, because, if the user throughput is to be distributed in the cell according to any specific strategy (e.g. fair throughput distribution), the allocation of cell resources between HSDPA and DCH might be very different.

The impact of the hard handover on the delay jitter for streaming flows mapped on HSDPA is another point requiring investigation. Again, the usage of the soft handover operation of Dedicated Channels to users located close to the cell edge may improve the performance.

But perhaps the most interesting topic that remains open after this Ph.D. investigation is the design of a Packet Scheduling policy capable of serving simultaneously mixed traffic with different QoS requirements. As described in Chapter 2, the services belonging to different traffic classes have dissimilar QoS attributes to be satisfied. Moreover, within the same traffic class the ARP attribute differentiates users according to their subscription. Furthermore, in the

case of interactive services, the THP parameter adds another dimension for traffic distinction. The MAC-hs design must be able to integrate the scheduling strategies for non-delay sensitive (described in Chapter 5) and for delay sensitive (described in Chapter 7) in a single algorithm that should as well satisfy the ARP and the THP attributes.

References

[1] Nokia Corporation. *Nokia HSDPA Solution.* White Paper, www.nokia.com. 2003.

[2] Ralph D. et al. *3G and Beyond – The applications Generation.* 3G Mobile Communications Technologies, 2002. Third International Conference on. 2002. pp. 433-438.

[3] UMTS Forum. *White Paper No 1. Evolution to 3G/UMTS Services.* www.umts-forum.org. August 2002.

[4] Frodigh M. et al. *Future-Generation Wireless Networks.* IEEE Personal Communications, October 2001. Volume 8. Issue 5. pp. 10-17.

[5] Huber J. *Quo Vadis Mobile Applications?* Proceedings of Momentum, August 2003.

[6] Sprint Corporation. http://www.sprint.com/pcsbusiness/compare/speed.html.

[7] Halonen T. et al. *GSM, GPRS and EDGE Performance.* Wiley, 2002.

[8] Siemens Corporation. *Comparison of W-CDMA and CDMA2000.* White Paper, www.siemens.com.

[9] Hedberg, T et al. *Evolving WCDMA.* Ericsson White Paper, March 2001. www.ericsson.com

[10] Holma H. et al. (Editors). *WCDMA for UMTS. Second Edition* Wiley, 2002.

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

[12] Helmersson K. et al. *Performance of Downlink Shared Channels in WCDMA Radio Networks.* Vehicular Technology Conference, 2001. VTC 2001 Spring. Volume 4. pp. 2690-2694.

[13] 3GPP. Technical Specification Group Services and System Aspects. *QoS Concept and Architecture (3GPP TS 23.107 version 5.9.0).*

[14] Reguera E. et al. *Enhanced UMTS Services and Applications: a Perspective beyond 3G.* IEE 5th European Personal Mobile Communications Conference. April 2003. seacorn.ptinovacao.pt/docs/papers/service2.pdf

[15] Montes H. et al. *Deployment of IP Multimedia Streaming Services in Third-Generation Mobile Networks.* IEEE Wireless Communications. Volume 9. pp. 84-92. October 2002.

[16] Mah B.A. *An Empirical Model of HTTP Network Traffic.* INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communication Societies. Volume 2. pp. 592-600. April 1997.

[17] Khaunte S.U. et al. *Statistical Characterization of World Wide Web Browsing Session.* Technical Report GIT-CC-97-17. College of Computing, Georgia Institute of Technology. 1997. http://citeseer.nj.nec.com/khaunte97statistical.html.

[18] Choi H. et al. *A Behavioural Model of Web Traffic.* Seventh International Conference on Network Protocols. pp. 327-334. November 1999.

[19] Molina M. et al. *Web Traffic Modeling Exploiting TCP Connections' Temporal Clustering through HTML-REDUCE.* IEEE Network. Volume 14. pp. 46-55. May 2001.

[20] Crovella M. et al. *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes.*

Networking, IEEE/ACM Transactions on, Volume 5. December 1997. pp. 835-846.

[21] Vicari N. *Measurements and Modeling of WWW-Sessions.* Report 187. Institute of Computer Science, University of Würzburg. Semtember 1997. http://www-info3.informatik.uni-wuerzburg.de/TR/tr184.pdf.

[22] Casilari E. et al. *Characterisation of Web Traffic.* Global Telecommunications Conference, Globecom'01 Volume 3. pp. 25-29. November 2001.

[23] Stuckmann P. *Performance Characteristics of the Enhanced General Packet Radio Service For the Mobile Internet Access.* 3G Mobile Communications Technologies, 2001. Second International Conference on. 2001. pp. 287-291.

[24] TSG-RAN Working Group 1 #21. *HSDPA Number of Users.* Source: Qualcomm. R1-01-0912. August 2001.

[25] Velez F. J. et al. *Mobile Broadband Services: Classification, Characterization, and Deployment Scenarios.* IEEE Communications Magazine, Volume 40, April 2002. pp. 142-150.

[26] 3GPP. Technical Specification Group Radio Access Network. *High Speed Downlink Packet Access; Overall UTRAN Description. (3GPP TR 25.855 version 5.0.0).*

[27] 3GPP. Technical Specification Group Services and System Aspects. *QoS Concept. (3GPP TR 23.907 version 5.7.0).*

[28] 3GPP. Technical Specification Group Services and System Aspects. *Service Aspects; Services And Service Capabilities. (3GPP TR 22.105 version 6.2.0).*

[29] 3GPP. Technical Specification Group Services and System Aspects. *Transparent end-to-end Packet-switched Streaming Service (PSS); General Description. (3GPP TS 26.233 version 4.2.0).*

[30] Montes H. et al. *An end-to-end QoS Framework for Multimedia Streaming Services in 3G Networks.* Personal, Indoor and Mobile Radio Communications, 2002. The 13th IEEE International Symposium on, Volume 4. 2002. pp. 1904-1908.

[31] 3GPP. Technical Specification Group Services and System Aspects. *Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs. (3GPP TS 26.234 version 5.3.0).*

[32] 3GPP. Technical Specification Services and System Aspects. *Transparent end-to-end Packet switched Streaming Service (PSS); RTP usage model. (3GPP TR 26.937 version 1.2.0).*

[33] 3GPP. Technical Specification Group Radio Access Network. *UTRAN Iub Interface NBAP signalling. (3GPP TR 25.433 version 5.5.0).*

[34] Stevens R *TCP/IP Illustrated, Volume 1. The Protocols.* Addison Wesley, 1994.

[35] Network Working Group. *RFC 1889. RTP: A Transport Protocol for Real-Time Applications.* January 1996.

[36] Varsa V. et al. *Long Window Rate Control for Video Streaming*. Proceedings of the 11th International Packet Video Workshop, 30 April- 1 May, 2001, Kyungju, South Korea.

[37] Crispin M. Network Working Group. *RFC 2060. Internet Message Access Protocol. Version 4Rev1*. December 1996.

[38] Wap Forum. *Wireless Application Protocol. White Paper*. http://www.wapforum.org/. June 2000.

[39] Wap Forum. *Wireless Application Protocol. WAP 2.0 White Paper.* http://www.wapforum.org/. January 2002.

[40] Kaaranen H. et al. *UMTS Networks.* Wiley, 2001.

[41] 3GPP. Technical Specification Group Radio Access Network. *Packet Data Convergence Protocol (PDCP) Specification. (3GPP TS 25.323 version 4.1.0).*

[42] 3GPP. Technical Specification Group Radio Access Network. *RLC Protocol Specification. (3GPP TR 25.322 version 4.1.0).*

[43] 3GPP. Technical Specification Group Radio Access Network. *MAC Protocol Specification. (3GPP TS 25.321 version 5.1.0).*

[44] Fielding R. et al. Network Working Group. *RFC 2616. Hypertext Transfer Protocol.* June 1999.

[45] Reyes-Lecuona A. et al. *A Page-Oriented WWW Traffic Model for Wireless System Simulations.* International Teletraffic Congress, ITC-16. Volume 3.b, pp. 287-291. June 1999.

[46] 3GPP. Technical Specification Group Radio Access Network. *Physical Layer Aspects of UTRA High Speed Downlink Packet Access. (3GPP TR 25.848 version 0.5.0).* May 2000.

[47] Cohen E. et al. *Managing TCP Connections Under Persistent HTTP.* Computer Networks, 31. pp. 1709-1723. March 1999.

[48] UMTS. Universal Mobile Telecommunications System. *Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS. (UMTS 30.03 version 3.2.0). May, 1998.*

[49] Huffaker B. *Round Trip Time Internet Measurements from CAIDA's Macroscopic Internet Topology.* http://www.caida.org/analysis/performance/rtt/walrus0202/. June, 2002.

[50] Hernandez-Valencia E. et al. *Transport Delays for UMTS VoIP.* Wireless Communications and Networking Conference, 2000. WCNC, September 2002. Volume 3. pp. 1552-1556.

[51] 3GPP. Technical Specification Group Radio Access Network. *Delay Budget Within the Access Stratum. (3GPP TR 25.853 version 4.0.0).* May 2001.

[52] Miller G. *The Nature of the Beast: Recent Traffic Measurements from an Internet Backbone.* April, 1998. http://www.caida.org/outreach/papers/1998/Inet98/Inet98.html.

[53] Chockalingam A., et al. *Performance of TCP on Wireless Fading Links With Memory.* Communications, 1998. The IEEE International Conference on, ICC 98. Volume 1. June 1998. pp. 595-600.

[54] Chockalingam A., et al. *Performance of TCP/RLP Protocol Stack on Correlated Fading DS-CDMA Wireless Links.* Vehicular Technology Conference, 1998. VTC 1998, May. Volume 1. pp. 363-367.

[55] Lefevre F., et al. *Optimizing UMTS Link Layer Parameters for a TCP Connection.* Vehicular Technology Conference, 2001. VTC 2001, Spring. Volume 4. pp. 2318-2321.

[56] Network Working Group. *RFC 1191. Path MTU Discovery.* November 1990.

[57] Jacobson V. *Congestion Avoidance and Control.* ACM Computer Communication Review, SIGCOMM, Proceedings of, August 1998. pp. 314-329.

[58] Allman M., et al. *Increasing TCP's Initial Window*. Experimental Protocol RFC 2414, September

1998, http://www.faqs.org/rfcs/rfc2414.html.

[59] Network Working Group. *RFC 2581. TCP Congestion Control.* April 1999.

[60] Network Working Group. *RFC 2988. Computing TCP's Retransmission Timer.* November 2000.

[61] Ameigeiras P. et al. *Impact of TCP Flow Control on the Radio Resource Management of WCDMA Networks.* Vehicular Technology Conference, 2002. VTC 2002 Spring. Volume 2. pp. 977-981.

[62] Bestak R., et al. *RLC Buffer Occupancy When Using a TCP Connection over UMTS.* Personal, Indoor and Mobile Radio Communications, 2002. The 13th IEEE International Symposium on, Volume 3. 2002. pp. 1161-1165.

[63] Fall K. et al. *Simulation-based Comparisons of Tahoe, Reno, and SACK TCP.* Computer Communication Review. July 1996. ftp://ftp.ee.lbl.gov/papers/sacks.ps.Z

[64] Ludwig R. et al. *The Eifel Algorithm: Making TCP Robust Against Spurious Retransmissions.* ACM Computer Communications Review, January 2000.

[65] Lakshman S., et al. *The Performance of TCP/IP for Networks with High Bandwidth-Delay Products and Random Loss.* Networking, IEEE/ACM Transactions on, Volume 7. June 1997. pp. 336-350.

[66] Kolding T. et al. *Performance Aspects of WCDMA Systems with High Speed Downlink Packet Access (HSDPA).* Vehicular Technology Conference, 2002. VTC 2002 Fall. Volume 1. pp. 477-481.

[67] 3GPP. Technical Specification Group Radio Access Network. *High Speed Downlink Packet Access (HSDPA); Overall Description. (3GPP TS 25.308 version 5.4.0).*

[68] Proakis J. *Digital Communications.* McGraw-Hill, 2nd Edition, 1989.

[69] 3GPP. Technical Specification Group Radio Access Network. *Radio Resource Control (RRC); Protocol Specification. (3GPP TS 25.331 version 5.5.0).*

[70] 3GPP. Technical Specification Group Radio Access Network. *Physical Layer Procedures (FDD). (3GPP TS 25.214 version 5.5.0).*

[71] Miyoshi K, et al. *Link adaptation method for High Speed Downlink Packet Access for W-CDMA.* WPMC Proceedings. Volume 2. pp. 455-460. 9th –12th September, 2001. Aalborg.

[72] TSG-RAN Working Group1. Source: Nokia. *HSDPA DL Timing.* 5th –7st, November2001.

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

[74] Frederiksen F. et al. *Performance and Modeling of WCDMA/HSDPA Transmission/H-ARQ Schemes.* Vehicular Technology Conference, 2002. VTC 2002 Fall. Volume 1. pp. 472-476.

[75] Holtzman J.M. *CDMA Forward Link Waterfilling Power Control.* Vehicular Technology Conference, 2000. VTC 2000 Spring. Volume 3. pp. 1663-1667.

[76] Jalali A. et al. *Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System.* Vehicular Technology Conference, 2000. VTC 2000 Spring. Volume 3. pp. 1854-1858.

[77] Elliott R.C. et al. *Scheduling Algorithms for the CDMA2000 Packet Data Evolution.* Vehicular

Technology Conference, 2002. VTC 2002 Fall. Volume 1. pp. 304-310.

[78] Liu X. et al. *Opportunistic Transmission Scheduling With Resource-Sharing Constraints in Wireless Networks*. Selected Areas in Communications, Journal on, Volume: 19. Oct 2001. pp 2053-2064.

[79] Hosein P. *QoS Control for WCDMA High Speed Data.* Mobile and Wireless Communications Network, 4th International Workshop on, September, 2002. pp. 169-173.

[80] 3GPP. Technical Specification Group Radio Access Network. *UTRAN Iub Interface User Plane Protocols for Common Transport Channel Data Streams. (3GPP TS 25.435 version 5.4.0).*

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

[82] D. Bertsekas, Gallager R. *Data Networks. Second Edition.* Prentice Hall. 1992.

[83] Berger L. et al. *Interaction of Transmit Diversity and Proportional Fair Scheduling.* Vehicular Technology Conference, 2003. VTC 2003 Spring. Volume 4. pp. 2423-2427.

[84] Holtzman J.M. *Asymptotic Analysis of the Proportional Fair Algorithm.* Personal, Indoor and Mobile Radio Communications, 2001 12th IEEE International Symposium on , Volume: 2, Sep 2001. pp. F-33-F-37.

[85] Kelly F. *Charging and Rate Control for Elastic Traffic.* Europeans Transactions on Telecommunications, Volume 8, 1997. pp. 33-37. http://www.statslab.cam.ac.uk/~frank/elastic.html

[86] Barriac G., et al. *Introducing Delay Sensitivity into the Proportional Fair Algorithm for CDMA Downlink Scheduling.* Spread Spectrum Techniques And Applications, 2002 IEEE Seventh International Symposioum on, Volume 3, 2002. pp. 652-656.

[87] Wang I.-J. et al. *Weighted Averaging and Stochastic Approximation.* Decision and Control, 1996., Proceedings of the 35th IEEE, Volume: 1. 11-13 Dic. Pp 1071-1076.

[88] Kolding T. *Link and System Performance Aspects of Proportional Fair Scheduling in WCDMA/HSDPA.* To Appear in Vehicular Technology Conference, 2003. VTC 2003 Fall.

[89] Sipila K. et al. *Estimation of Capacity and Required Transmission Power of WCDMA Downlink Based on a Downlink Pole Equation.* Vehicular Technology Conference, 2000. VTC 2000 Spring. Volume 2. pp. 1002-1005.

[90] Kourtis S. et al. *Downlink Shared Channel: An Effective Way For Delivering Internet Services in UMTS.* 3G Mobile Communications Technologies, 2002. Third International Conference on. 2002. pp. 479-483.

[91] Ghosh A. et al. *Shared Channels for Packet Data Transmission in W-CDMA.* Vehicular Technology Conference, 1999. VTC 1999 Fall. Volume 2. pp. 943-947.

[92] Rappaport T. *Wireless Communications. Principles and Practice.* Prentice Hall, 1996.

[93] Yavuz M. et al. *TCP over Wireless Links with Variable Bandwidth.* Vehicular Technology Conference, 2002. VTC 2002 Fall. Volume 3. pp. 1322-1327.

[94] Peisa J. et al. *TCP Performance over HS-DSCH.* Vehicular Technology Conference, 2002. VTC 2002 Spring. Volume 2. pp. 987-991.

[95] Schieder A. et al. *Resource Efficient Streaming Bearer Concept for GERAN.* Wireless Personal

Communications, 2002. The 5th International Symposium on. Volume 2. pp. 858-862.

[96] Furuskär A. et al. *Performance of WCDMA High Speed Packet Data.* Vehicular Technology Conference, 2002. VTC 2002 Spring. Volume 3. pp. 1116-1120.

[97] Moulsley T.J. *Performance of UMTS High Speed Downlink Packet Access for Data Streaming Applications.* 3G Mobile Communications Technologies, 2002. Third International Conference on. 2002. pp. 302-307.

[98] Kelly F.P., et al. *Rate Control For Communication Networks: Shadow Prices, Proportional Fairness and Stability*. Journal of the Operational Research Society, vol. 49, pp. 237 -- 252, 1998.

[99] Lu S., et al. *Fair Scheduling in Wireless Packet Networks.* Networking, IEEE/ACM Transactions on, Volume 7. August 1999. pp. 473-489.

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

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

[102] Recommendation ITU-R M.1225. *Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000.* 1997.

[103] *Adaptive DCS Antenna, The Tsunami Project*. Technical Report, Celwave R.F., March 1996.

[104] Klingenbrunn T. *Downlink Capacity Enhancement of UTRA FDD Networks.* Ph.D. Thesis Dissertation. January 2001. Center for Person Kommunikation. Aalborg University.

[105] Ramiro-Moreno J. *Description of a Light WCDMA Dynamic Network Simulator for Testing Advanced Antenna Concepts in Release 5*. Nokia Internal Document. July, 2002.

[106] Laiho J. et al. (Editors). *Radio Network Planning and Optimisation for UMTS.* Wiley, 2002.

[107] Ramiro-Moreno J. et al. *Network Performance of Transmit and Receive Antenna Diversity in HSDPA under Different Packet Scheduling.* Vehicular Technology Conference, 2003. VTC 2003 Spring. Volume 2. pp. 1454-1458.

[108] Holma H. et al. (Editors). *WCDMA for UMTS.* Wiley, 2000.

Appendix A

Convergence Properties of Round Robin

The present section treats the convergence properties of the Round Robin scheduler in a time shared channel (such as HS-DSCH).

Let consider a set of coexisting users $\{i\}$ in the aforementioned time-shared channel. Let assume a blind Packet Scheduling strategy that does not make use of the channel quality information, and whose target is to distribute the time resources between the users. Let assume that, with the present scheduling algorithm S_1, every user has a link utilization factor of $\{\rho_i\}$ (see section 5.3.3 for the definition of the user link utilization factor). Then, the average user throughput is a linear function of the user's link utilization:

$$\lambda_i = \rho_i \cdot \overline{R_i} \qquad i = 1,..,N \tag{A.1}$$

where λ_i is average user throughput, ρ_i is the link utilization factor of user i, and $\overline{R_i}$ is the average user throughput of user i if it has full link utilization.

Let assume that S_1 applies an even distribution of the time resources: $\forall i, \rho_i = \rho / N$. Let further assume that the time shared channel is fully utilized (i.e. the overall link utilization equals 1):

$$\rho = \sum_i \rho_i = 1 \qquad i = 1,..,N \tag{A.2}$$

The set of throughputs $\{\lambda_i\}$ is proportionally fair if it is feasible and for no other feasible vector of throughputs $\{\lambda_i^*\}$ the aggregate of proportional changes is zero or negative [85]:

$$\sum_{\forall i} \frac{\lambda_i^* - \lambda_i}{\lambda_i} \leq 0 \qquad i = 1,..,N \tag{A.3}$$

Under previous assumptions, it can be demonstrated that S_1 is proportionally fair. Let assume any other blind scheduling policy S_2, which has a time resource distribution equal to S_1 except for users j and k who obtain a different user's link utilization:

$$\begin{aligned} \Delta + \rho_j &= \rho_j^* \\ \rho_k - \Delta &= \rho_k^* \end{aligned} \tag{A.4}$$

where ρ_j^* and ρ_k^* is the link utilization of users j and k respectively, and Δ is a value that represents the link utilization variation provided by the scheduling strategy S_2.

Applying the aggregate of proportional changes to the scheduling strategy S_1:

$$\begin{aligned} &\frac{\rho_j^* \cdot \overline{R_j} - \rho_j \cdot \overline{R_j}}{\rho_j \cdot \overline{R_j}} + \frac{\rho_k^* \cdot \overline{R_k} - \rho_k \cdot \overline{R_k}}{\rho_k \cdot \overline{R_k}} = \\ &\frac{\rho_j^* - \rho_j}{\rho_j} + \frac{\rho_k^* - \rho_k}{\rho_k} = \\ &\frac{\Delta}{\rho_j} - \frac{\Delta}{\rho_k} = \Delta \cdot \left(\frac{1}{\rho_j} - \frac{1}{\rho_k} \right) = 0 \end{aligned} \tag{A.5}$$

which demonstrates that S_1 is proportionally fair. Note that, if the link utilization factor of users j and k is not even for S_1, then it is possible to find Δ, so that the equation (A.3) yields a positive number.

Appendix B

HSDPA Network Simulator Model

The present section describes the modelling of the quasi-dynamic network level simulator for HSDPA.

B.1 Simulated Network Layout

The cell layout is a regular grid comprising a central HSDPA cell and 13 surrounding cells (see Figure B.1 of the appendix). The full detailed of all the dynamics is only simulated in the central cell of the network layout, while the remaining cells only provide a source of interference to the users served by the central cell.

B.2 Propagation Model, Mobility, Transmission Power and Receiver Modelling

- *Mobility, Cell Layout and Propagation Model*: Users are created in the central cell according to a uniform call arrival rate that is computed to fit a certain offered HSDPA cell load. The offered load is simply computed as the call arrival rate multiplied by the average size of the retrieved file. For NRT services the retrieved file corresponds to the packet call size, while for streaming users it corresponds to the average video bit rate multiplied by the mean call duration. Every user requests a single a packet call transfer, and dies right after the file download. Once created, UEs do not vary their

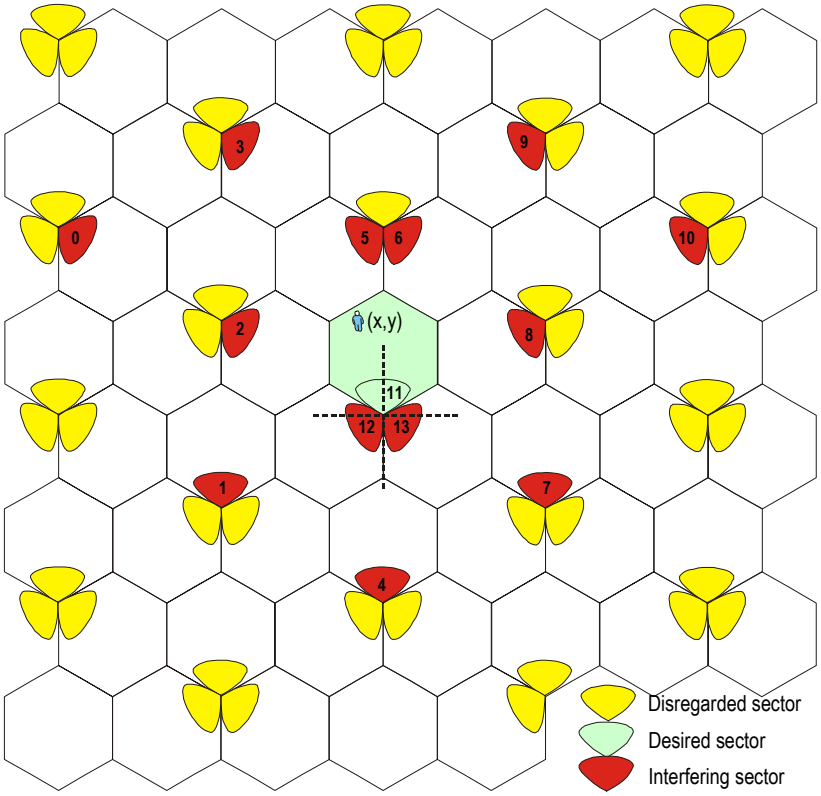


Figure B.1: Simulated Network Layout.

geographical location with time, and therefore, their deterministic path loss and shadow fading remains invariable during the lifetime of the user. This simplification avoids the implementation of hard handover for the HSDPA connections and the soft handover for the associated DCH connections. The shadow fading is assumed to be uncorrelated between connections belonging to the same sector, and between the links from different base stations that interfere a certain UE. Although users do not modify their geographical position, they are still assumed to move around within a short range, which is modelled by a fast fading with independently fading Rayleigh processes, whose power delay profile is described by the ITU Pedestrian A model (Vehicular A will also be used though not as default) [102]. A speed dependent Doppler effect is included in every tap of the power delay profile, and the user default speed is 3 km/h. The antenna radiation pattern is obtained from [103]:

$$Gain(\phi) = \begin{cases} G \cdot \cos^{3.2}(\phi) & |\phi| \le \frac{\pi}{2} \\ G / R & Otherwise \end{cases} \qquad \text{(B.1)}$$

where ϕ represents the location angle of the UE relative to the broadside direction of the corresponding antenna. The broadside antenna gain G is set to 18dB, whereas the front-to-back ratio is set to 33dB.

- *Transmission Power and Receiver Modelling*: The power in every cell of the is composed by the following factors:

$$P^{BS} = P^{HSDPA} + P^{CPICH} + P^{A\_DCH} + P^{NC} \quad \text{(B.2)}$$

Every cell transmits a constant HS-DSCH power P^{HSDPA}. The CPICH channel is modelled through a constant transmission power P^{CPICH}. In the central cell, users keep a low bit rate associated DPCH channel for signalling and link adaptation purposes whose contribution to the total cell transmission power is P^{A_DCH}. The remaining non-controllable cell power from DCH channels (modelling speech traffic bore by DCH power controlled channels) is modelled with a Lognormal distributed process P^{NC}, whose average is selected to maintain an average Node B transmission power approximately double than the HS-DSCH power. A conventional ideal Rake receiver is assumed to capture the power of the independently fading taps of the power delay profile. Maximal Ratio Combining (MRC) is applied after the channel compensated symbols from every finger to finally obtain the energy per symbol to noise ratio (Es/No). The Es/No is computed by performing geometrical averaging of the Es/No across the 3 slots that comprise the TTI. The mapping from the averaged symbol energy to noise rate into a corresponding block erasure rate is done as in [74]. For the Hybrid-ARQ protocol the default number of Stop And Wait (SAW) channels is 4, and the maximum number of Hybrid-ARQ retransmissions is 4. The Packet Scheduler always serves the hybrid-ARQ retransmissions with the highest priority. If after 4 retransmissions the coded word is still unsuccessfully decoded, another MCS is selected for the new transmission and the number of retransmissions of this SAW channel is reset.

B.3 Simple Traffic Modelling for NRT Services

- *Traffic Model*: The packet call is simply modelled as a single payload whose size is constant and equal to 100 kbytes. The data is fully available in the Node B, waiting to be served by the PS. With such a simple modelling, the connection related and the HSDPA architecture modelling is not necessary (Server, Internet, CN, RNC, and Iub interface). Of special interest is the fact that the lack of RLC modelling challenges the NRT traffic requirement on a very low SDU error rate. To cope with such need, when the maximum number of Hybrid-ARQ retransmissions is reached, and a new coded word transmission is started with a different MCS.

- *Blocking, Dropping and Outage Criterion*: As in the streaming model case, if a new call is created when all the DCH channels are already occupied, the incoming call is blocked. The Node B drops a user if it achieves a user throughput less than 9.6 kbps. The dropping criterion is checked every frame, except for the first 10 seconds of the call. The outage definition is different from the streaming case, i.e., a user contributes to the outage if it is either blocked or dropped or it achieves a data rate less than 64 or 128 or 384kbps. Note the linear weight of the three metrics (blocking, dropping, and percentage of users not achieving the minimum data rate) in the computation of the outage, which is relevant when evaluating the cell capacity at a certain outage.

PARAMETER	ANALYTICAL DESCRIPTION	VALUE	REF.
Simulation step	WCDMA slot level	0.666 msec	[104]
Spatial distribution of users in central cell	Uniformly distributed		
Call arrival rate	Uniformly distributed	According to offered load	[105]
Path loss	Single slope	Exponent = 3.5	[104]
Receiver Noise Power	Thermal Noise	-99.9 dBm	[106]
Shadow fading	Log normal distributed & invariable with time	Std = 8 dB	[105]
Site to site distance	Interference Limited Cell	2.8 km	[105]
Power delay profile	ITU Multipath Model	Default Pedestrian A / Vehicular A	[102]
Mobile speed	Short term mobility	3 km/h	
Number of HS-DSCH codes	≈ 43% of code resources	7	[107]
Maximum cell transmission power		20 Watts	[108]
HS-DSCH power	≈ 50% of Average cell Transmission power	8 Watts	[107]
HS-SCCH	Not simulated	0 Watts	
Own cell non controllable power	Log normal distributed	Average = 5 Watts/ Std = 2dB	
Perch power	10% of Max Tx Power	2 Watts	[106]
Maximum number of associated DCH channels	Maximum number of users in the cell	32	[88]
Maximum power for associated DCH channels		0.6 Watts	
BLER target for associated DCH channels	8 kbps bit rate channel	1%	
Number of HARQ-SAW channels	L1 RTT equals 4 TTIs	4	
Maximum number of HARQ retx		4	

Table B.1: Default System Level Parameters for Quasi-Dynamic Simulations. Propagation Model, Mobility, Transmission Power and Receiver Modelling.

B.4 Simple TCP Protocol Modelling for NRT Services

The network model and the underlying assumptions of the simulation tool employed to evaluate the performance of TCP traffic over HSDPA is described next:

- *Traffic Model*: A very simple modelling of the TCP protocol and higher layers is implemented. A packet call of a typical HTTP file size is assumed to be fully available for TCP to transfer. See in section B.6 a sensibility analysis of the HSDPA capacity with the packet call size. The TCP implementation only includes the slow start and the modelling of the receiver window. It simply assumes that after the reception of a

PARAMETER	ANALYTICAL DESCRIPTION	VALUE
Packet call size	Constant size	100 kbytes
MSS	Headers removed	536 bytes [34]
Rtt	Internet (150 ms) + CN (150 ms)	Default 300 ms [10]
Fixed network segment loss probability	Lossless links	0.0
RLC operation mode		Disabled (section 6.2.1)
TCP receiver memory capabilities	Hard split division to all users	32 kbytes
Initial congestion window		1 MSS
Delayed Ack	Disabled	0 ms
Iub flow control period	Round Robin among all users	5 ms
Number of Iub data transfers per period		4
Bit Rate Averaging Window for PF	~ 10 λ at 3km/h	1.5 sec [76]

Table B.2: Default System Level Parameters for Quasi-Dynamic Simulations. Simple TCP Protocol Modelling for NRT Services.

correct segment acknowledgement, the server opens the congestion window by MSS, as done by the slow start. At any time, the congestion window can not exceed the receiver window. No TCP/IP headers are included in the aforementioned packets. After generated by the server, the TCP segments cross the IP and CN networks (both modelled as a single FIFO with constant delay) and reach the HSDPA network. Once transmitted by the Node B of HSDPA, the TCP segments are reassembled and stored in the UE buffer. As commented 6.2.1, in HSDPA, the RLC retransmissions have a very negative in the overall round trip time suffered by the TCP segments due to the significant Node B. As concluded from section 6.2.1, the maximum number of RLC retransmissions must be very limited to avoid triggering TCP time outs. Due to the limited number of RLC retransmissions, and the low residual error rate from the physical layer, it is decided to simplify the modelling by not including the RLC layer. Note, that to cope with the residual error rate from the physical layer, a new coded word transmission is started with a different MCS after reaching the maximum number of Hybrid-ARQ retransmissions.

- *HSDPA Architecture Modelling*: Only the central cell of the network layout is modelled in very detail. There is a single RNC that serves all links of the central HSDPA cell under study. The RNC merely operates as an infinite FIFO that stores the data packets provided by the CN until the flow control functionality requests the RNC to release the TCP packets. TCP segments are not fragmented into lower layer PDUs in the RNC. The flow control performs a periodic data transfer request from the Node B to the RNC every 5 ms. The RNC is allowed to reply with no more than 4 data transfers (one data transfer for one UE data flow), although there is no limit in the maximum amount of data to be transferred. Note that this Iub flow control modelling does not impose any limitation in terms of instantaneous Iub bandwidth capabilities. The default Node B buffering capabilities equal 10Mbits, and this available memory is

hard split between the maximum number of users to be mapped into HSDPA (32 by default).

- *Blocking, Dropping*: The blocking and dropping criteria are the same as in section B.3.

B.5 Traffic and Architecture Modelling for Streaming Services

The network model and the underlying assumptions of the simulation tool employed to evaluate the performance of CBR encoded streams over HSDPA is described next:

- *Traffic Model*: Only CBR encoded video streams are considered. At first, a server entity generates RTP packets according to the CBRP transmission data trace described in section 2.4.2. Note that the CBRP data trace is assumed to be a fair model of the CBR encoded stream because both traffic flows show a quasi-constant bit rate pattern. The average bit rate of the encoded data stream equals 64 kbps. For 128 kbps video clip simulation cases, the RTP packet size is simply doubled. No RTP/UDP/IP headers are included in the aforementioned packets. After generated by the server, the RTP packets cross the IP and CN networks (both modelled as a single FIFO with constant delay) and reach the HSDPA network. Once transmitted by the Node B of HSDPA, the RTP packets are reassembled and stored in the UE buffer. At the UE only the Dejitter buffer is considered, and no Predecoder or Postdecoder buffers are considered. In case of Dejitter buffer overflow, the incoming data is simply discarded, while in case of Dejitter buffer underflow, no UE rebuffering mechanism is carried out (the situation will simply be counted as outage).

- *HSDPA Architecture Modelling*: The RNC, Node B, and Iub modelling is identical to the one in B.4. Note that no rate control mechanism that ensures the GBR of a certain flow is included in the Iub flow control model. The default Node B buffering capabilities equal 20Mbits, and this available memory is hard split between the maximum number of users to be mapped into HSDPA (32 by default). Note that 10 Mbits of Node B buffering is expected to be a more realistic capability, but the default is selected in order not to impose any limitation to the system throughput for either 64 or 128 kbps bearers. If a RTP packet (or part of it) is stored in the Node B buffer for longer than the discard timer, then it is discarded without notification.

- *Blocking, Dropping and Outage Criterion*: The blocking criterion is the same as in section B.3. The Node B drops a user if more than 20% percent of its RTP packets are discarded (due to overcoming the discard timer). The dropping criterion is checked every frame, except for the first 10 seconds of the call. A user contributes to the outage if it is either blocked or dropped or if it plays out correctly less than 95 % RTP packets of the total video clip.

PARAMETER	ANALYTICAL DESCRIPTION	VALUE
Audio component of PSS	Not considered (section 2.4)	0 kbps
Video clip duration	Truncated exponentially distributed	Average = 28.9 sec, cut off = 1 minute and 15 sec
Video Encoding	Based on CBRP trace (section 2.4.2), [32]	CBR
Default video bearer bit rate	Inherent from CBRP trace (section 2.4.2)	64 kbps
Initial UE Dejitter buffering	60% of total buffering [36]	5 seconds
Predecoder & Postdecoder UE buffers	Not considered (section 7.5)	
Internet + CN delay	Constant	150 ms
RLC operation mode	Section 7.2	Unacknowledged
Buffer size at Node B	Hard split division to all users	20 Mbits
Iub flow control period	Round Robin among all users	5 ms
Number of Iub data transfers per period		4
Node B Scheduler	Default	M-LWDF Algorithm [101]
M-LWDF Bit Rate Averaging Window	~ 10 λ at 3km/h	1.5 sec [76]

Table B.3: Default System Level Parameters for Quasi-Dynamic Simulations. Traffic and Architecture Modelling for Streaming Services.

B.5.1 Considerations about the Outage Criterion

The outage criterion for streaming services defined in previous bullet point is a relevant factor in the present Ph.D. thesis because it is used to define the cell capacity. For streaming services, as described in Chapter 2, the UMTS QoS specification defines the transfer delay for the 95% percentile of the distribution of the delay for all delivered SDUs. The specification relaxes the demands imposed on the network by disregarding the 5% of the SDUs with longest transmission delay. These SDUs could have been transferred under unsatisfactory conditions (e.g. poor instantaneous radio channel conditions due to a received signal fade). The UMTS QoS specification further defines the guaranteed bit rate for streaming bearers but does not specify any percentile for the satisfaction of this attribute.

The service provision in cellular networks under all possible radio propagation conditions can be too costly in terms of radio resource efficiency. The lognormal distribution of the shadow fading can induce poor average propagation conditions for users in the cell edge (see distribution of the G Factor in section B.7). Providing service to such users may require a large amount of radio resources and severely impair the cell capacity. For this reason, the application of a certain outage in the service provision is a typical procedure in the process of radio network planning [106].

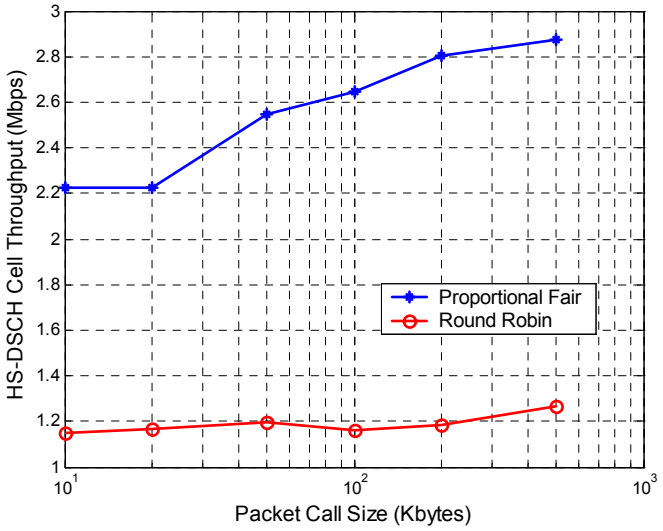


Figure B.2: Influence of the Packet Call Size on the HS-DSCH Cell Throughput for Round Robin and Proportional Fair in Pedestrian A, 3km/h.

Similarly, the investigation in Chapter 7 applies the outage principle in the satisfaction of the transfer delay attribute for streaming bearers. See the sensibility of the cell capacity to the outage level in the results of Chapter 7 (for example Figure 7.7). Due to this strong sensibility, the selection of the actual level of outage is of considerable importance. As it can be seen in [106], an outage value of 5% is commonly used in the process of radio network planning. For this reason, the same level of outage has been selected for Chapter 7, though most of the results of the chapter are also depicted for outage levels up to 2%.

B.6 Performance Sensibility of the Proportional Fair Algorithm with the Packet Call Size

The present section aims at giving a brief overview of the influence of the packet call size on the convergence of the Proportional Fair algorithm.

The Proportional Fair algorithm computes the priority of every user as described in section 5.3.2:

$$P_i = \frac{R_i(t)}{\lambda_i(t)} \qquad i = 1,..,N \tag{B.3}$$

where $P_i(t)$ denotes the user priority, $R_i(t)$ is the instantaneous data rate experienced by user i if it is served by the Packet Scheduler, and $\lambda_i(t)$ is the average user throughput. Section 5.3.2.1 gives a brief description of the averaging alternatives when computing the denominator of equation (B.3).

As described in the previously mentioned section, the performance of the fast scheduling algorithm strongly depends on the window length employed in the throughput averaging

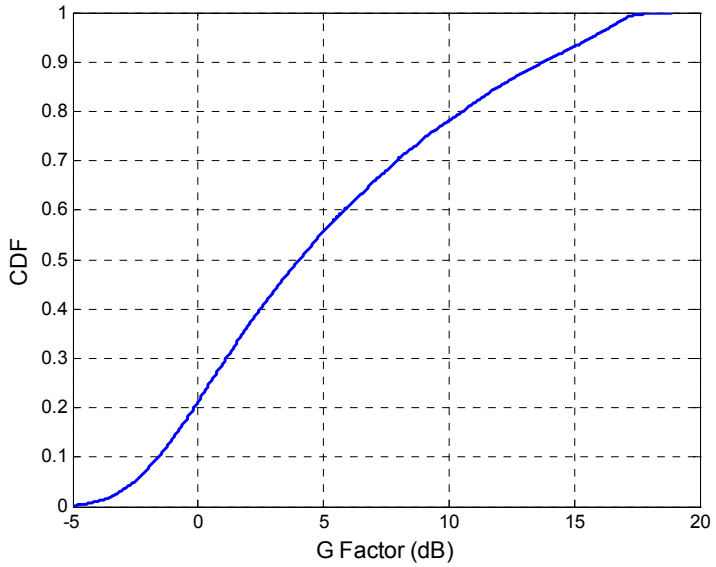


Figure B.3: G Factor Distribution for the System Model and Parameters Described in B.2.

process. However, the initialisation also plays an important role in the efficiency of the scheduling algorithm. If the average user throughput is initialised to zero, the user will enjoy a privileged priority during the initial downloading period until the user throughput stabilizes. This initial convergence period represents a source of inefficiency whose overall weight in the system performance depends on the total download time of the user. The total time required by a user to download its packet call can easily be controlled with the size of the packet call to be downloaded.

Figure B.2 depicts the HS-DSCH cell throughput as a function of the packet call size for Round Robin and Proportional Fair in Pedestrian A at 3km/h. The simulations assume a traffic model as described in section B.3. The user throughput averaging is computed as described in equation (5.5) (which implicitly involves the zero initialisation).

The results clearly show the how the packet call size can diminish the significant effect of the zero initialisation on the performance of the Proportional Fair algorithm. For example, a packet call size increase from 20 to 100 kbytes represents a 19% system throughput improvement.

B.7 Distribution of the G Factor

Figure B.3 plots the distribution of the G Factor (also referred to as the Geometry Factor). This factor is computed as the total received power from the serving base station divided by the received interference from other cells plus thermal noise. The resulting G Factor of Figure B.3 implicitly assumes a hard handover operation. A high value of the G Factor is obtained when the mobile is close to the Node B and, typically, a low value at the cell edge.

Appendix C

Performance Indicator Definitions

[PI 1.1] *Average willingness to pay*.

It is defined as the amount of money to be paid by the end user for the consumption of a certain wireless data service measured in Euro/month. This metric is averaged over all the subjects that the survey is conducted on.

[PI 1.2] *Data rate requirement of 3G application*.

It is defined as the requirement imposed by a given 3G application on the average bit rate capabilities of the end to end connection in order to provide a certain QoS to the end user. This metric is typically measured in kbps.

[PI 1.3] *Peak data rate*.

It is defined as the theoretical maximum bit rate capabilities that the physical layer of a certain technology can provide to a given user. This metric is typically measured either in kbps or Mbps.

[PI 1.4] *Average user throughput*.

Let define $\lambda_i[n]$ as the instantaneous (per transmission time interval) user throughput of user *i*:

$$\lambda_i[n] = \frac{b_i[n]}{TTI} \qquad n = n_1,..n_2$$

where *b* denotes the number of data bits correctly delivered to a certain user by the physical layer of a certain technology in the transmission time interval number *n* during the life time of the user connection (n_1, n_2), and *TTI* represents the transmission time interval duration.

Let define λ_i the user thoughput of user *i* as the time average of the instantaneous user throughput:

$$\lambda_i = \frac{1}{n_2 - n_1 + 1} \cdot \sum_{m=n_1}^{n_2} \lambda_i[m]$$

For a given average cell load and system technology, the user throughput λ_i is a random variable that depends on numerous factors such as the path loss experienced by the user during the interval (n_1, n_2), the shadow fading, the number of remaining users in the cell, the interference conditions, etc. Assuming that the statistical properties of all the users are the same (e.g. equal probability of having a certain path loss or shadow fading), it could be stated that every the user throughput λ_i follows a probability density function f_λ.

The average user throughput is defined as the statistical mean (ensemble average) of the user throughput:

$$\overline{\lambda} = \int_{-\infty}^{\infty} \lambda \cdot f_\lambda \cdot d\lambda$$

This metric is typically measured either in kbps or Mbps.

[PI 1.5] *Network throughput*.

Let define the instantaneous network throughput as the sum of all the individual throughputs of all the existing users in the network. See definition of user throughput in [PI 1.4].

$$\Omega[n] = \sum_{i=1}^{N} \lambda_i[n] \qquad n = 0,...,\infty$$

where $\Omega[n]$ denotes the instantaneous network throughput, and *N* is the total number of users coexisting in the system at the instant *n*.

The instantaneous network throughput $\Omega[n]$ is a stochastic random process. The network throughput Ω is defined as the ensemble average of the instantaneous

network throughput for a given average offered network load.

$$\Omega = E\{\Omega[n]\}$$

In this report this metric will be measured in Mbps, though it could possibly be measure per area unit (i.e. cell or km^2).

[PI 1.6] *Spectral efficiency*.

It is defined as the ratio between the average system (network) throughput over a certain area and the system bandwidth. See definition of network throughput in [PI 1.5]. This metric is typically measured either in Mbps/cell/MHz or Mbps/km^2/MHz.

[PI 1.7] *Service response time*.

This definition applies to data services that follow a request-response pattern such as Interactive services. The service response time is defined as the period elapsed since the end user requests the service (e.g. a web page) until the end of the message reception. This metric is typically measured either in seconds or milliseconds.

[PI 1.8] *Cell throughput*.

Let define the instantaneous cell throughput as the sum of all the individual throughputs of all the existing users in the cell. See definition of user throughput in [PI 1.4]. The cell throughput is defined as the ensemble average of the instantaneous cell throughput for a given average offered cell load.

This metric is typically measured either in kbps or Mbps.

[PI 1.9] *Cell capacity*.

It is defined as the cell throughput when the cell load is such that the blocking rate tends to 100%. See definition of cell throughput in [PI 1.8]. The blocking criterion employed in the HSDPA investigation is described in Appendix B. This metric is typically measured either in kbps or Mbps.

[PI 1.10] *Network capacity*.

It is defined as the network throughput when the network load is such that the blocking rate tends to 100%. See definition of network throughput in [PI 1.5]. The blocking criterion employed in the HSDPA investigation is described in Appendix B. This metric is typically measured either in kbps or Mbps.

[PI 3.1] *RLC throughput*.

The RLC throughput is defined as the number of data bits correctly delivered to a certain user by the RLC layer per time unit, and this metric is averaged over time during the lifetime of the connection. Note that this metric is completely connection specific (i.e. user specific). Note as well, that Chapter 3 assumes an

ideal closed loop and outer loop power control operation, which implies that this metric could be assumed to be approximately independent of the user local area in the cell. This metric is typically measured in kbps.

[PI 4.1] *Raw data symbol energy to interference ratio (EsNo)*.

It is defined as the instantaneous (per Transmission Time Interval) narrowband symbol energy divided by the noise plus interference power after despreading and RAKE combining. In this Ph.D. the EsNo metric is geometrically averaged over all the slots comprised in a TTI. This metric is typically measured in decibels.

[PI 4.2] *Data bit energy to interference ratio (Data bit EbNo)*.

It is defined as the average narrowband decoded data bit energy divided by the noise plus interference power after all physical layer processing including channel coding, physical channel overhead, etc. This metric is typically measured in decibels.

[PI 4.3] *Geometry factor (G Factor)*.

The Geometry factor is defined as the ensemble average of the random process I_{or}/I_{oc} over a local area.

Since the distribution of the random process I_{or}/I_{oc} is unknown, the G Factor is rather estimated. The estimation of the G Factor is based on Monte Carlo simulations where the statistics of the ratio I_{or}/I_{oc} for a given user are collected (i.e. the estimation is based on a realization of the random process I_{or}/I_{oc}). Finally, the estimation of the G Factor is computed by taking a temporal averaging of the I_{or}/I_{oc} collected statistics over the lifespan of the user.

Throughout this Ph.D. thesis, all the allusions to the Geometry Factor refer to the above mentioned estimation by means of temporal averaging.

The G Factor is typically measured in decibels.

[PI 5.1] *First transmission effective throughput*.

It is defined as the number of data bits correctly delivered to a certain user by the physical layer with a given MCS for a given instantaneous EsNo per TTI. This metric is statistically averaged by taking the ensemble average. The First Transmission Effective Throughput can be computed as PDR x (1-BLER), where PDR represents the peak data rate for a given MCS and BLER is the Block Erasure Rate. This metric is typically measured either in kbps or Mbps.

[PI 5.2] *HS-DSCH cell throughput*.

Let define the instantaneous cell throughput as the sum of all the individual instantaneous throughputs of all the existing users mapped onto the HS-DSCH in the cell. See definition of user throughput in [PI 1.4]. The cell throughput is

defined as the ensemble average of the instantaneous cell throughput for a given average offered cell load. This metric is typically measured either in kbps or Mbps.

[PI 6.1] *Average Node B buffering for a UE data flow*.

Let define the instantaneous Node B buffering $B_i[n]$ of the user *i* as the amount of data bits belonging to the user *i* stored by the MAC-hs buffer. Note that $B_i[n]$ represents a random process with multiple possible realizations.

The average Node B buffering for the user *i* is defined as the ensemble average of the random process $B_i[n]$:

$$B_i = E\{B_i[n]\}$$

This metric is typically measured kbytes.

[PI 6.2] *Average Node B queueing delay per user*.

Let define the queueing delay D_i^n of the packet *n* of the user *i* as the time elapsed since the packet arrives to the MAC-hs queue until the instant that the last bit of the packet abandons the queue.

In steady-state, it can be assumed that the statistical properties of the queueing D_i^n of the packet *n* of the user *i* are equal to the properties of the queuing delay for any other packet of the same UE data flow. Note that, for TCP based flows, this assumption only holds if the properties of the connection do not vary for the different packets (e.g. if the congestion window remains about constant).

Let f_{Di} denote the probability density function of the queuing delay of all the packets of the user *i*. Then, the average queuing delay for a UE data flow is defined as:

$$\overline{D_i} = \int_{-\infty}^{\infty} D_i \cdot f_{Di} \cdot dD_i$$

This metric is typically measured either in seconds or milliseconds.

[PI 6.3] *Average Node B buffering*.

Let $B[n]$ be the instantaneous Node B buffering that equals the sum of instantaneous Node B buffering $B_i[n]$ of all the users existing in the cell at instant n.

$$B[n] = \sum_{i=1}^{N} B_i[n]$$

where *N* is the total number of users coexisting in the system at the instant *n*. The average Node B queueing is defined as the ensemble average of the

instantaneous Node B buffering:

$$B = E\{B[n]\}$$

This metric is typically measured kbytes, Mbytes, or Mbits.

[PI 6.4] *Median user throughput*.

Let f_λ be the probability density function of the user throughput for all the users in the cell. Note in [PI 1.4] that the user throughput is a random variable that depends on random factors such as the location of the user in the cell. The median user throughput is the median of the probability density function f_λ. This metric is typically measured either in kbps.

[PI 6.5] *Average Node B queueing delay*.

Let define Node B queuing delay D^n of any packet as the time elapsed since the packet arrives to the MAC-hs queue until the instant that the last bit of the packet abandons the queue. The average Node B queuing delay is the ensemble average of the Node B queuing delay of any packet D^n over all possible packets.

$$D = E\{D^n\}$$

This metric is typically measured either in seconds or milliseconds.

[PI 6.6] *Mean number of alive users*.

A user is defined to be alive when it is mapped on the HS-DSCH and it occupies one associated DPCH channel, i.e. when the user has been admitted in the system regardless of whether he has or not data queued in the Node B to transmit. The number of alive users in the cell $A[n]$ is a random process.

For a given load, the mean number of alive users in the cell is defined as the ensamble average of the random process $A[n]$:

$$A = E\{A[n]\}$$

[PI 7.1] *Maximum queuing plus transmission delay for a UE data flow*.

Let define the queueing plus transmission delay Ψ_i^n of the RTP packet n of the user i as the time elapsed since the packet arrives to the MAC-hs queue until the instant that the last bit of the packet is successfully transmitted.

Let ψ_i^n denote the sample of the random variable Ψ_i^n for the RTP packet n of the user i. The maximum queuing plus transmission delay for the UE data flow m_i is defined as the maximum of all the sampled queueing plus transmission delays over all RTP packets of the UE data flow i.

$$m_i = \max_n \left\{ \psi_i^n \right\}$$

It is interesting to observe that the maximum queuing plus transmission delay for a UE data flow is completely dependent on the specific characteristics of the user (such as experienced path loss, shadow fading, etc). Figure 7.4 a) depicts the distribution of m_i over all data flows.

This metric is typically measured either in seconds or milliseconds.

[PI 7.2] *Average queuing plus transmission delay for a UE data flow*.

For CBR flows, in steady-state, it can be assumed that the statistical properties of the queueing plus transmission delay Ψ_i^n of the RTP packet n of the user i (see [PI 7.1]) are equal to the properties of the queuing plus transmission delay for any other RTP packet of the same UE data flow. Let $f_{\Psi i}$ denote the probability density function of the queuing plus transmission delay of all the packets of the user i. Then, the average queuing plus transmission delay for a UE data flow is defined as:

$$\overline{\Psi_i} = \int_{-\infty}^{\infty} \Psi_i \cdot f_{\Psi i} \cdot d\Psi_i$$

This metric is typically measured either in seconds or milliseconds.

[PI 7.3] *Average number of discarded packets in the Node B*.

Let define ζ_i as the the total number of RTP packets of the data flow of user i that reach the MAC-hs functionality of the Node B in order to be transmitted through the radio interface. Let define ζ_i^* as the number of RTP packets of the data flow of user i that are discarded in the MAC-hs functionality of the Node B. The average number of discarded packets in the Node B ζ is defined as:

$$\zeta[n] = \frac{\sum_{i=1}^{N} \zeta_i}{\sum_{i=1}^{N} \zeta_i^*} \qquad n = 0,..,\infty$$

where N the overall number of users that have ever been mapped into the HS-DSCH until the instant n.

[PI 7.4] *Overall average queuing plus transmission delay for a UE data flow*.

As defined in [PI 7.2], the average queuing plus transmission delay for the user i is a random variable that depends on depends on numerous factors such as the path loss experienced by the user during its lifetime, the shadow fading, the number of remaining users in the cell, the interference conditions, etc. The overall average queuing plus transmission delay for a UE data flow is defined

as the ensemble average of $\overline{\Psi_i}$ over all users.

$$\overline{\Psi} = E\left\{\overline{\Psi_i}\right\}$$

This metric is typically measured either in seconds or milliseconds.

[PI 7.5] *Maximum Node B buffering for a UE data flow*.

Let $b_i[n]$ denote the realization $B_i[n]$ actually experienced by user *i*. See the definition of the instantaneous Node B buffering for a UE data flow $B_i[n]$ in [PI 6.1]. The maximum Node B buffering for the UE data flow b_i is defined as the maximum of the realization experienced by user *i*:

$$b_i = \max_n \{b_i[n]\}$$

This metric is typically measured either in kbytes.

Appendix D

Reliability and Statistical Significance of the HSDPA System Level Simulation Results

Section 1.6 argued the need for computer aided system level simulations as the assessment methodology for the HSDPA technology evaluation. The requirement of absolute cell capacity figures under realistic conditions for the present Ph.D. investigation imposes some challenges on the accuracy of the simulation results. Here, the reliability and the statistical significance of the HSDPA system level simulation results are discussed in the following two sections.

D.1 Reliability of the System Level Simulation Results: A Comparative Evaluation

There exists some publications in the available literature that provide absolute HSDPA cell capacity figures based on computer aided simulation tools. All these publications do not treat the concept of QoS, and their network operation only provides a best effort service. Among all of them, the network scenario studied in [11] is the most similar to the one considered in the present Ph.D. investigation. Table D.1 gives a description of the simulation assumptions in [11] and the resulting cell capacity figures.

Simulation assumptions in [11]	
HS-DSCH power	75% of Node-B power
Number of HS-DSCH codes	15 (SF = 16)
HS-DSCH cell coverage	90%
Maximum number of associated DCH channels	32
Mobile speed	3km/h
Packet call size	400 kbits
Node B Peak Code Domain Error (PCDE)	-36 dB (SF = 256)
HSDPA Cell capacity in [11] **(ITU vehicular A scenario)**	
Round Robin	1.4 Mbps
Proportional Fair algorithm	2.1 Mbps

Table D.1: Simulation Assumptions and HSDPA Cell Capacity Results in [11].

As described in section 8.4, the results of the network level simulator employed in this Ph.D. investigation yield in macro cell (Vehicular A) scenario a maximum cell capacity of around 1.44 Mbps and 850 kbps for the Proportional Fair and Round Robin algorithms respectively. Note that the main differences of the network set-up considered in [11] with the one described in the Appendix B are the HS-DSCH power allocation, the number of available HS-DSCH codes, the packet call size and the cell coverage.

A reasonable estimation of the cell capacity to be obtained under the network simulation model described in Appendix B, but with a HS-DSCH power allocation of 75% of total Node B power, could be achieved by simply rescaling the results given in section 8.4 relative to a larger amount of power (i.e. 75% of the total Node power). This rescaling can be assumed as a fair approximation in a power limited environment if the number of HS-DSCH codes is sufficient. Such a rescaling leads to a maximum cell capacity of approximately 2.2 Mbps and 1.3 Mbps for the Proportional Fair and Round Robin algorithms respectively. These results closely match the HSDPA cell capacity figures provided in [11]. However, in [11] the HS-DSCH is assumed to give coverage to 90% of the cell area, which is expected to provide higher cell capacity than a full HS-DSCH coverage set-up. On the contrary, the lower packet call size and the Node B hardware imperfections in [11] are likely to reduce the cell capacity. Though it is difficult to quantitatively estimate these influences without further system level simulations, it is expected that these effects partially cancel each other.

Summarizing, the reasonably good agreement between the HSDPA network set-ups and the cell capacity figures of this Ph.D. investigation and of reference [11] provides a source of reliability on the accuracy of the network level simulation tool developed for this Ph.D. investigation.

D.2 Statistical Significance of the HSDPA System Level Simulation Results

The statistical significance of an experiment result gives the probability that the outcome occurred by pure chance, and that in the considered physical phenomenon no such a result exists. More specifically, the statistical significance represents the probability of error that is involved in accepting the observed result as valid, that is, as representative not only of our sampled statistics, but the entire physical phenomenon. Obviously, the larger the set of sampled statistics of the given physical phenomenon, the lower it is the probability of committing such an error.

Here, a simple model for evaluating the statistical significance of the HSDPA system level results will be used. The analysis will be applied to the HSDPA system level results that compare the HS-DSCH cell throughput versus the percentage of users that incur outage. Such results are depicted in Figure 5.9- 5.11 of Chapter 5, and in Figure 7.7, 7.10 and 7.11 of Chapter 7.

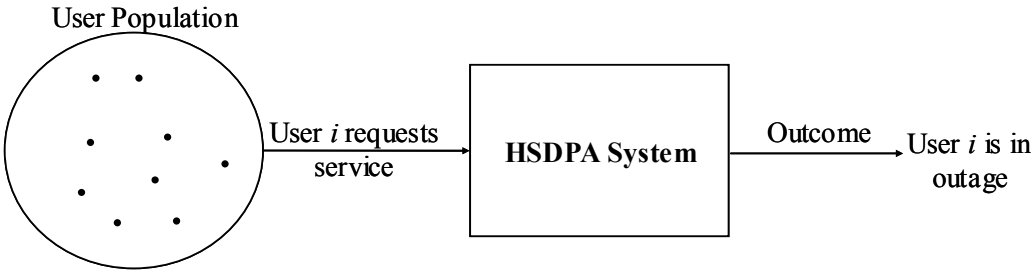


Figure D.1: Evaluation Model of the Statistical Significance.

The simple model for evaluating the statistical significance of the simulation results will assume a population of N users that are to request service to the HSDPA system under consideration. See the evaluation model Figure D.1. In this model, the HSDPA system will be considered as a "black box" that may or may not provide an acceptable service to every requesting user. If the system does not provide an acceptable service to a requesting user, this user will be considered to be in "outage". See in Appendix B the outage definition for the analysis carried out in Chapter 5 and 7. Let assume that the probability that a user that is requesting service will not experience an acceptable service is equal to p (i.e. the outage probability is equal to p). Let further assume that the provision of a successful service to a certain user is an event that is independent and identically distributed to the outcome of the service provision to any other user of the entire population.

Let X be the random variable that counts the total number of users that incur outage from the overall user population. Under the previous assumptions, it can be stated that X follows a binomial probability density function whose mean is:

$$\bar{x} = E\{X\} = N \cdot p \tag{D.1}$$

and whose variance equals:

$$\sigma_x^{\,2} = N \cdot p \cdot (1 - p) \tag{D.2}$$

Then, the standard deviation (relative to the mean) of the variable that counts the total number of users that incur outage is equal to:

$$\frac{\sigma_x}{\bar{x}} = \sqrt{\frac{1-p}{N \cdot p}} \tag{D.3}$$

The simulation campaigns carried out for the above mentioned results comprised a population of users of 4000 users per simulation. Every single point in the HS-DSCH cell throughput curves is the result of a cumulative process from three different realizations obtained from simulations with different seeds. Hence, the overall population for every single point in the HS-DSCH cell throughput curves comprised 12000 users. Note that the most exigent level of outage in the above mentioned HS-DSCH cell throughput figures typically was 2%.

For an outage probability of 2%, the relative error (standard deviation relative to the mean) of the variable that counts the number of users in outage equals 6% for an overall population of 12000 users. A relative error of 6% for such an exigent outage probability could be considered as an acceptable accuracy. For an outage probability of 5%, the relative error (standard deviation relative to the mean) of the variable that counts the number of users in outage equals 4% for an overall population of 12000 users.