Traffic Engineering Concepts for Cellular Packet Radio Networks with Quality of Service Support

Von der Fakultät für Elektrotechnik und Informationstechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Ingenieur
Peter Stuckmann

aus Eslohe

Berichter: Universitätsprofessor Dr.-Ing. Bernhard Walke
Universitätsprofessor Dr. Petri Mähönen

Tag der mündlichen Prüfung: 23. Juni 2003

ABSTRACT

Cellular packet radio networks are presently introduced and further developed to support mobile Multimedia applications and mobile Internet access. The *General Packet Radio Service* (GPRS) and *Enhanced GPRS* (EGPRS), the packet services of the *Global System for Mobile Communications* (GSM), are dominant standards for packet data support in evolved second-generation cellular systems.

This thesis presents traffic engineering concepts for cellular packet radio networks that are applied but not limited to the standards GPRS and EGPRS and can be integrated into the radio interface capacity planning process. They are based on a comprehensive performance evaluation of Internet and Multimedia applications for the relevant cell scenarios, system parameter settings and protocol options that are expected for the evolution of GPRS and EGPRS networks. Additionally, optimized algorithms for traffic management and quality of service management, comprising connection admission control as well as scheduling, are developed. Their performance is evaluated and compared to existing implementations. From the traffic performance analysis, based on the existing and optimized system models, traffic engineering rules are derived. They can be used for quantitative radio interface dimensioning by network operators, equipment vendors or system integrators. Examples for the application of the traffic engineering rules are presented for GPRS evolution scenarios and an EGPRS introduction scenario. While the examples are based on the existing standards GPRS and EGPRS, the concepts developed in this thesis are valid for all packet radio networks.

The traffic performance evaluation of mobile packet data services is a complex problem, because non-Poissonian traffic characteristics, dynamic behavior of lower layer protocols and the effects of the radio channel have to be considered. The author has solved this problem by the development of the emulation tool GPRSIM. This comprises the implementation of the GPRS and EGPRS protocols and load generators for typical GPRS/EGPRS usage. The tool allows to study the existing and evolved systems in their natural environments with the appropriate radio coverage, mobility and typical traffic volumes and permits to analyze own approaches in the improvements and introduction of new features.

KURZFASSUNG

Zellulare Paketfunknetze befinden sich in der Einführung und werden weiterentwickelt, um mobile Multimedia-Anwendungen und den mobilen Internetzugang zu ermöglichen. *General Packet Radio Service* (GPRS) und *Enhanced GPRS* (EGPRS), die Paketdatendienste des *Global System for Mobile communication* (GSM), stellen die dominierenden Standards zur paketorientierten Datenübertragung in weiterentwickelten zellularen Mobilfunksystemen der zweiten Generation dar.

Diese Arbeit stellt Dimensionierungskonzepte für zellulare Paketfunknetze vor, die anwendbar aber nicht beschränkt sind auf die Standards GPRS und EGPRS und die in den Funknetzplanungsprozess integrierbar sind. Sie basieren auf einer umfassenden Leistungsbewertung von Internet und Multimedia-Anwendungen für die relevanten Zellszenarien, Systemeinstellungen und Protokolloptionen, die für die Evolution von GPRS und EGPRS erwartet werden. Zusätzlich werden optimierte Algorithmen zur Verkehrs- und Dienstgüteverwaltung entwickelt, die sowohl Abfertigungs- als auch Verbindungsannahmestrategien umfassen. Ihre Leistungsfähigkeit wird bewertet und mit bereits existierenden Implementierungen verglichen. Ausgehend von der Leistungsanalyse auf Basis der Modelle für existierende und für die optimierten Systeme werden Dimensionierungsregeln abgeleitet, die von Netzbetreibern, Infrastrukturanbietern oder Systemintegratoren für die quantitative Dimensionierung von Funkschnittstellen herangezogen werden konnen. Die Anwendung der Dimensionierungsregeln wird anhand von Beipielen für ein GPRS-Evolutionsszenario und ein EGPRS-Einführungsszenario veranschaulicht. Obwohl die Beispiele auf den konkreten Standards GPRS und EGPRS basieren, behalten die in dieser Arbeit entwickelten Konzepte für alle Paketfunknetze ihre Gültigkeit.

Die Leistungsanalyse von mobilen Paketdatendiensten stellt ein komplexes Problem dar, weil nicht-Poissonsche Verkehrscharakteristiken, das dynamische Verhalten unterer Schichten und die Effekte des Funkkanals modelliert werden müssen. Der Autor hat dieses Problem mit der Entwicklung des Emulationswerkzeugs GPRSim gelöst, welches sowohl die Implementierung der GPRS- und EGPRS-Protokolle als auch Lastgeneratoren für die Nutzung von GPRS/EGPRS umfasst. Das Werkzeug erlaubt die Untersuchung der existierenden und weiterentwickelten Systeme in ihren natürlichen Umgebungen mit der zugehörigen Funkabdeckung, Mobilität und dem typischen Verkehrsaufkommen. Es schafft außerdem die Grundlage für die Analyse eigener Ansätze zur Verbesserung und zur Einführung neuer Funktionen.

CHAPTER 1

Introduction

With the development of the *Global System for Mobile communication* (GSM) standard for digital cellular mobile radio networks in the late 1980s in Europe and their introduction in the 1990s a new mass market with several million subscribers worldwide has been created. Besides the growth of the subscriber numbers, the technological evolution of GSM is going on. New services and applications have been developed and standardized and are presently integrated into GSM networks. Though its principal use is for mobile telephony, mobile data services are becoming more and more popular. The success of text messaging in Europe and the growth in both subscriber numbers and the usage of the i-mode service in Japan, which enables the delivery of Internet-like content on mobile phones, has led to very high market potential estimations for packet-oriented mobile data services. Licenses for *Third-Generation* (3G) systems, based on *Universal Mobile Telecommunication System* (UMTS), which are seen as the successor of GSM-based *Second-Generation* (2G) systems, were granted in 1999–2002 in Europe to offer the needed radio capacity for data services with higher data rates and for enhanced speech services.

UMTS is aiming to realize peak bit rates of up to 144 kbit/s with wide coverage, up to 384 kbit/s in hotspots, and up to 2 Mbit/s in indoor scenarios. A stepwise way in the direction of 3G, however, will already be performed by the extension and the further development of existing cellular systems. The advantage of such an evolution process is the faster availability of such services, since the infrastructure of existing 2G systems can be used. Furthermore, the opportunity is given to prepare the customers for new, so-called 3G services.

First packet-switched services based on the *General Packet Radio Service* (GPRS) have been available in Europe since 2001 and many countries worldwide will introduce GPRS in the next years. Due to this new service, mobile data applications with peak bit rates of up to 117 kbit/s and typical user data rates of 25–64 kbit/s are offered and established on the market. To realize higher data rates the *European Telecommunications Standards Institute* (ETSI) has developed the *Enhanced Data rates for GSM Evolution* (EDGE) standard. The packet-oriented part *Enhanced General Packet Radio Service* (EGPRS) offers a peak bit rate of up to 384 kbit/s and typical user data rates of 40–100 kbit/s by means of modified modulation, coding and medium access schemes.

1.1 Objectives

Whereas just after the service introduction minimal cell configurations were chosen supporting only a basic availability of GPRS, with increasing data traffic load during the next years GSM/GPRS cell capacity will have to be extended. For these evolution scenarios traffic engineering guidelines are required. They should describe the relationship between the offered traffic and the radio resources that have to be allocated to reach a desired quality of service for the different applications.

For traffic engineering in circuit-switched networks the Erlang-B-Formula has been successfully applied over decades, while for packet-switched cellular radio networks such an applicable traffic engineering model is still missing. The analytical description of statistical multiplexing and Internet and Multimedia traffic modeling are more complex than for circuit-switched networks. Although results for packet multiplexer systems are available, e.g., for *Asynchronous Transfer Mode* (ATM) networks, the specifics of the radio interface have not been included into the models so far.

The goal of this thesis is the development of traffic engineering rules for cellular packet radio networks based on GPRS and EDGE. They are based on traffic models for typical mobile applications. Load generators, representing these traffic models, are developed and integrated into a simulation environment with the prototypical implementation of the EGPRS protocols and models for the radio channel, which were also developed in the framework of this thesis. With this simulation tool a comprehensive performance evaluation is carried out that leads to the traffic engineering rules.

Additionally, quality of service management algorithms are developed for best-effort scheduling and traffic class support to be deployed in the base station RLC/MAC layer of cellular packet radio networks. The performance of the proposed integrated scheduling schemes is evaluated, basing on the example of the EGPRS standard. These schemes can be implemented in GPRS and EGPRS networks without changing the standard and can also be employed in other cellular packet radio networks.

1.2 Outline

In Chapter 2 an overview of the requirements for the evolution of cellular packet radio networks is given. After the introduction of the existing 2G and 3G standards and the related evolution paths, the requirements for the cost-effective evolution of these networks are discussed, namely the QoS provisioning of different traffic classes ensured by traffic engineering and traffic management.

In Chapter 3 the GPRS design approach, the extension of the GSM logical architecture for GPRS and the GPRS radio interface, is presented. The protocols for data transmission are explained in a top-down approach. Starting with GTP and BSSGP, the protocols of the GPRS air interface (SNDCP, LLC, RLC/MAC, PLL) are described in detail. Protocol functions, state machines, data structures and algorithms are introduced.

The GPRS control plane, which realizes control functions, namely mobility management, session management and QoS management, is also discussed. Starting with the GPRS mobility management, it is explained how the system keeps track of the mobile station's location and its state, how databases are updated with this information and how cell change procedures are supported. In the section that deals with session management, the logical context between the mobile station and external packet data networks based on TCP/IP or X.25, is introduced. Subsequent to this, QoS management in GPRS networks is explained. Here it becomes clear how the network establishes QoS contracts with the MS, differentiating applications and subscribers according to their QoS requirements, and how these contracts are complied by the network.

Chapter 4 discusses the enhancement of GSM/GPRS by EDGE functionality at the air interface. The new modulation and coding schemes are introduced, followed by the protocol extensions *Link Adaptation* (LA) and *Incremental Redundancy* (IR).

Chapter 5 presents typical existing and predicted applications and the traffic characteristics together with the related service platforms. After an overview of existing Internet applications like WWW, e-mail and FTP, applications developed for mobile devices like WAP-based applications including the MMS are explained. Following this, future applications like Audio and Video Streaming as well as Voice over IP are described and related traffic models are proposed.

In Chapter 6 the emulation tool used for the presented traffic engineering analysis is described. After the description of the model architecture the different modules are explained. First the traffic generators for data and voice traffic are presented. Then the protocol implementations are depicted and the channel and mobility models are introduced. The tools for graphical presentations and the graphical user interface are described as well. Finally a validation of the tool is presented comprising an analytical estimation of the TCP performance for typical WWW and e-mail object sizes in a simplified GPRS scenario and traffic performance measurements in a real operational GPRS network.

In Chapter 7 the applicability of analytical concepts for the traffic engineering of cellular packet radio networks is discussed. The prerequisites of this have been developed in the framework of this thesis from analytical models for fixed packet networks available in literature. Analytical approaches in other publications will also be regarded. By comparison of analytical results with the results gained by the GPRS emulator, presented in Chapter 6, it becomes clear that analytical models of cellular packet networks as known today, allow a fast estimation of the general system capabilities. On the other hand it is shown that they are limited in their suitability for quantitative packet radio interface traffic engineering, when complex scenarios with heavy-tailed traffic sources, TCP influence and radio interface-specific mechanisms like scheduling and link adaptation have to be considered.

The comprehensive traffic performance analysis, based on the emulation tool GPRSim, is presented in Chapter 8. Here, the traffic performance of Internet applications over GPRS and EGPRS regarding fixed and on-demand channels as well as different protocol options and channel conditions are presented.

In Chapter 9 the traffic performance of different evolving applications like WAP and Streaming in comparison and in traffic mix scenarios are discussed.

Methods and techniques for traffic engineering and traffic management are presented in Chapter 10. Reverting to the results in Chapter 8 and 9, traffic engineering rules are derived for GPRS introduction and evolution scenarios with fixed and on-demand PDCHs. Next, traffic management techniques are introduced that were developed in the framework of this thesis, comprising scheduling and admission control algorithms for best-effort services and traffic class support. The performance of these algorithms is evaluated. Finally the influence of traffic management on traffic engineering rules is discussed for an EDGE introduction scenario.

The conclusions in Chapter 11 sum up the main insights and contributions of this thesis.

Evolution of Cellular Radio Networks

Since the introduction of digital cellular radio networks in the 1990s, which were the successors of analog cellular telephone systems, the subscriber numbers and usage of mobile telephony services have grown exponentially worldwide. This step from *First-* to *Second-Generation* (2G) mobile radio networks has led to a number of incompatible systems for different regions.

The European standard GSM is based on the *Time Division Multiple Access* (TDMA) concept and uses eight time slots on 200 kHz carrier frequencies. GSM is operating at 900 MHz and 1800 MHz in Europe and in other countries worldwide and at 1900 MHz in America.

In the US two additional groups of systems have been put in place: the TDMA systems IS-54 (also called D-AMPS) and the further development IS-136 (Digital PCS) are in part similar to the GSM, while IS-95 (cdmaOne) is based on *Code Division Multiple Access* (CDMA). With a carrier frequency bandwidth of 1.23 MHz, IS-95 systems are also called *Narrowband-CDMA* (N-CDMA) systems. IS-95 is also used in South America, Africa and Asia.

Another standard was developed in Japan. *Personal Digital Cellular* (PDC) is based on TDMA and works with three time slots per 25 kHz channel at 800 MHz and 1500 MHz.

All these systems were designed to support speech services with data rates of 5–15 kbit/s. With the rapid increase of the Internet traffic volume in telecommunication networks and the success of the mobile data services *Short Message Service* (SMS) in Europe and i-mode in Japan, a growing demand for mobile data services has been predicted. Against this background *Third-Generation* (3G) cellular radio networks, which were standardized in the 1990s, will be introduced in the next few years.

2.1 The IMT-2000 Family of Systems

The *Internationl Telecommunication Union* (ITU) has defined requirements for the development of 3G systems. The main focus is the support of data services with high data rates (144 kbit/s outdoor and 2 Mbit/s indoor), efficient support of asymmetric traffic, packet-switched transmission at the air interface and high spectrum efficiency [ETSI (1996)]. To enable *Global Multimedia Mobility* (GMM), a family of compatible systems should be defined, which results from the evolution of the existing incompatible systems. Six out of 15 proposals were accepted for the *International Mobile Telecommunications at 2000 MHz* (IMT-2000) family of systems, which can be grouped into four categories:

W-CDMA: *Wideband Code Division Multiple Access* (W-CDMA) systems comprise the *Frequency Division Duplex* (FDD) component of the *Universal Mobile Telecommunication System* (UMTS) and the US standard *cdma2000*.

TD-CDMA: The *Time Division—Code Division Multiple Access* (TD-CDMA) group of systems contains the *Time Division Duplex* (TDD) component of UMTS and the Chinese *Time Division—Synchronous Code Division Multiple Access* (TD-SCDMA) system, which is now integrated in the UMTS-TDD mode.

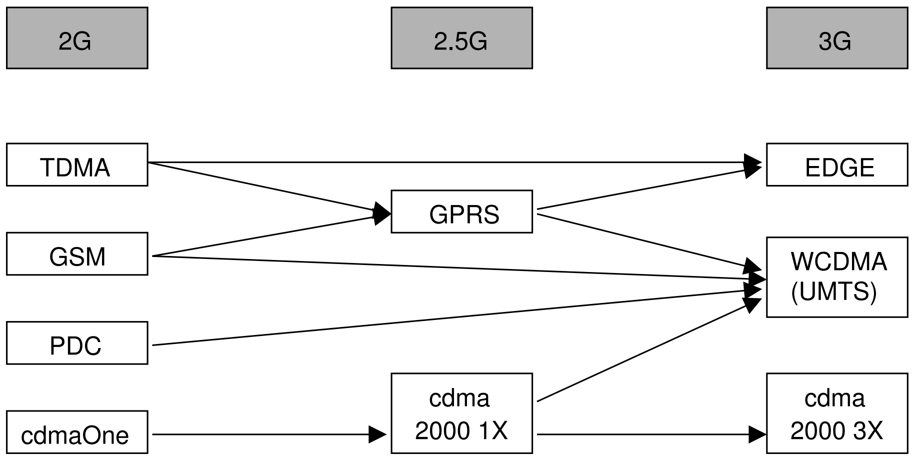


Figure 2.1: Possible migration paths from 2G to 3G systems

TDMA: The *Enhanced Data rates for GSM Evolution* (EDGE) concept, the further development of GSM and IS-136, has been chosen under the name *Universal Wireless Communications* (UWC)-136 to be accepted for the IMT-2000 family.

FD-TDMA: *Frequency Division—Time Division Multiple Access* (FD-TDMA) is the further development of the European *Digital Enhanced Cordless Telecommunications* (DECT) standard for cordless telephony and has been accepted, too.

From these alternatives of IMT-2000 systems for 3G solutions follows that no unique standard can be expected worldwide. To realize global multimedia mobility, one important goal of IMT-2000, dual-mode or even multi-mode terminals will be needed always to be connected independently from the home-used standard of a customer. In addition, 3G terminals for a long duration will require the implementation of a 2G air interface, typically GSM/GPRS, to be able to connect independently from the roll-out status of 3G systems.

For the migration path from 2G cellular radio networks to 3G systems different scenarios are possible (see Figure 2.1). In Europe, UMTS will be introduced as the successor of GSM, while *GSM/EDGE Radio Access Networks* (GERANs) have the potential to complement *UMTS Terrestrial Radio Access Networks* (UTRANs) in areas with small subscriber density [Halonen et al. (2002); Stuckmann (2002b)].

Initially, the ETSI has been given the responsibility for producing the UMTS and the evolved GSM standards. In 1999 the *3rd Generation Partnership Project* (3GPP) was formed to coordinate the efforts of the standardization institutes ETSI (Europe), *Association of Radio Industries and Businesses* (ARIB) (Japan), *Telecommunication Industry Association* (TIA) (USA), *Chinese Wireless Telecommunications Standards* (CWTS) (China), *Telecommunication Technology Association* (TTA) (Korea) and *Telecommunication Technology Committee* (TTC) (Japan).

The 3GPP is organized in several *Technical Specification Groups* (TSGs) responsible for *Core Network* (CN), *Services and System Aspects* (SSA), *Radio Access Network* (RAN), *Terminal Aspects* (TA), and *GSM/EDGE Radio Access Network* (GERAN).

For the development of cdma2000 standards the *3rd Generation Partnership Project 2* (3GPP2) was formed out of the standardization bodies ARIB, TIA, and CWTS.

2.2 Requirements for Cost-effective Network Evolution

The success of evolving cellular packet radio networks is heavily dependent on their ability to offer a range of services in an economic and reliable way. Essential requirements are traffic management mechanisms, capable of providing necessary *Quality of Service* (QoS) support, and traffic engineering procedures, which ensure that these mechanisms are used in a cost-effective way. It is increasingly recognized that traffic management and traffic engineering are complementary and mutually dependent functions. While various scheduling and admission control algorithms provide service differentiation, it remains to specify how resources have to be adequately engineered to meet quantitative quality objectives. In the following, the two essential requirements for the evolution of cellular packet radio networks, namely traffic engineering and traffic management, are discussed, presenting the problem, the solution ideas and the methodology.

2.2.1 Radio Interface Traffic Engineering

Traffic engineering procedures ensure that the network is designed and upgraded in a cost-effective way. They should be based on the *traffic-performance* relation, linking network capacity, traffic demand and realized performance, and should assure that the network has sufficient capacity to handle the offered traffic.

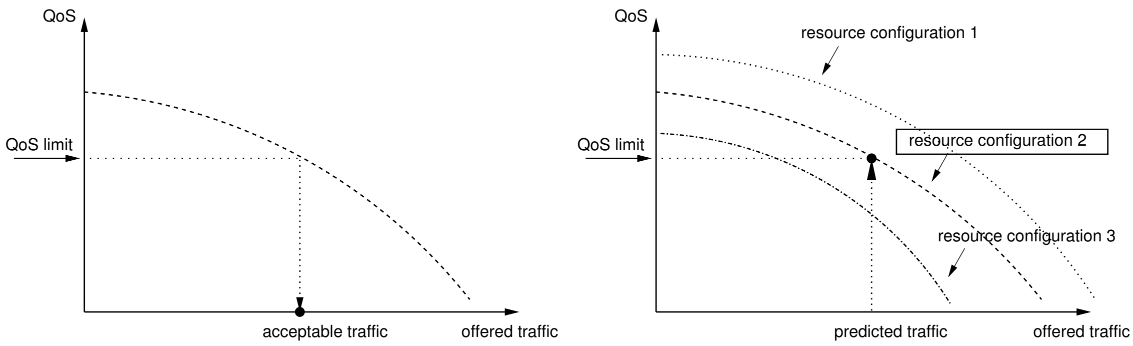
Traffic engineering is fundamental to the design of circuit-switched networks like the telephone network. The traffic-performance relation here is given by the Erlang loss formula which gives the probability of call blocking when a certain volume of traffic is offered to a given number of circuits [Kleinrock (1975)].

For packet-switched networks the objective of theoretical or simulation studies is to define simple network engineering procedures like applying the Erlang formula in circuit-switched networks [Roberts (2001)].

For mobile packet data services like GPRS the amount of radio resources available for packet data traffic is most critical for the realized performance. The resources can be defined by a number of fixed or on-demand *Packet Data Channels* (PDCHs) to be allocated for GPRS in the case of sufficient capacity for both speech and packet data traffic or by the number of additional transceivers. These have to be installed in the existing GSM base stations, if the traffic demand for speech and data is exceeding the acceptable traffic that can be carried by the existent hardware.

To achieve the goal of defining simple network engineering procedures that are usable in practice, dimensioning graphs for application-specific performance measures are proposed that are valid for the cell and load scenarios of interest. With these graphs, describing the traffic-performance relation, the offered traffic that can be served under a given number of radio resources (see Figure 2.2(a)) or the number of resources, necessary for a given offered traffic (see Figure 2.2(b)), can be estimated.

For interactive download-oriented applications like *World Wide Web* (WWW) and e-mail the throughput performance is of interest. For transaction-oriented applications with small objects like WAP the application response time is perceived by the user. For Streaming applications, however, the delay and delay variation is critical. Since in GPRS networks no strict QoS guarantees are supported and no delay-critical applications will be introduced initially, traffic engineering rules will be based on mean values first of all. For the network evolution like the introduction of streaming applications, the introduction of radio interface enhancements or for traffic engineering of related packet radio networks, the same methodology might be applied in using stricter QoS measures like throughput or delay quantiles.



(a) Estimation of capacity (b) Estimation of needed resources

Figure 2.2: Dimensioning graphs

To ensure the applicability in practice, the traffic engineering rules themselves should be simple and only based on the user number and traffic volume during the busy hour. However, the dimensioning graphs themselves should be taken from accurate models for the protocol stacks, the traffic pattern and the radio channel as close as possible to reality.

2.2.1.1 Related Work

Several papers concerning GPRS performance analysis have been published in the last years [Brasche and Walke (1997); Cai and Goodman (1997); Kalden et al. (2000)].

In [Brasche and Walke (1997)] potential applications and elementary concepts of GPRS are described and simulation results for system throughput, frame transfer delay and blocking rate are presented. For traffic load generation the respective models suggested by ETSI for the performance evaluation and optimization of proposed protocols during the standardization of GPRS like FUNET, Railway and Mobitex [Walke (2001)] were applied. Internet traffic was not taken into account.

The same ETSI traffic models were used in [Cai and Goodman (1997)] for the performance evaluation of GPRS where performance and system measures were presented, based on simulation with a focus on the comparison of multislot vs. single slot operation and on traffic model effects.

In [Kalden et al. (2000)] WWW applications over GPRS are examined. They present simulation results for system throughput and packet bit rate for different average C/I values, different fixed coding schemes and for 1, 2, 4 and 8 fixed PDCHs. The aim of this work is to highlight the capabilities of GPRS and not to solve the traffic engineering problem.

All these papers do not contain any result for on-demand channel configurations with coexisting circuit-switched speech traffic being the typical situation during the introduction and evolution of GPRS networks.

Further papers were published considering on-demand channels like [Budka (1997); Lindemann and Thümmler (2001); Litjens and Boucherie (2000); Ni and Häggmann (1999); Ni et al. (1999)].

In [Lindemann and Thümmler (2001)] an analytical model for the GPRS radio interface, based on a continuous-time Markov chain, is presented to derive the number of packet data channels needed under a given traffic load in order to reach an appropriate QoS. Since the traffic model is a Markov-modulated Poisson Process, the effects of Internet traffic that is characterized by a heavy-tailed distribution and the dynamic control by TCP cannot be regarded. Additionally the QoS perceived by the user, e.g., the throughput during transmission periods, is not determined in this work.

In [Litjens and Boucherie (2000)] an analytical model based on a continuous-time Markov chain is used to derive a closed-form expression for the expected sojourn time of a GPRS data call. The Internet traffic characteristics are not regarded and the TCP behavior is much simplified in this model. The use of the results for traffic engineering is not outlined.

In [Ni and Häggmann (1999); Ni et al. (1999)] the impact of GPRS traffic on speech services, namely the blocking probability, is presented in typical GSM/GPRS cell configurations. The capacity remaining for GPRS and the resultant GPRS performance are estimated. This work is founded on the ETSI traffic models mentioned above. Results for Internet applications are not given.

In [Budka (1997)] the availability of channels for packet data transmission is calculated without consideration of the specifics of GPRS. The impact on the performance of Internet applications has not been examined in this work.

2.2.1.2 Methodology

Analytical and algorithmic models for the performance analysis of packet-switched radio networks are under development [Vornefeld (2001)], but the full details of the GPRS protocol stacks of the radio interface and the fixed network and of the Internet protocols including the characteristics of TCP currently cannot be adopted into the models that are analytically tractable. The idea of using a dedicated channel for packet data services can be related to a traditional Reservation-Aloha approach as presented in [Goodman and Wei (1991); Lam (1980)]. For this concept well known queueing theory exists to derive its performance, if the traffic is characterized by Markovian arrivals without elastic behavior. If heavy-tailed traffic in combination with TCP occurs, these techniques no longer apply. For these traffic characteristics new traffic theory needs to be developed [Roberts (2001)].

Since GPRS networks are presently introduced in the field, traffic engineering rules and related performance results are soon needed so that capacity and performance estimations are possible in advance to the GPRS introduction and its evolution.

Measuring the traffic performance in the existing GPRS network is unsuitable, since a scenario with well-defined traffic load is hard to set-up. The evaluation of the performance based on measurement is very difficult, and the analysis of the behavior of different protocol options is impossible in an existing radio network.

Therefore computer simulation based on the prototypical implementation (called emulation) of the GPRS and the Internet protocols in combination with stochastical traffic generators for the regarded applications and models for the radio channel are chosen as the methodology to get the needed results rapidly. A number of samples, which is sufficient for traffic engineering, has been used. Per simulation several thousands of web and e-mail sessions have been simulated per mobile station in the scenario. The LRE algorithm [Schreiber (1988); Schreiber and Görg (1996)] has been applied to ensure a relative error interval of less than 5% for each value range of a distribution function for the variables considered.

2.2.1.3 Scenarios

For the introduction and the evolution of GPRS and EGPRS, scenarios with a relatively small number of users per radio cell and small traffic volumes per user have to be regarded.

The number of PDCHs per radio cell is initially limited to 8 and this limit will not increase significantly in the next years, since speech traffic has to be served with the accustomed quality and the spectrum for GSM services is limited. With this capacity limit for GPRS the maximum IP throughput per cell will not exceed 80 kbit/s even in overload situations. This means that scenarios with 1–40 active stations per radio cell are adequate for traffic engineering.

Since GPRS terminals will only support a few timeslots to receive in parallel, the performance for browsing applications is limited, so that operators are predicting only small traffic volumes per user (active or non-active) during the busy hour. While the active stations regarded in the simulation scenarios should offer traffic volumes of several kbit/s during their sessions, the mean traffic volume per user in a planning scenario, where hundreds of users per radio cell are regarded, either active or inactive, is much smaller, i.e., in the order of several kbyte/h.

Traffic patterns of interest are characterized by Web browsing and e-mail with small object sizes, because applications with higher QoS requirements initially can not be served by GPRS. Transaction-oriented applications like WAP are less important for traffic engineering because of the very small offered traffic compared to WWW and e-mail. Mobility should be regarded as part of the simulation model, whereas the effects of mobility management, especially cell change procedures, are neglected in this traffic engineering approach. In contrast to the GSM system where the handover decision is made by the network based on measurement reports sent by the mobile terminal, the cell change is usually not prepared and triggered by the network. It is decided by the mobile station, when it detects that it has entered a new cell by reading the system information on the broadcast channel. Thus transmission interruptions during cell changes are natural in GPRS environments and need not necessarily be taken into account for access network dimensioning.

2.2.2 Traffic Management

To elaborate a concept for a traffic management functionality based on application QoS requirements, it is necessary to define the implications of the term QoS itself.

In this context, QoS shall be defined as “the collective effect of service performance, which determines the degree of satisfaction of a user of the service” [ITU-T SG 2 (1994)].

It is composed of numerous parameters. In the following the most important ones will be described.

2.2.2.1 QoS Characteristics

Technically, QoS refers to an aggregation of system performance measures. The five most important of these are [Black (2000); Dutta-Roy (2000); ITU-T SG 2 (1994)]:

Availability: The availability of a network, its components, or even a service, should ideally approximate 100%. Even a value like 99.8% means about an hour and a half of down time per month. Thus, consumer satisfaction largely depends on this parameter.

Throughput: This is the effective data transfer rate available to an application, measured in bit/s. It depends on – but is explicitly not the same as – the maximum capacity of the network. Multiple connections sharing a transmission link, packet errors and losses during transfer, overhead imposed by protocol headers as well as characteristics of the nodes on the transmission path, such as buffer capacity or processing power, lower the throughput at disposal for an application.

Packet loss: Network elements like switches and routers are equipped with buffered queues for adopting to link congestion to some extent. However, if a link remains congested for too long, this will result in a buffer overflow and thus a loss of data. In a mobile radio network packets may additionally get lost because of the special conditions on the radio interface. Both cases usually result in a retransmission of the packet increasing the total transmission time.

Latency: Latency or delay is understood as the time taken by data to travel from its source to its destination. Thus, it may also be referred to as end-to-end delay, and is an important aspect of the perceived QoS. Since long delays reduce the interactivity of communication, interactive real-time (RT) applications are especially affected, while non-interactive RT applications show more sensitivity to a variation in delays, also called jitter. Non-real-time (NRT) applications are usually not delay-sensitive.

Various components add up to the end-to-end delay of a packet on a transmission path:

- *Transmission delay:* the time it takes to put all bits of a packet onto the link
- *Propagation delay:* the time it takes for a bit to traverse a link (e.g., at the speed of light)
- *Processing delay:* the time it takes to process a packet in a network element (e.g., routing it to the output port)
- *Queueing delay:* the time a packet must wait in a queue before it is scheduled for transmission

In a mobile radio environment there may be additional delays adding to the delays mentioned, caused by random access or paging mechanisms.

Jitter: Jitter, or latency variation, may be induced by various causes, e.g., variations in queue length, variations in the processing time needed to reorder packets that arrived out of order because they traveled over different paths, and variations in the processing time needed for reassembly of packets segmented before being transmitted. Again, interactive RT applications, especially, are sensitive to delay jitter as well as non-interactive RT applications. The latter may be able to adjust their playback point, i.e., the time offset between playback of consecutive packets based on changes in the jitter value. They are called “adaptive” applications then. Packets arriving after their playback point has passed are generally not useful to the application, and are, in most cases, discarded.

Table 2.1: Varied sensitivities of network traffic types

Traffic type	Sensitivities			
	Bandwidth	Loss	Delay	Jitter
Voice	Very low	Medium	High	High
E-commerce	Low	High	High	Low
Transactions	Low	High	High	Low
E-mail	Low	High	Low	Low
Telnet	Low	High	Medium	Low
Casual browsing	Low	Medium	Medium	Low
Serious browsing	Medium	High	High	Low
File transfers	High	Medium	Low	Low
Video conferencing	High	Medium	High	High
Multicasting	High	High	High	High

As mentioned above, applications vary significantly in their QoS requirements (see Table 2.1, [Dutta-Roy (2000)]). Interactive RT applications like *Voice over IP* (VoIP) have the most stringent demands on system performance, especially concerning delay and delay jitter. While non-interactive RT applications, like streaming audio or video, largely depend on small jitter values and, to a certain extent, on packet delay, NRT applications, like FTP or e-mail, are in most cases delay and jitter insensitive, but need a throughput as high as possible. Compared to delay or jitter, occasional loss of packets does not have a strong impact on the performance of RT applications of any kind, and solely reduces the throughput, because of packet retransmissions, when regarding NRT applications. Network availability should be as high as possible, but is not a parameter under the influence of mechanisms for QoS support. Thus, primarily such mechanisms have to concentrate on optimizing delay and throughput measures on behalf of application requirements.

2.2.2.2 QoS Support

The fundamental idea is that traffic can be differentiated and classified into different classes of service. The granularity of differentiation may be a small set of classes (e.g., simple priority) or could be as fine as each application flow. Some control must determine how much traffic of each class is allowed into the network, considering the available resources – this may be done statically (provisioning) or dynamically (signaling for resource reservations). Additionally, network elements must manage the processing and queueing of packets in such a way that appropriate, differentiated services are provided to the packets.

There are two sub-categories of traffic management:

Reservation-based: In this model resources for traffic are explicitly identified and reserved. Network nodes classify incoming packets and use the reservations to provide differentiated services. Typically, a dynamic resource reservation setup protocol is used in conjunction with admission control to set up reservations. Further, the nodes use intelligent processing, e.g., *Random Early Detection* (RED), and queueing mechanisms, e.g., *Weighted Fair Queueing* (WFQ), to service packets.

Reservation-less: Here, no resources are explicitly reserved. Instead, traffic is differentiated into a set of classes, and network nodes provide priority-based treatment of these classes. It may still be necessary to control the amount of traffic in a given class that is allowed to be injected into the network to preserve the QoS being provided to other packets of the same class.

2.2.2.3 QoS Support in Cellular Packet Radio Networks

The GPRS specification follows the reservation-less approach defining a set of QoS parameters that are combined to QoS profiles. It is designed to meet the requirements of one kind of traffic class each. Connection admission control can be performed to ensure that the QoS, negotiated for the packet data flows already in the system, remains unchanged, and that any kind of traffic will be served at least to a certain degree.

Operational cellular packet radio networks only include very basic MAC scheduling algorithms based on simple round robin and only realize best-effort services without QoS support for different applications and subscribers with their specific QoS requirements. Admission control to avoid overload situations is also not realized.

2.2.2.4 Related Work

In the last years algorithms were proposed for adaptive best-effort scheduling in wireless networks making use of the information on the link quality for different flows [Jeong et al. (1999); Yoshimura et al. (1998)]. Protocol aspects, e.g., the interworking of RLC and MAC protocols have not been addressed yet.

For traffic class specific scheduling several algorithms have been proposed. They can be subdivided into priority schemes and bandwidth sharing schemes. For bandwidth sharing schemes several concepts are available designed for ATM switches and IP routers such as *Weighted Fair Queueing* (WFQ), *Weighted Round Robin* (WRR) and *Deficit Weighted Round Robin* (DWRR) [Shreedhar and Varghese (1995)]. Hybrid scheduling approaches have been developed to support real-time applications together with background traffic composed of several traffic classes [Shreedhar and Varghese (1995)].

Older scheduling algorithms for queueing systems, based on the job duration or queue length of each connection, e.g., *Shortest Jobs First* (SJF) currently cannot be applied effectively for Internet traffic, since the TCP flow control hides the information how much data has to be transmitted for each session from the data link layer.

2.2.2.5 Proposed Solutions for QoS Support

The contribution of this work in the area of QoS support is the application and extension of the concepts described in Section 2.2.2.4 for best-effort scheduling and traffic class support to be deployed in the base station RLC/MAC layer of cellular packet radio networks.

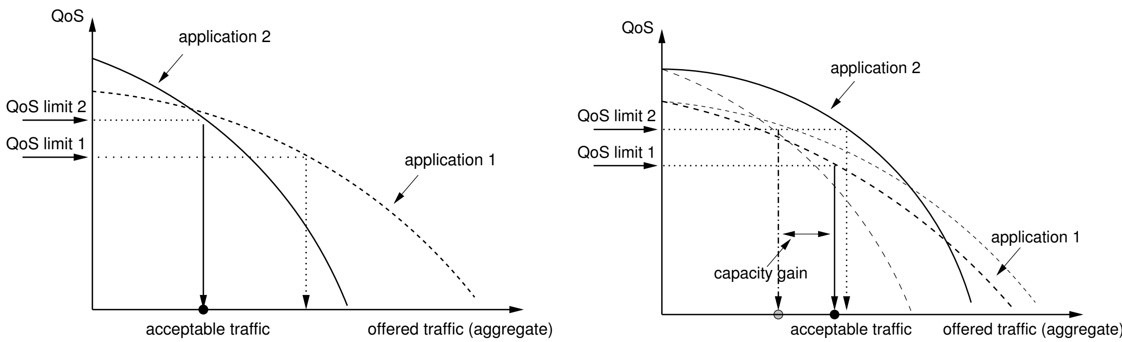
In the first step a MAC scheduler for efficient best-effort scheduling is proposed. A new feature is introduced that considers the urgency of RLC blocks to transmit into the decision, which logical connection is to be scheduled. Additionally the use of information on the link quality, which is already available in packet radio networks to support link adaptation, is proposed to realize adaptive scheduling of different logical connections.

To support different traffic and subscriber classes, priority algorithms are compared with bandwidth sharing algorithms. The interworking with a *Connection Admission Control* (CAC) algorithm is discussed. Parameters are proposed that can be configured by the operator both in single- and in multi-vendor environments for the maximum number of admitted sessions per traffic class and for the related weights of the traffic classes in the MAC scheduler. The concepts for advanced best-effort scheduling and traffic class scheduling can be integrated by using the proposed best-effort scheduling algorithm inside each traffic class queue.

The performance of the proposed integrated scheduling algorithms, based on the paradigm of the EGPRS standard, is evaluated. These algorithms can be implemented in GPRS and EGPRS networks without changing the standard and can also be employed in other cellular packet radio networks.

2.2.3 Mutual Dependency of Traffic Management and Traffic Engineering

If no QoS support is available, the application with the most critical QoS requirement determines the needed resources or the maximum offered traffic that can be served. In Figure 2.3(a) application 2 passes its QoS limit already with a relatively small overall system load, while the performance of application 1 is still significantly higher than the related QoS limit under the same system load. Using traffic management functions the performance of the most critical application can be improved, while the performance of other applications will decrease. In this way, the acceptable offered traffic can be maximized by the optimized parameterization of the QoS functions. This is shown in Figure 2.3(b). The



(a) Acceptable traffic without traffic management

(b) Acceptable traffic with traffic management

Figure 2.3: Mutual dependency of traffic management and traffic engineering

performance of the critical application 2 is increased so that both applications reach their QoS limit under nearly the same system load. With this optimization more traffic can be carried by the system, while all QoS requirements are still met. This has an effect on the traffic engineering process which is usually determined by the application with the most critical QoS requirements.

CHAPTER 3

General Packet Radio Service

The main goal of integrating the *General Packet Radio Service* (GPRS) into the *Global System for Mobile communication* (GSM) is to use the GSM radio resources more efficiently than is possible with the existing GSM phase 2 services. This is realized by allowing a number of logical connections per bearer to co-exist and by multiplexing packet data of these connections onto the GSM physical channels (see Chapter 2).

GPRS has been standardized by the ETSI as part of the GSM Phase 2+ development. The phase 2+ specifications are defining the implementation of packet switching within GSM, which is essentially a circuit-switched (CS) technology.

Packet-switching means that GPRS radio resources are used only when users are actually sending or receiving data. Rather than dedicating a radio channel to a mobile data user for a fixed period of time, the available radio resource can be shared between several users. The actual number of users supported depends on the application being used and how much data is being transferred. Through multiplexing of several logical connections to one or more GSM physical channels, GPRS reaches a flexible use of channel capacity for applications with variable bit rates.

Additionally it brings *Internet Protocol* (IP) capability to the GSM network and enables the access to a wide range of public and private data networks using industry standard data protocols such as *Transport Control Protocol* (TCP)/*Internet Protocol* (IP) and X.25.

The standardization of the GPRS was broadly concluded in 1997. However, details were regularly discussed in the boards of standardization and the GPRS standard was regularly modified. First proposals for a packed-oriented data service in GSM were published in 1991 [Brasche and Walke (1997); Decker (1993, 1995); Walke et al. (1991a,b)].

The first GPRS-based services have been available since 2001 in Europe. Many countries worldwide will introduce GPRS in the next few years. With these new services mobile multimedia applications with net bit rates of up to 117 kbit/s will be offered and established on the market.

3.1 Design Approach

The motivation to define a GSM packet-switched data service in 1994 was motivated by the growing number of dedicated packet radio networks. The provision of a circuit-switched channel for bursty traffic was felt to be inefficient. Applications in mind were fleet management, logistics, telematics and mobile offices [Hillebrandt (2002)].

In order to realize economically priced packet services, it was the premise to change the GSM components as little as possible and to develop the new service in consideration of the limits of the existing GSM tele- and bearer services [Brasche (1999); Stuckmann (2002b)].

GPRS integrates a packet-based air interface into the existing circuit-switched GSM network. The GSM infrastructure is not to be replaced. A couple of new network elements have been added (see Section 3.2).

The GPRS specification does not provide an upper limit for the amount of data which can be transmitted per access. GPRS was initially designed for

- the frequent, regular transmission of short data packets up to 500 bytes, and
- the irregular transmission of short to medium-sized data packets up to a few kbyte.

The basic approach to integrate the packet data service into the GSM standard represents the reservation and the logical subdivision of certain GSM channels. The number of channels allocated for GPRS can be dynamically adapted to the load situation in the respective radio cell.

3.2 Logical Architecture

The existing GSM network does not provide sufficient functionality to realize a packet data service. Integrating GPRS into a GSM network requires the addition of components which provide the packet-switched service (see Figure 3.1). Hence, the GSM network is extended by two additional nodes:

Gateway GPRS Support Node (GGSN): The GGSN serves as the interface towards external *Packet Data Networks* (PDNs) or other *Public Land Mobile Networks* (PLMNs). Here, switching functions are fulfilled, e.g., the evaluation of *Packet Data Protocol* (PDP) addresses and the routing to mobile subscribers via the SGSN.

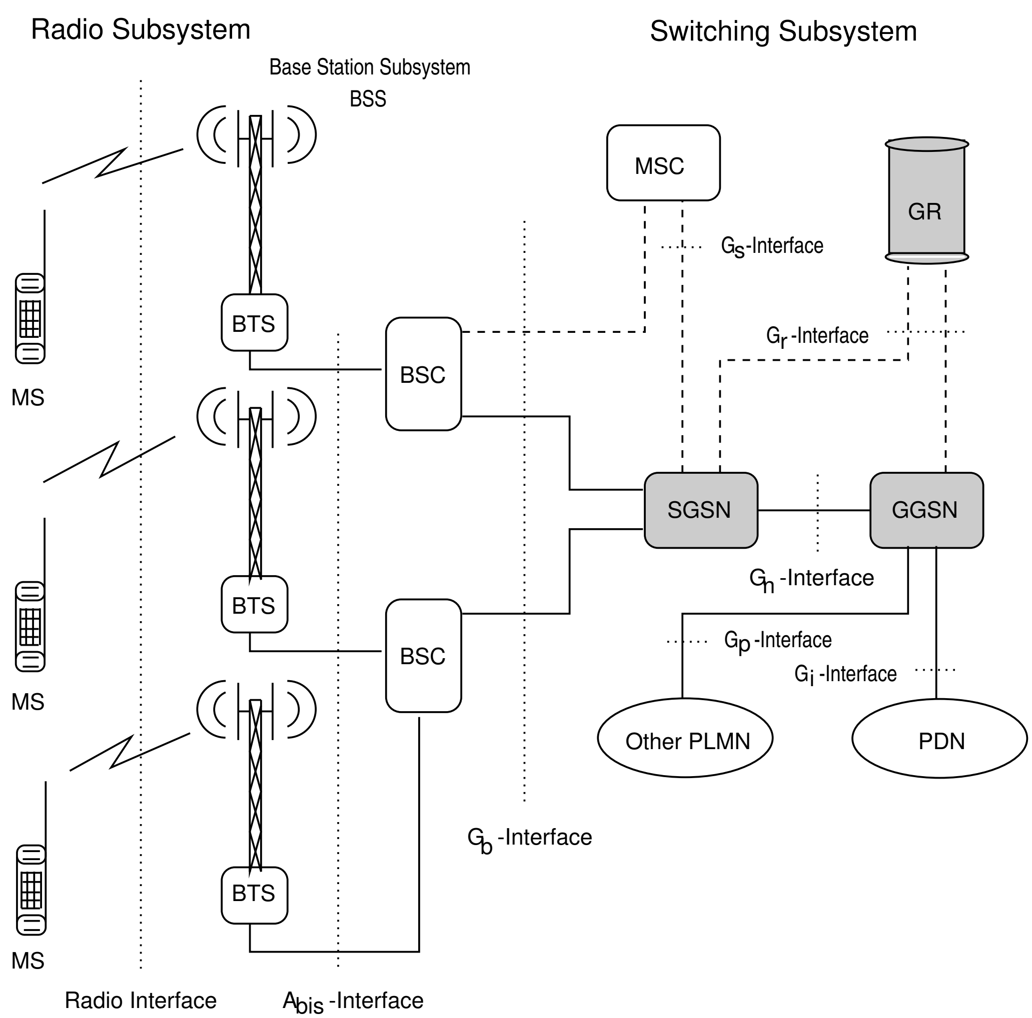


Figure 3.1: GPRS logical architecture

Serving GPRS Support Node (SGSN): The SGSN represents the GPRS switching center in analogy to the *Mobile-services Switching Center* (MSC). Packet data addresses are evaluated and mapped onto the *International Mobile Subscriber Identity* (IMSI). The SGSN is responsible for the routing inside the packet radio network and for mobility and resource management. Furthermore, it provides authentication and encryption for the GPRS subscribers.

For the communication between the SGSN and the GGSN within one PLMN, the Intra PLMN *IP Version 6* (IPv6) or *IP Version 4* (IPv4) backbone is used. SGSN and GGSN encapsulate and decapsulate, respectively, the packets using a special protocol called the *GPRS Tunneling Protocol* (GTP) that operates on top of standard TCP/IP protocols. The SGSN and GGSN functions may also be combined into one physical node.

3.3 Radio Interface

The radio interface U_m is located between MS and BSS. The GSM recommendations combine a *Frequency Division Multiplex* (FDM) and a *Time Division Multiplex* (TDM) scheme (see Chapter 2).

The communication over the GPRS radio interface comprises functions of the *Radio Link Control* (RLC)/*Medium Access Control* (MAC) layer and the physical layer. For the transfer of layer-2 messages over the radio interface the GSM logical channel concept is reused.

3.3.1 Channel Concept

The basic approach to integrate the packet data service into the GSM standard represents the reservation and the logical subdivision of certain GSM channels for GPRS. The physical channels dedicated to packet data traffic are called *Packet Data Channels* (PDCHs).

Similar to GSM, logical channels are mapped onto physical channels using a cyclically recurring multiframe structure. Two 26-frame multiframes are assembled to one GPRS 52-frame multiframe. A 52-frame multiframe represents one physical GPRS channel consisting of 12 radio blocks and four `Idle` bursts, each block comprising four radio bursts distributed on time slots with the same *Timeslot Number* (TN) in consecutive TDMA frames (see Figure 3.2 and Figure 3.3).

Since one TDMA frame represents eight physical channels to be shared by GSM and GPRS, there have to be strategies on how to distribute these resources, whether fixed or on-demand (see Figure 3.4). In [3GPP TSG GERAN (2002e)] the principles concerning this task are specified. The specific details of implementation are left to manufacturers and network operators.

3.3.1.1 The Master-slave Concept

One or more PDCHs are run as *Master Packet Data Channels* (MPDCHs) (see Figure 3.3), and provide *Packet Common Control Channels* (PCCCHs) that carry the necessary control and signaling information to initiate a packet data transmission. These PCCCHs have to be run if signaling information is not transmitted via existing *Common Control Channels* (CCCHs).

Furthermore PDCHs operate as slaves and are utilized for transmission of user data over *Packet Data Traffic Channels* (PDTCHs) and associated control over *Packet Associated Control Channels* (PACCHs).

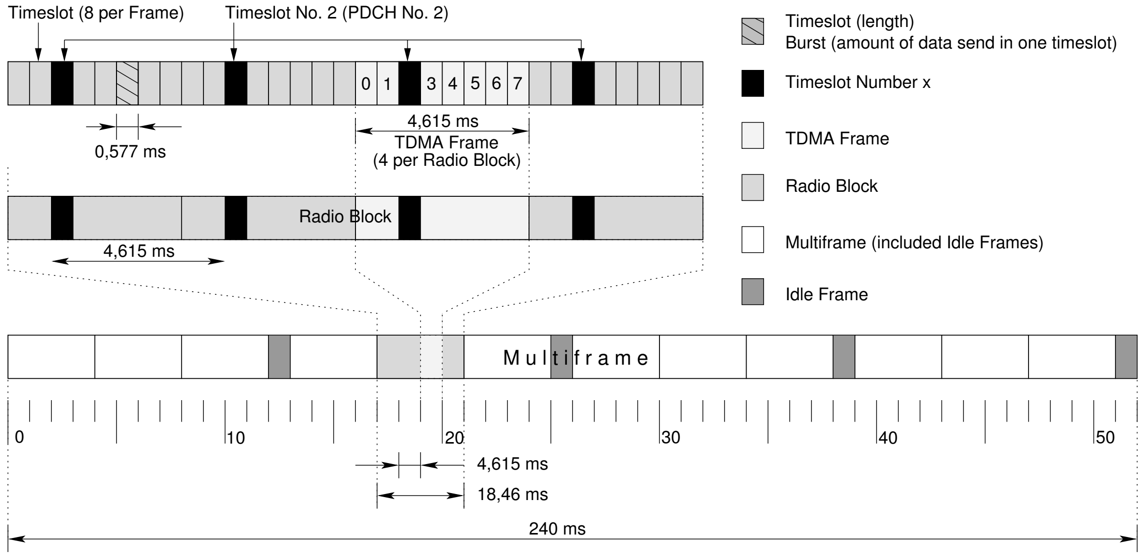


Figure 3.2: Duration of a time slot, TDMA frame, radio block and 52-frame

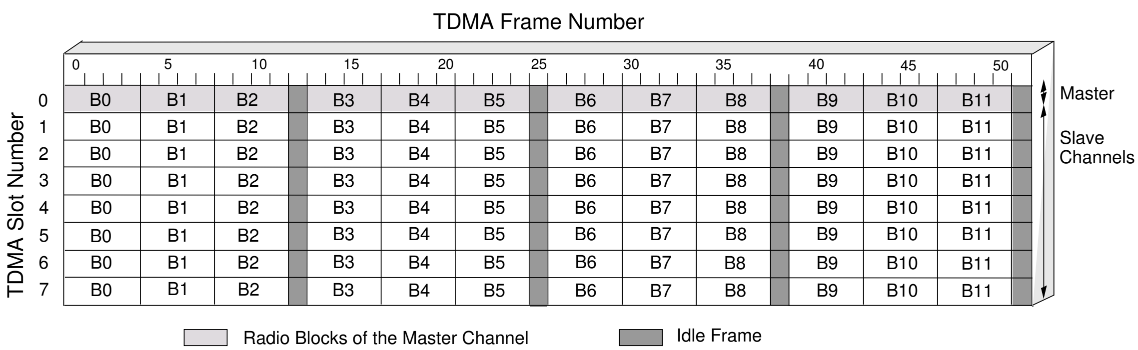


Figure 3.3: GPRS channel structure

3.3.1.2 The Capacity-on-demand Concept

GPRS does not require fixed allocated PDCHs. Capacity assignment for packet data transmission can be done according to actual demand. The decision about the number of fixed and on-demand PDCHs is left with the radio network operator.

There are several mechanisms to increase or diminish the number of actually assigned PDCHs on a capacity-on-demand base.

Load monitoring The PDCH utilization is supervised by a load monitor instance that should be implemented as part of the MAC functionality.

Dynamic allocation of PDCHs Unused channels, whether by GSM or GPRS, may be allocated as PDCHs to increase the GPRS QoS. PDCHs are released in order to meet the obligations of higher priority services.

Release of PDCHs Fast release of PDCHs is an important criterion for a pool of radio resources to be dynamically shared by circuit as well as packet switched services. To achieve this goal, there are the following possibilities:

- Channel release is delayed until there are no more active packet data flows on the regarded PDCHs.
- Each user having allocated the PDCHs to be released has to be notified, e.g. by `Packet Resource Reassignment` messages from the network side.

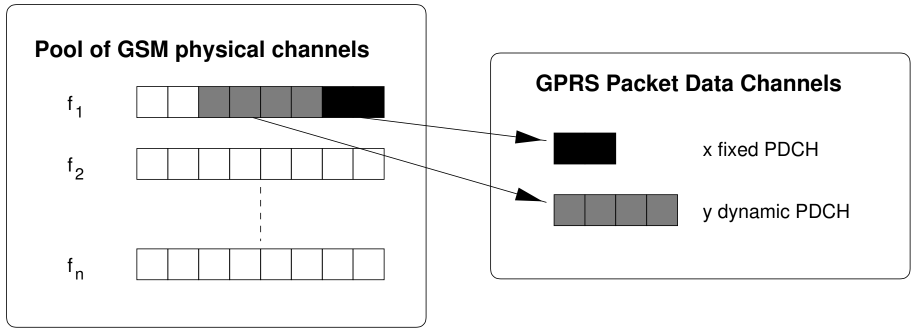


Figure 3.4: Assignment of GSM physical channels for GPRS

- The channel release notification is broadcast. There has to be one PDCHs monitored by all MSs to assure the reception of this notification.

In practice, a combination of the above methods can be implemented. In case of an MS not receiving a channel release notification there will be temporary channel interference because of the MS using a channel provided for other services. However, the MS will terminate the packet data flow as soon as it receives an inappropriate response on the downlink, and initiate a retransmission on a different PDCH.

3.3.2 Logical Channels

The packet data logical channels are mapped onto the physical channels dedicated to packet data (PDCHs).

Table 3.1 lists the GPRS logical channels and their functions. A detailed description for each channel is presented below.

3.3.2.1 Packet Common Control Channel (PCCCH)

The PCCCH comprises the following logical channels for GPRS common control.

PRACH: the *Packet Random Access Channel* (PRACH) is used by the MS in the uplink direction to initiate uplink transfer, e.g., for sending data or a paging response. In other words the channel is used by the MS to initiate a packet transfer or respond to paging messages. On this channel an MS transmits access bursts with long guard times. On receiving access bursts, the BSS assigns a value for the *Timing Advance* (TA) algorithm to each terminal.

PPCH: the *Packet Paging Channel* (PPCH) is used in the downlink direction to page an MS prior to downlink packet transfer. The PPCH uses paging groups in order to allow usage of *Discontinuous Reception* (DRX) mode.

PAGCH: the *Packet Access Grant Channel* (PAGCH) is used in the establishment phase of the packet transfer to send a resource assignment in the downlink direction to an MS requesting a packet transfer. Additionally, resource assignment for a downlink packet transfer can be sent on a PACCH if the MS is currently involved in a packet transfer.

PNCH: the *Packet Notification Channel* (PNCH) is used in the downlink direction to send a *PTM Multicast* (PTM-M) notification to a group of MSs prior to a PTM-M packet transfer. The notification has the form of a resource assignment for the packet transfer. DRX mode will be provided for monitoring the PNCH. Furthermore, a `PTM-`

Table 3.1: GPRS logical channels

Group	Channel	Name	Direction	Function
PCCCH	PRACH	Packet Random Access Channel	UL	random access
	PPCH	Packet Paging Channel	DL	paging
	PAGCH	Packet Access Grant Channel	DL	access grant
	PNCH	Packet Notification Channel	DL	multicast
PBCCH	PBCCH	Packet Broadcast Control Channel	DL	broadcast
PTCH	PDTCH	Packet Data Traffic Channel	UL/DL	data
	PACCH	Packet Associated Control Channel	UL/DL	assoc. control

`M new message` indicator may optionally be sent on all individual paging channels to inform MSs interested in PTM-M when they need to listen to the PNCH.

3.3.2.2 Packet Broadcast Control Channel (PBCCH)

The *Packet Broadcast Control Channel* (PBCCH) broadcasts packet data specific system information. If the PBCCH is not allocated, the *Broadcast Control Channel* (BCCH) transmits the system information to all GPRS terminals in a radio cell.

3.3.2.3 Packet Traffic Channel (PTCH)

The PTCH is allocated to carry user data and associated control information and is subdivided into the following channels.

PDTCH: the *Packet Data Traffic Channel* (PDTCH) is a channel allocated for data transfer. It is temporarily dedicated to one MS or in the *Point-To-Multipoint* (PTM) case to a group of MSs. In multislot operation, one MS may use multiple PDTCHs, e.g., more than one PDTCH, simultaneously for individual packet transfer.

PACCH: the *Packet Associated Control Channel* (PACCH) is used to convey control information related to a given MS, such as an acknowledgement message. The PACCH also carries resource assignment and reassignment messages comprising the assignment of capacity for PDTCHs and for further occurrences of PACCH. One PACCH is associated with one or several PDTCHs assigned to one MS.

3.3.3 Mapping of the Logical Channels

The mapping of the logical channels is defined by the multiframe structure. As described in Section 3.3.1 the multiframe structure for a PDCH consists of 52 TDMA frames, divided into 12 blocks (of four frames each) and four idle frames according to Figure 3.5. The mapping of the logical channels onto the radio block periods is defined by means of the ordered list of blocks (B0, B6, B3, B9, B1, B7, B4, B10, B2, B8, B5, B11). With this list a certain number of channels per PDCH is uniformly distributed over the multiframe. For example one PBCCH and three PAGCHs on the master channel would be allocated to B0, B6, B3 and B9 of the first PDCH.

PDCH that contain PCCCHs are indicated on the BCCH. The number of PDCHs containing PCCCHs is defined by the `BS PCC CHANS` parameter. The following restrictions are defined in [3GPP TSG GERAN (2002d)] concerning PCCCHs.

Only one PDCH—the master channel—contains one or more PBCCHs All other channels do not contain PBCCHs and are called slave channels. The logical channel PAGCH can only be mapped on the master channel and the slave channels that contain PCCCHs.

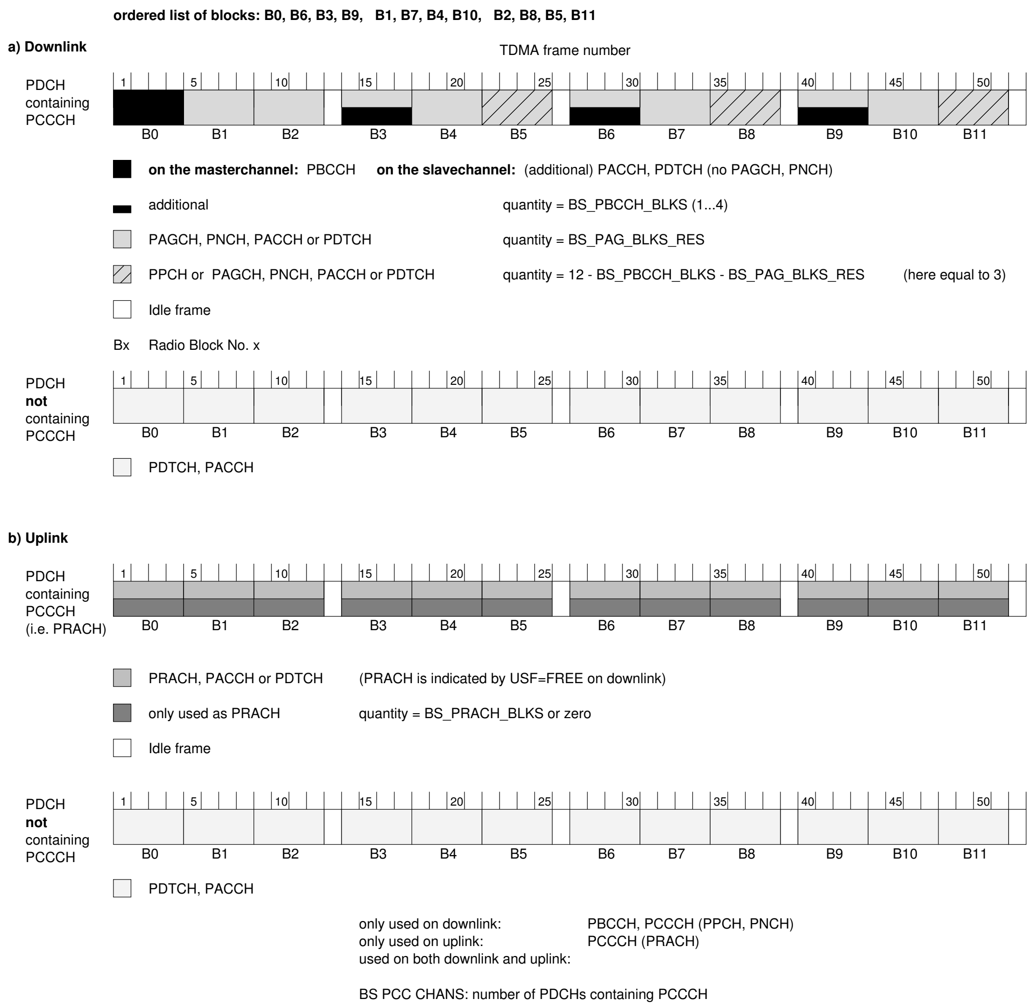


Figure 3.5: Mapping of the logical channels onto the radio blocks

The response message to a `Packet Channel Request` has to be sent on the same PDCH that was used for the `Packet Channel Request`.

The PRACH can only be mapped on an uplink PDCH containing PCCCHs. On PRACHs, access bursts are used, while on all other packet data logical channels, radio blocks comprising four normal bursts are used. The only exception are messages on uplink PACCHs, which can comprise four consecutive access bursts to increase robustness. The detailed mapping rules of the single channels as defined in [3GPP TSG GERAN (2002d)] are described in the following (see Figure 3.5).

3.3.3.1 Downlink

Mapping of PBCCHs: On the downlink of the master channel the first block in the ordered list of blocks (B0) is always used as a PBCCH. If required up to three more blocks on the same PDCH can be used as additional PBCCHs. The parameter `BS PBCCH BLKS` indicates the total number of PBCCH blocks (from one up to four) on the downlink master channel. Therefore on this PDCH the first `BS PBCCH BLKS` blocks in the ordered list of blocks are reserved for PBCCHs.

Any additional PDCH containing PCCCHs is indicated on the PBCCH. On these PDCHs the `BS PBCCH BLKS` first blocks in the ordered list of blocks are used as

PDTCHs or PACCHs in the downlink.

Mapping of PAGCHs, PNCHs, PDTCHs and PACCHs: On any PDCH containing PCCCHs (with or without PBCCH), the next `BS PAG BLKS RES` blocks in the ordered list of blocks are used as PAGCHs, PNCHs, PDTCHs or PACCHs. The `BS PAG BLKS RES` parameter is broadcast on the PBCCH and ranges from 0 up to `12 - BS PBCCH BLKS`.

Mapping of PPCHs: The remaining blocks in the ordered list of blocks are used as PPCHs. But they can also be used as PAGCHs, PNCHs, PDTCHs or PACCHs, if no paging messages are waiting for transmission.

3.3.3.2 Uplink

Mapping of PRACHs: On a PDCH containing PCCCHs, all blocks in the multiframe can be used as a PRACH—indicated by an Uplink State Flag in the previous downlink message—or as a PDTCH or a PACCH.
Optionally the `BS PRACH BLKS` first blocks in the ordered list of blocks are only used as PRACHs. The `BS PRACH BLKS` parameter is broadcast on the PBCCH. The remaining blocks in the multiframe are used as PRACH, PDTCH or PACCH.

Mapping of PACCHs and PDTCHs: On a PDCH that does not contain a PCCCH, i.e., that does not contain a PRACH (`BS PRACH BLKS`=0), all blocks are used as PDTCHs or PACCHs.

3.4 GPRS User Plane

The GPRS protocol architecture follows the *International Standards Organization* (ISO)/ *Open Systems Interconnection* (OSI) reference model [ISO/IEC JTC1 (1994)]. This reference model provides a common basis for the coordination of standards development for the purpose of systems interconnection, while allowing existing standards to be placed into perspective within the overall reference model. It divides the functions of a system into seven layers. The communication between the corresponding partner entities of each layer is specified by communication protocols comprising the syntax, semantics, timing and pragmatics of the exchanged data units.

The GPRS user plane, also called transmission plane, comprises all protocols used for user data transmission. In GPRS networks the user plane can be considered separate from the control plane, since communication is ongoing in two phases. While a so-called *Packet Data Protocol* (PDP) context, a logical context between an MS and an external *Packet Data Network* (PDN) based on TCP/IP or X.25, can be set up without actually transmitting user data, the user plane will only be utilized when the MS is actually transmitting or receiving user data. Nevertheless some protocols of the user plane are reused to carry signaling messages (see Section 3.5).

The kernel of the GPRS standard is the *Radio Link Control* (RLC)/*Medium Access Control* (MAC) protocol that operates on top of the GSM channel structure (see Section 3.3.1) and realizes the reliable packet-oriented transmission between *Mobile Station* (MS) and *Base Station Subsystem* (BSS) by statistical multiplexing of several logical connections on the PDCHs available for GPRS in the radio cell. While RLC/MAC contains already an *Automatic Repeat Request* (ARQ) scheme to handle transmission errors on the air interface, the *Logical Link Control* (LLC) protocol provides a reliable ciphered link between MS and SGSN applying flow control and error handling mechanisms known from link layer protocols in fixed networks like *Integrated Services Digital Network* (ISDN). The ARQ scheme in LLC is not designed to handle transmission errors on the air interface, but for handling frame losses and errors in the fixed network or for link recovery after a cell change.

The *Base Station Subsystem GPRS Protocol* (BSSGP), which is operating on top of a network service such as *Frame Relay* (FR), is responsible for flow control between BSS and SGSN. To support different network layer standards that may be used in GPRS networks the *Sub-Network Dependent Convergence Protocol* (SNDCP) adapts the network layer *Protocol Data Units* (PDUs) to the GPRS logical link. Finally the *GPRS Tunneling Protocol* (GTP) is used as part of the GPRS user plane to realize a transport service for IP datagrams by establishing an IP tunnel through the GPRS *Core Network* (CN) between SGSN and GGSN.

This section will describe all GPRS-specific protocols (see shaded blocks in Figure 3.6) that are involved in user data transmission in a top-down approach.

3.4.1 GPRS Tunneling Protocol (GTP)

In GPRS networks the interworking with TCP/IP or X.25 based networks is foreseen [3GPP TSG CN (2002d)] on the G_i interface. Between SGSN and GGSN on the G_n interface, encapsulated data packets are transmitted with the help of the GTP. GTP signaling is used by the *GPRS Mobility Management* (GMM)/*Session Management* (SM) (see Section 3.5) to create, modify and delete tunnels [3GPP TSG CN (2002c)]. GTP uses the *User Datagram Protocol* (UDP) for the transmission in the core network.

Tunnels are used to carry encapsulated PDUs between a given GSN pair for individual mobile stations. The *Tunnel Identifier* (TID), part of the GTP header, indicates to which tunnel a particular datagram belongs. The TID specifies the PDP context (see Section 3.5). Like this, datagrams injected into the GPRS network at the GGSN are tunneled to the correct SGSN area, where the MS is presently located, which means that mobility-related routing is completely supported by the GTP. In other mobile core networks, e.g., networks based on the 3GPP2 specifications, this function is realized by *Mobile IP* [Perkins (1996)].

The GTP comprises two different entities, the GTP-U entity that is responsible for user data tunneling and the GTP-C entity responsible for signaling between SGSN and GGSN (see Section 3.5.2) using path, tunnel, location and mobility management messages.

3.4.2 Base Station Subsystem GPRS Protocol (BSSGP)

On the G_b interface, the BSSGP [3GPP TSG GERAN (2002c)] provides a connectionless link between BSS and SGSN. This protocol's main task is the flow control for the downlink transfer of LLC PDUs. No flow control is performed in the uplink direction. The SGSN is expected to accept all data that is sent to it—the buffers and link capacity have to be dimensioned according to this premise to avoid loss of uplink data.

3.4.2.1 Flow Control between SGSN and BSS

Each cell in the coverage area of one BSS is virtually connected to the SGSN by a *BSSGP Virtual Connection* (BVC), identified by a *BSSGP Virtual Connection Identifier* (BVCI). The BSS maintains one queue for each BVCI that contains the LLC PDUs for that particular cell. There are several ways to further split these queues, e.g., into one queue per MS, or into one queue that serves all LLC PDUs belonging to a certain delay or precedence class. The queues are filled with downlink LLC PDUs and emptied by the BSS forwarding these PDUs to the MSs addressed. The SGSN is regularly informed by the BSS about the maximum queue size available for each BVC and MS at a given time, and also about the

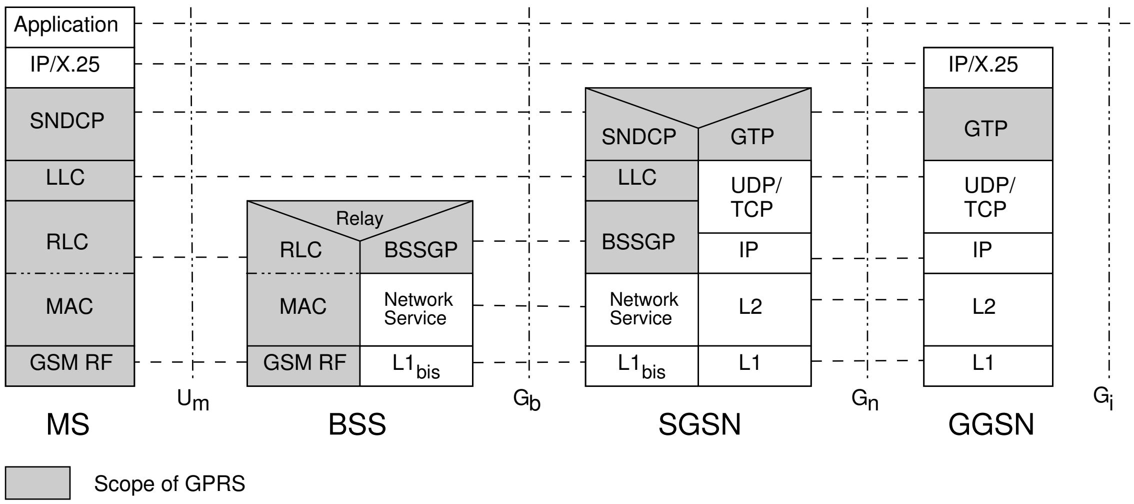


Figure 3.6: GPRS user plane

rate at which the available queue size increases. Thus, the SGSN is enabled to estimate the maximum allowable throughput per BVC as well as per MS belonging to the BVC.

Within the SGSN there are queues provided, similar to those in the BSS. Here the downlink LLC PDUs are queued and scheduled as long as the maximum allowable throughput per BVC and per MS is not exceeded. By this mechanism, it is ensured that both the MS and the BSS are able to handle the incoming traffic.

3.4.2.2 BSS Context

The SGSN can provide a BSS with information related to ongoing user data transmission. The information related to one MS is stored in a BSS context. The BSS may contain BSS contexts for several MSs. A BSS context contains a number of BSS *Packet Flow Contexts* (PFCs). Each BSS PFC is identified by a *Packet Flow Identifier* (PFI) assigned by the SGSN. A BSS PFC is shared by one or more activated *Packet Data Protocol* (PDP) contexts with identical or similar negotiated QoS profiles. The data transmission related to PDP contexts that share the same BSS PFC comprises one packet flow.

Three packet flows are predefined, and identified by three reserved PFI values. The BSS does not negotiate BSS PFCs for these pre-defined packet flows with the SGSN. One pre-defined packet flow is used for best-effort service, one is used for SMS, and one is used for signaling. The SGSN can assign the best-effort or SMS PFI to any PDP context. In the SMS case, the BSS handles the packet flow for the PDP context with the same QoS that it handles SMS with.

The combined BSS QoS profile for the PDP contexts that share the same packet flow is called the *Aggregate BSS QoS Profile* (ABQP). The ABQP is considered to be a single parameter with multiple data transfer attributes. It defines the QoS that must be provided by the BSS for a given packet flow between the MS and the SGSN, i.e., for the U_m and G_b interfaces combined. The ABQP is negotiated between the SGSN and the BSS.

A BSS packet flow timer indicates the maximum time for which the BSS may store the BSS PFC. The BSS packet flow timer is started when the BSS PFC is stored in the BSS and when an LLC frame is received from the MS. When the BSS packet flow timer expires the BSS deletes the BSS PFC.

When a PDP context is activated, modified, or deactivated (see Section 3.5.2.3), the SGSN may create, modify, or delete BSS PFCs.

3.4.3 Sub-Network Dependent Convergence Protocol (SNDCP)

Network layer protocols are intended to be capable of operating over services derived from a wide variety of subnetworks and data links. GPRS supports several network layer protocols providing protocol transparency for the users of the service. Therefore, all functions related to transfer of network layer PDUs are carried out in a transparent way by the GPRS network entities. SNDCP provides this adaptation of different network layers to the logical link by using *Network layer Service Access Point Identifiers* (N-SAPIs) to map different PDPs onto the services provided by the LLC layer. Additionally SNDCP supports compression of redundant user data and protocol control information. The main functions of the SNDCP layer are [3GPP TSG CN (2002b)]:

- Multiplexing of several PDPs.
- N-PDU buffering.
- Compression and decompression of user data and control information (e.g., TCP/IP header compression).
- Segmentation and reassembly of network PDUs.

The N-PDUs are buffered in the SNDCP layer before they are compressed, segmented and transmitted to the LLC layer. For acknowledged data transfer, the SNDCP entity buffers an N-PDU until successful reception of all segments of the N-PDU have been confirmed. The confirmation is carried out by the underlying layer using a confirm service primitive. N-PDUs buffered at the transmitting side, which which have been acknowledged by the receiver side are cleared. For unacknowledged data transfer, the SNDCP deletes an N-PDU immediately after it has been delivered to the LLC layer. Acknowledged or unacknowledged mode on SNDCP level only means that the SNDCP layer is using the acknowledged or unacknowledged data service from the LLC layer. In the SNDCP layer itself no error handling functions are performed.

The protocol control information compression method is specific for each network layer protocol type. TCP/IP (IPv4) header compression is specified in [Jacobson (1990)]. For data compression [ITU-T (1990)] may be used.

After the SNDCP header is added the compressed N-PDU is segmented if it is longer than the maximum payload size for an LLC frame.

3.4.4 Logical Link Control (LLC)

The LLC layer [3GPP TSG CN (2002a)] provides a highly reliable and ciphered link between the MS and the SGSN. It includes functions for:

- Provision of one or more logical connections.
- Sequence control.
- Error detection and correction.
- Flow control.
- Ciphering.

The protocol for the LLC sublayer is based on well-known data link protocols like the *Link Access Procedure on the D-channel* ($\mathrm{LAPD_m}$) used in GSM and the *High level Data Link Control* (HDLC) protocol standardized by ISO for *Open Systems Interconnection* (OSI).

The key modifications can be summarized as follows:

Variable frame length The GPRS protocol architecture allows a variable frame length at the LLC level. Therefore frame delimiters and bit stuffing are not necessary but an additional field is required in the frame header for specification of the frame length.

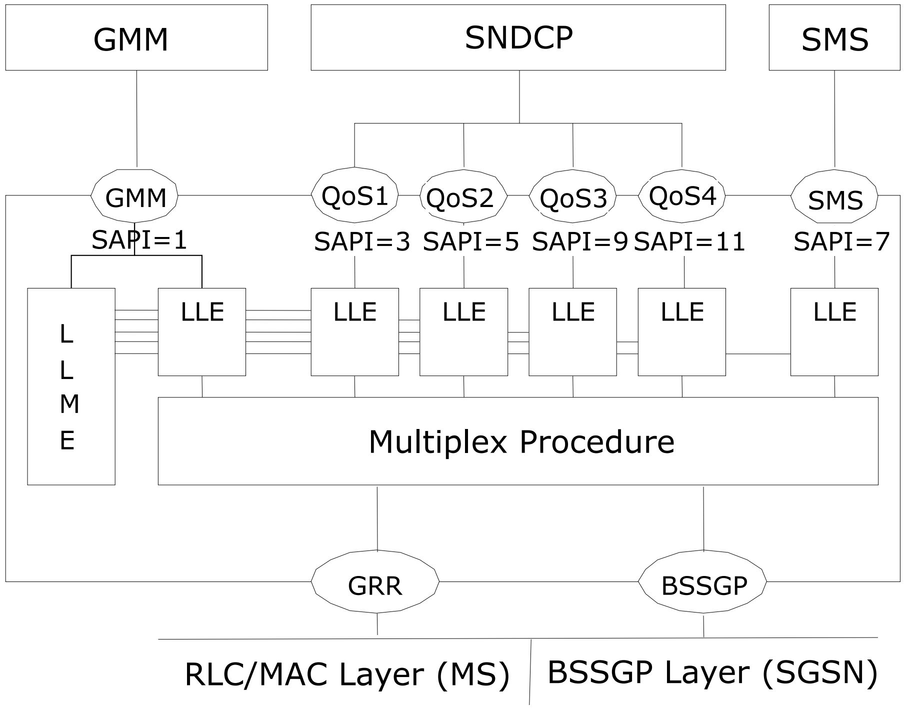


Figure 3.7: LLC layer structure

Prioritized Service Access Points (SAPs) The priority classes in GPRS are considered in the introduction of new SAPs with priorities.

3.4.4.1 Layer Entities and Service Access Points

The LLC layer provides six SAPs to the upper layer:

- Four SAPs are dedicated to the SNDCP that manages data packet transmission; one SAP exists for each QoS class
- One SAP is dedicated to *GPRS Mobility Management* (GMM) (see Section 3.5)
- One SAP is dedicated to *Short Message Service* (SMS)

The related services are performed by different logical link entities (see Figure 3.7) that are managed by the logical link management entity. Corresponding to four radio priority levels, four *Service Access Point Identifiers* (SAPIs) are required for packet data transmission. SAPIs of the LLC layer identify a point at which LLC services are provided by an LLC entity to a layer-3 entity. With the introduction of service or subscriber differentiation, it is also necessary that flow control and error correction is performed on a priority basis.

3.4.4.2 LLC Frame Structure

The LLC frame format is shown in Figure 3.8. The frame header consists of the address field and the control field and ranges from two to 37 octets. The address field consists of a single octet containing the SAPI (see Section 3.4.4.1) and the identifier of the logical link, the *Data Link Connection Identifier* (DLCI). The control field typically consists of between one and three octets (see Figure 3.9). LLC frames can be classified into information frames (I frames) containing sequence numbers for acknowledged data transmission, unnumbered

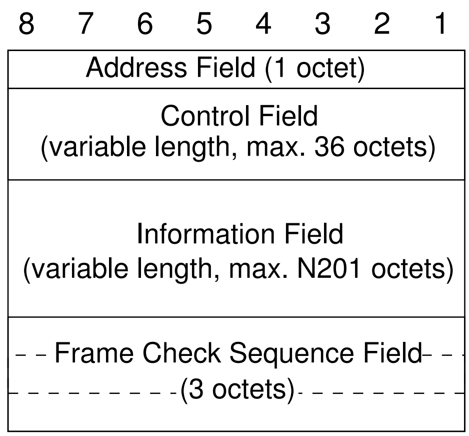


Figure 3.8: LLC frame format

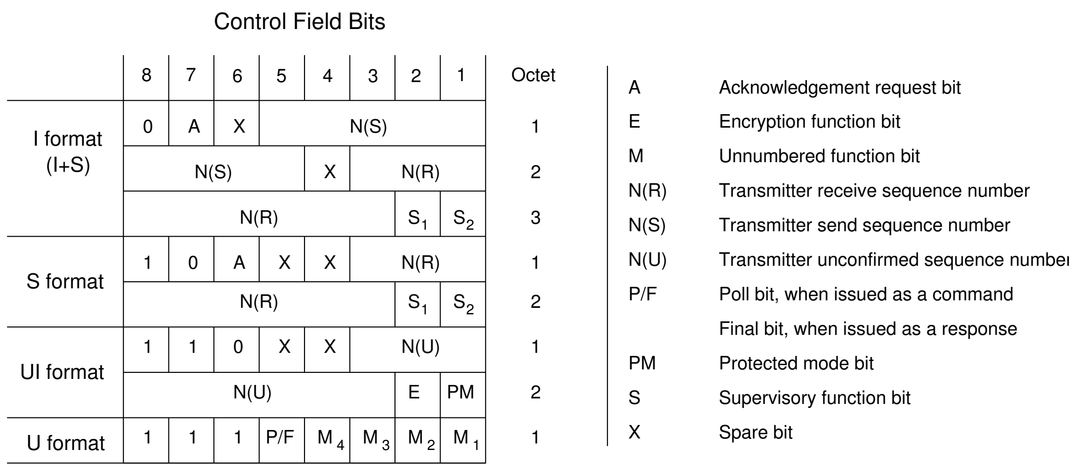


Figure 3.9: LLC control field

information frames (UI frames) for unacknowledged data transmission and supervisory frames (S frames) and unnumbered frames (U frames) containing LLC control messages. The *Selective Acknowledgement* (SACK) supervisory frame additionally includes a bitmap field of variable length of up to 32 octets. The information field, if present, follows the control field. The maximum length of the information field depends on the SAPI. The *Frame Check Sequence* (FCS) field consists of a 24-bit *Cyclic Redundancy Check* (CRC) code that is used to detect bit errors in the frame header and information field.

3.4.4.3 LLC Frame Transmission

The LLC layer performs an ARQ protocol based on retransmissions after timeouts or frame loss detections and optionally uses bitmap-based selective acknowledgements.

3.4.4.3.1 LLC Modes

LLC can operate in acknowledged and unacknowledged mode. In unacknowledged mode the network layer PDUs are transmitted in *Asynchronous Disconnected Mode* (ADM) using UI frames (see Figure 3.10). Neither LLC error recovery nor reordering procedures are defined, but transmission and format errors are detected. Duplicate UI frames are discarded.

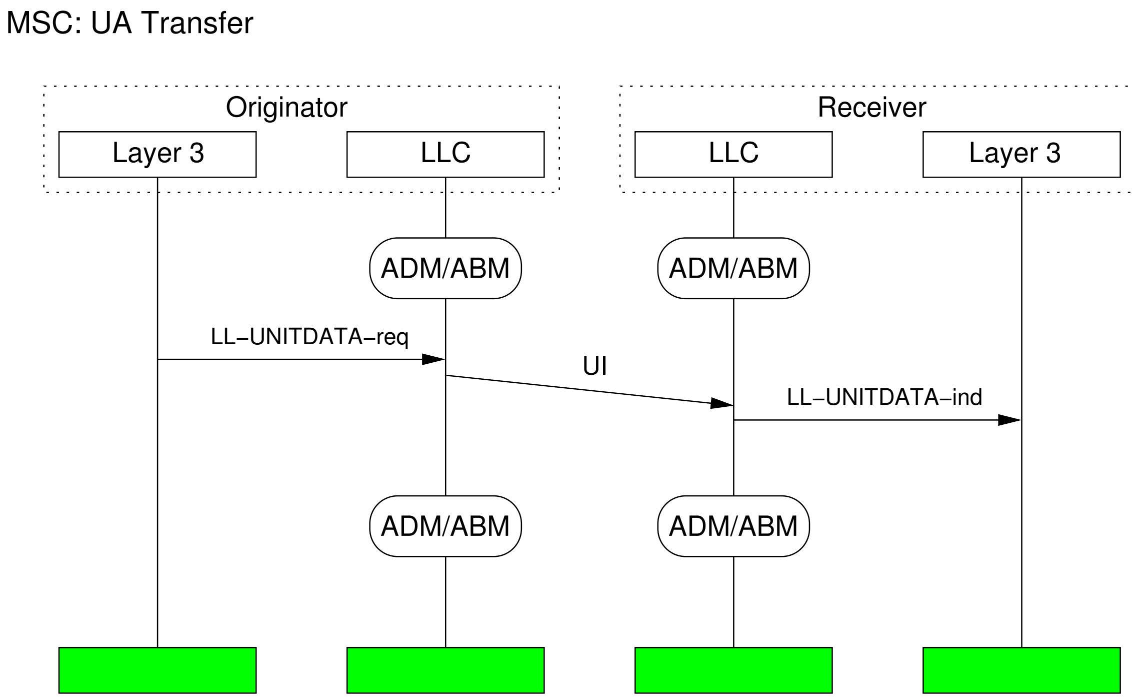


Figure 3.10: LLC unacknowledged mode

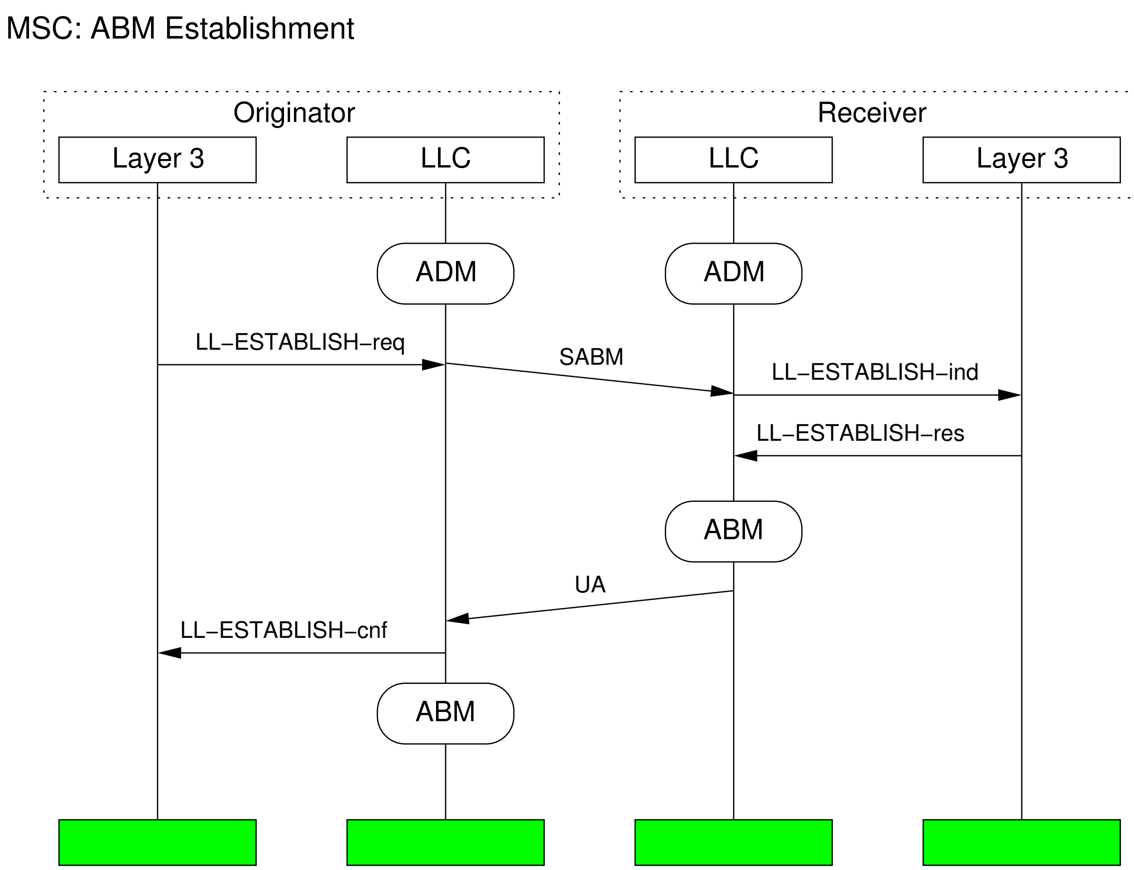


Figure 3.11: LLC acknowledged mode

Flow control procedures are not defined in unacknowledged mode. In acknowledged mode network layer information is transmitted within numbered I frames that are acknowledged by the peer entity. Error recovery and reordering procedures based on retransmission of unacknowledged I frames are specified. Acknowledged operation requires that the *Asynchronous Balanced Mode* (ABM) has been initiated in an establishment procedure using the *Set Asynchronous Balance Mode* (SABM) command (see Figure 3.11).

3.4.4.3.2 Flow Control

The flow control between the LLC peer entities is determined by the LLC window size that ranges from 2 to 16 and a modulus of 64 frames.

To perform vertical flow control between LLC and RLC/MAC, i.e. not to overload the RLC/MAC layer with I frames, a maximum buffer size is introduced. The actual buffer size is incremented with the length of the information field of a transmitted I frame and decremented with the size of an acknowledged LLC frame. The value of the buffer size shall never exceed the maximum. If a new I frame to transmit would make the buffer size exceed the maximum, then the LLC entity shall not transmit any new I frames, but still may retransmit I frames as a result of the error recovery procedures.

3.4.5 Radio Link Control (RLC) and Medium Access Control (MAC)

The idea of packet-switched services in GSM is to multiplex several users onto one single physical channel in order to use its capacity in a more efficient way. Furthermore, one single MS can use more than one PDCH simultaneously to increase the data rate. The maximum number of PDCHs that can be used in parallel is determined by the multislot capabilities of the MS, which can range in uplink and downlink from multislot capability 1 (one PDCH) up to multislot capability 8 (all eight PDCHs). The *multislot class* specified in the GPRS standard is a combination of the multislot capabilities in the uplink and downlink.

The *Data Link Control* (DLC) layer at the mobile U_m interface is divided in two sublayers: the RLC layer and the MAC layer, see Figure 3.6. The RLC sublayer provides a reliable logical connection between MS and BSS, while the MAC sublayer controls the access to the physical medium, the radio link.

3.4.5.1 Multiplexing Principles

The radio interface consists of asymmetric and independent uplink and downlink channels (see Section 3.3.2). The downlink carries packet data from the network to multiple MSs and does not require contention arbitration. The uplink is shared among multiple MSs and requires contention control procedures.

The access to the uplink is realized with a *Slotted-Aloha* based reservation protocol. A `Packet Channel Request` sent by an MS on the *Packet Random Access Channel* (PRACH) is responded by a `Packet Uplink Assignment` message indicating the uplink resources reserved for the MS.

LLC frames are segmented into RLC data blocks. At the RLC/MAC layer, a selective ARQ protocol between the MS and the network provides retransmission of erroneous RLC data blocks. When a complete LLC frame is successfully transferred across the RLC layer, it is forwarded to the peer LLC entity.

A *Temporary Block Flow* (TBF) is introduced as a virtual connection to support the unidirectional transfer of LLC PDUs on the packet data physical channels. The TBF is the allocated radio resource on one or more PDCHs and comprises a number of RLC/MAC blocks carrying one or more LLC PDUs. A TBF is temporary and maintained only for the duration of the data transfer, i.e., until there are no more RLC/MAC blocks to be transmitted and, in RLC acknowledged mode, all of the transmitted RLC/MAC blocks have been successfully acknowledged by the receiving entity.

A *Temporary Flow Identity* (TFI) is assigned for each TBF by the network. The MS assumes that the TFI value is unique among the concurrent TBFs in each direction. The same TFI value may be used concurrently for TBFs in opposite directions. An RLC/MAC block associated with a certain TBF comprises the TFI. The TBF is identified by the TFI together with the direction in which the RLC data block is sent and in case of an RLC/MAC control message additionally with the message type.

An *Uplink State Flag* (USF) is included in the header of each RLC/MAC block on a downlink PDCH. The USF may be used on the PDCH to allow multiplexing of radio blocks from different MSs and comprises three bits at the beginning of each radio block that is sent on the downlink. It enables the coding of *eight* different USF states which are used to multiplex the uplink traffic. On PCCCH, one USF value is used to denote PRACH (USF = FREE). The *seven* other USF values are used to reserve the uplink for different MSs.

Three medium access modes are supported:

- Dynamic Allocation (characterized by a USF)
- Extended Dynamic Allocation (supports the multislot functionality)
- Fixed Allocation (no uplink multiplexing during the TBF)

The Extended Dynamic Allocation medium access method extends the Dynamic Allocation medium access method to allow higher uplink throughput. The network allocates to the MS a subset of 1 to N consecutive PDCHs, where N depends on the multislot class of the MS.

In `packet idle mode` the MS monitors the relevant paging subchannels on the PCCCH. In `packet idle mode` no TBF exists and the upper layer may require the transfer of an LLC PDU, which implicitly triggers the establishment of a TBF and the transition to the `packet transfer mode`.

In `packet transfer mode` a TBF provides a physical point-to-point connection on one or more packet data physical channels for the unidirectional transfer of LLC PDUs between the network and the MS. Continuous transfer of one or more LLC PDUs is possible. Two parallel TBFs may be established in opposite directions for each MS.

3.4.5.2 RLC/MAC Block Structure

An RLC/MAC block consists of a MAC Header and an RLC data block or RLC/MAC control block, respectively. In the PLL a *Block Check Sequence* (BCS) is added for error detection. The RLC/MAC block together with the BCS are building one radio block that is transmitted in one PDCH in one radio block period (see Section 3.3.1). The RLC/MAC block structure is shown in Figure 3.12.

The RLC data block consists of RLC header, an RLC data field and spare bits. Each RLC data block may be encoded using any of the available channel coding schemes CS-1, CS-2, CS-3 or CS-4. The size of the RLC data block for each of the channel coding schemes is shown in Table 3.2. Two octets are always used in the RLC header for TFI, *Final Block Indicator* (FBI) and *Block Sequence Number* (BSN). The field for the `Length Indicator` and the `Extension`-Bit needs at least one octet. One additional octet is needed for each additional LLC frame. The remaining space is used for user data.

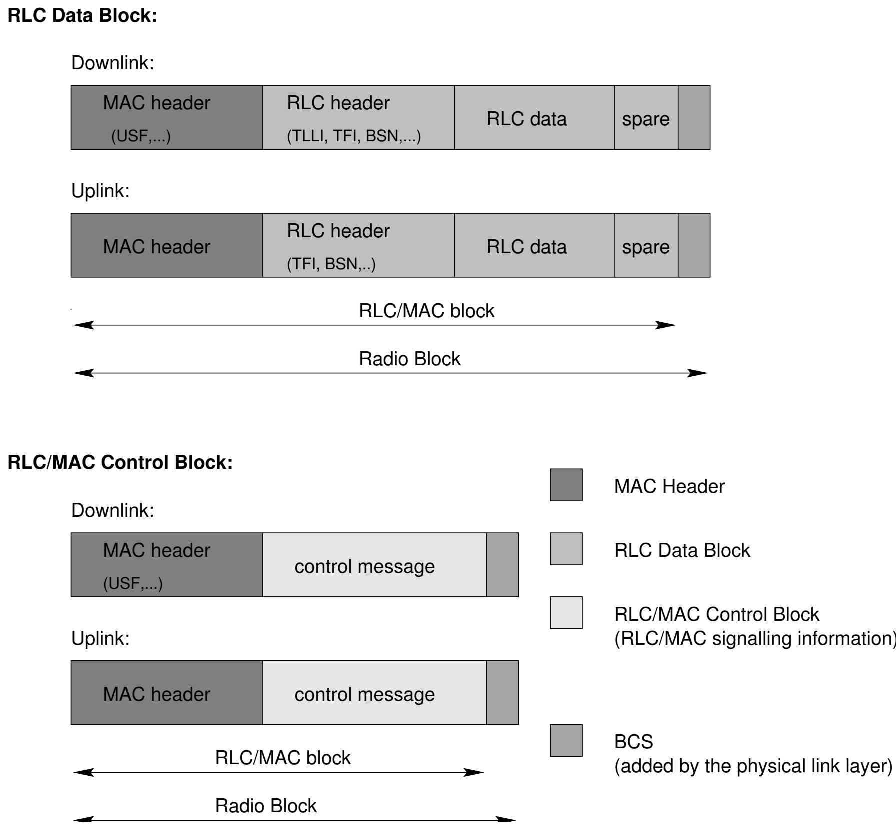


Figure 3.12: Structure of the radio block and RLC/MAC block

Table 3.2: RLC block size

Coding Scheme	RLC info field size [byte]	RLC data block size (without spare bits) [byte (bit)]	Spare bits [bit]	RLC/MAC block size [bit]
CS-1	20	22 (176)	0	181
CS-2	30	32 (256)	7	268
CS-3	36	38 (304)	3	312
CS-4	50	52 (416)	7	428

The RLC/MAC blocks containing an RLC/MAC control block always consists of 23 bytes, one byte for the MAC header and 22 bytes for the RLC/MAC control block containing the control message.

Each radio block carrying an RLC data block can be coded with one of the four different Channel Coding Schemes CS-1 to CS-4. For a radio block containing an RLC/MAC control block always the coding scheme CS-1 is used. The exception is messages that use the access burst, e.g., the `Packet Channel Request` message. The four different coding schemes are explained in detail in Section 3.4.5.6.

3.4.5.3 RLC Functions

The RLC functions in GPRS provide an interface towards the LLC layer, especially the segmentation and reassembly of LLC PDUs into RLC data blocks depending on the used *Coding Scheme* (CS) (see Table 3.2). *Backward Error Correction* (BEC) is used to enable the selective retransmission of uncorrectable PDUs. The BCS for error detection in GPRS is provided already by the Physical Link Layer (see Section 3.4.6).

There are two possible modes provided by the RLC/MAC layer: the acknowledged mode, used for reliable data transmission, and the unacknowledged mode, mainly used for real-time services, such as video or voice, where time delay is most critical, whereas bit errors are less important unless they pass the accepted range. For RLC acknowledged mode, a selective ARQ protocol is used between the MS and the BSS, which provides retransmission of erroneous RLC data blocks. As soon as a complete LLC frame is successfully transferred across the RLC layer, it is forwarded to the peer LLC entity.

3.4.5.3.1 Data Transmission in RLC Acknowledged Mode

In this mode, the transmitting side numbers the RLC data blocks with the *Block Sequence Number* (BSN) to supervise the acknowledged transmission of the RLC blocks. With *Acknowledgement* (ACK)/*Negative Acknowledgement* (NACK) messages on uplink and downlink (`Packet Uplink/Downlink ACK/NACK`), the receiving side can request retransmissions of RLC blocks if needed. A bitmap-based selective ARQ scheme is used with a window size of 64 and a modulus of 128.

Since the radio resource management is located in the BSS, the transmission of the acknowledgement messages in both directions is controlled by the BSS. The MS sends the ACK/NACK message in the reserved radio block which is requested from the peer entity by polling. `Packet Uplink ACK/NACK` messages can be transmitted directly by the BSS on the PACCH. In case of a negative acknowledgement, only those blocks listed as erroneous are retransmitted. In the case that the TBF ends before the LLC frame is transmitted completely, the missing part has to be transmitted during the new TBF.

3.4.5.3.2 Data Transmission in RLC Unacknowledged Mode

This mode comprises no retransmissions. The BSN is now only used to reassemble the RLC blocks. The `Packet ACK/NACK` messages are sent to convey the necessary control signaling, as for instance the monitoring of the channel quality. The MS transmits RLC data blocks without receiving a `Packet ACK/NACK` message until the *Window Size* (WS) is reached. Then it starts the timer `T3182` to wait for a `Packet Uplink ACK/NACK` message. On receipt of this message, the timer will be stopped. On expiry of the timer, an abnormal release with cell reselection will be performed.

3.4.5.4 MAC Functions

The GPRS MAC layer is responsible for providing efficient multiplexing of data and control signaling on the uplink and the downlink. The multiplexing on the downlink is controlled by the downlink scheduler, which has knowledge of the active MSs in the system and of the downlink traffic. Therefore, an efficient multiplexing on the PDCHs can easily be made. On the uplink, the multiplexing is controlled by channel reservation to individual MSs. This is done by resource requests, which are sent by the MS to the network, which then has to schedule the uplink PDCHs. Additionally, a contention resolution is needed when several MSs attempt simultaneously to access the control channel for resource requests. Collision detection and recovery are included when using the Slotted Aloha reservation protocol for recovery, which means that the MS has to wait a random backoff time before a new access attempt is done.

The MAC procedures include the provision of a TBF, which allows a point-to-point transfer of data in one cell between the BSS and the MS. As a physical connection between two radio resource entities, the TBF is used to transport the RLC data blocks on the PDCHs. The TBF is maintained only for the duration of the data transfer and is released as soon as all data is sent (RLC send queue is cleared). A TFI is assigned to the TBF by the network to associate the MS with the current TBF.

An *Uplink State Flag* (USF) is used by the network to control the multiplexing of different mobile stations on uplink PDCHs. It is included in the header of each RLC/MAC PDU on a downlink PDCH. The USF indicates which station is allowed to transmit during the next uplink radio block period (see Section 3.4.5.1).

3.4.5.4.1 TBF Setup

Both the network and the MS can initiate the establishment of a *Temporary Block Flow* (TBF) on the PCCCH allocated in the cell. The access is carried out on the PCCCH in either one or two phases, whereas two-phase access is used if the requested RLC mode is unacknowledged mode to ensure a safe establishment or if more than one time slot is requested by the MS. A one-phase access procedure is used if the amount of data to send fits into eight or fewer RLC/MAC blocks using CS-1.

3.4.5.4.1.1 Uplink TBF Setup

The MS enters the packet access procedure by sending a `Packet Channel Request` message on the PRACH and entering the `packet transfer mode` (see Figure 3.13). This `Packet Channel Request` message contains parameters required to indicate the mobile station's request for radio resources and the type of access needed. Access persistence control on PRACH can be steered either by the network (network steered method) or by the mobile station (mobile station steered method) to avoid collision failure.

In one-phase access, the network responds to the `Packet Channel Request` with the `Packet Immediate Assignment` (`Packet Uplink Assignment`), reserving the resources on PDCHs for uplink transfer of a number of radio blocks.

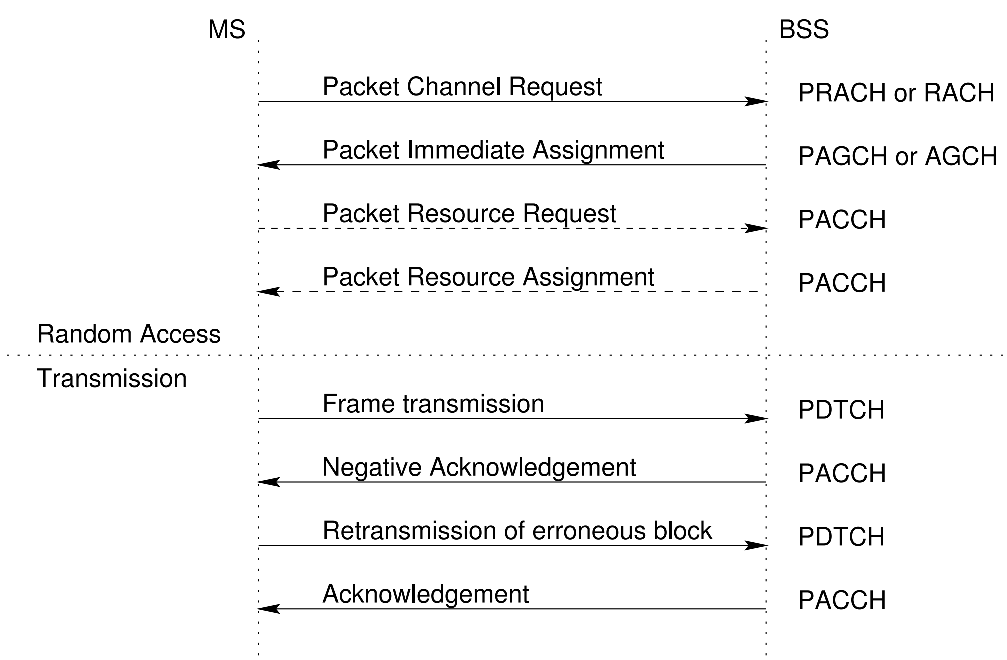


Figure 3.13: Uplink TBF establishment and data transmission

During two phase access, the network responds to the `Packet Channel Request` with `Packet Immediate Assignment` (`Packet Uplink Assignment`), which reserves the uplink resources for transmitting the `Packet Resource Request`. The `Packet Resource Request` message carries the complete description of the requested resources for the uplink transfer. Thereafter, the network responds with a `Packet Resource Assignment` message (`Packet Uplink Assignment`), reserving resources for the uplink transfer.

If there is no response to the `Packet Channel Request` within a predefined time period (`T3168`), the MS retries after a random back-off time. On receipt of a `Packet Channel Request`, the network sends a `Packet Uplink Assignment` message on the same PCCCH on which the network has received the `Packet Channel Request` message.

Packet data traffic is bursty in nature. Sometimes, the BSS will receive more channel requests than it can serve within a certain time limit. To avoid the repeat of `Packet Channel Requests` the sender is notified with a `Packet Queueing Notification` that its message is correctly received and will be handled later.

Efficient and flexible utilization of the available spectrum for packet data traffic with one or more PDCHs in a cell can be obtained using a multislot channel reservation scheme. Blocks from one MS can be sent on different PDCHs simultaneously, thus reducing the packet delay for transmission across the air interface. The bandwidth may be varied by allocating one to eight time slots in each TDMA frame depending on the number of available PDCHs, the multislot capabilities of the MS, and the current system load.

3.4.5.4.1.2 Downlink TBF Setup

A BSS initiates a packet transfer by sending a `Packet Paging Request` on the PPCH on the downlink in order to determine the location of an MS (if it is not already known). The MS responds to the paging message by initiating a procedure for page response very similar to the packet access procedure described earlier. If the location is known, the paging procedure is followed by the `Packet Resource Assignment` for downlink frame transfer containing the list of PDCHs to be used. The last procedure is also performed before downlink transfer, if the location of the MS is already known.

3.4.5.4.2 RLC Block Transfer

3.4.5.4.2.1 Uplink

When the MS receives the complete uplink assignment, it begins to monitor the assigned PDCHs for the USF value. If there is already a TBF running, the MS waits for the moment of the TBF starting time, which is specified in the `Packet Uplink Assignment` message. Then the MS starts to use the new parameters. Otherwise, if there is no TBF running, the MS begins to monitor the PDCH for USFs as soon as the starting time expires.

It is possible to set the `RLC-DATA-BLOCKS-GRANTED` information element in the `Packet Uplink Assignment` message to allow the MS to send only a specified number within the TBF.

When transmitting RLC blocks on the uplink, the first three RLC data blocks of the uplink TBF contain a *Temporary Logical Link Identifier* (TLLI) field in the RLC data block header. Each time the MS detects an assigned USF value on an assigned PDCH, one RLC block will be transmitted on the same PDCH in the next block period.

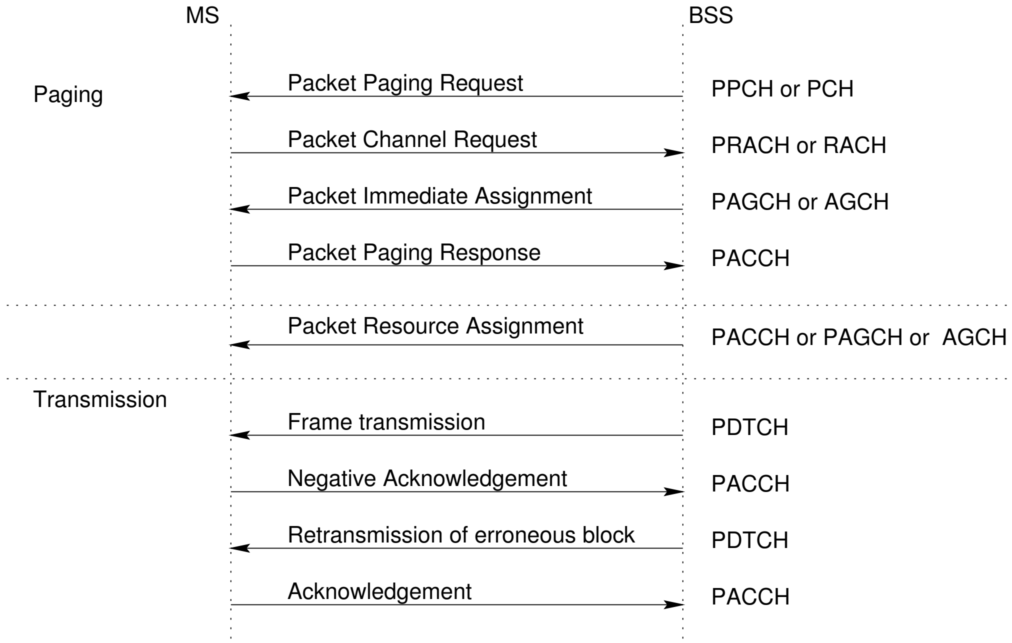


Figure 3.14: Downlink TBF establishment and data transmission

3.4.5.4.2.2 Downlink

After the reception of the `Packet Downlink Assignment` message, the MS starts the timer `T3190` to define when to stop waiting for valid data from the network.

Receiving a valid RLC data block without the FBI bit set to 1, the MS resets and restarts timer `T3190` to define when to stop waiting for valid data from the network.

3.4.5.5 RLC/MAC Scheduling

RLC/MAC scheduling in the (E)GPRS BSS can be subdivided into three steps: the selection of the traffic class (see Section 2.2.2.2 and Section 3.5.3.3), scheduling of the next TBF inside the selected traffic class and scheduling of the next RLC block of the selected TBF (see Figure 3.4.5.5).

3.4.5.5.1 Traffic Class Scheduling

The MAC scheduler classifies the incoming radio resource requests of established TBFs regarding the application and subscription of the MSs (see Section 3.5.3). For example, the TBF can be classified in one of three subscriber classes, *Gold* service, *Standard* service and *Best-effort* service. In case the TBF belongs to a *Standard* subscriber it might additionally be classified according to the application QoS profile to one of the four standard traffic classes, *Conversational*, *Streaming*, *Interactive*, or *Background*. Within the resulting six traffic class queues which are also called *TBF queues* only the identifiers of the TBFs are registered. The traffic class scheduler only has the information about TBFs which are requesting a data transfer and does not have information on the amount of data to be transmitted for each TBF. The traffic class scheduler selects a traffic class queue to be served by applying a class scheduling algorithm. This algorithm is not specified in the standard and can be optimized by the system designer. The algorithm can be, e.g. a priority algorithm or a bandwidth sharing algorithm.

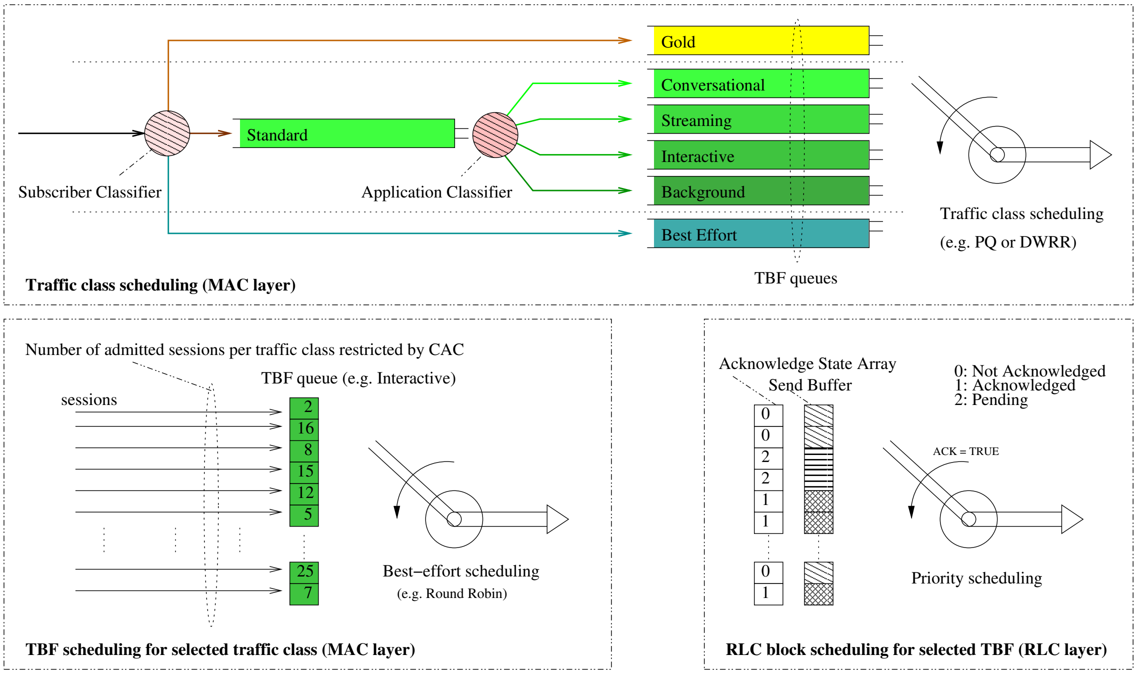


Figure 3.15: Scheduling of traffic classes, TBFs and RLC blocks

3.4.5.5.2 TBF Scheduling

Once a traffic class queue containing all TBF identifiers of this traffic class has been selected by the traffic class scheduler, the TBF scheduler selects one TBF of this TBF queue applying the TBF scheduling algorithm. This algorithm is also implementation-specific. As an example a *Round Robin* (RR) algorithm can be applied. The TBF scheduler only has the information that a TBF is established and has neither information on the amount of data to transmit nor if the TBF actually has data available. So the scheduler starts with the first TBF listed in the queue and checks if it has been allocated to the regarded PDCH. If not, the scheduler continues with the following TBF. In case the TBF is able to use the regarded PDCH the related RLC entity is polled for data until it reaches the predefined RR quantum or until it has no more radio blocks to transmit. Then the following TBF of the same class queue is served if the same traffic class is still selected by the traffic class scheduler. Typically the RR quantum is in the order of 1–20 radio blocks.

3.4.5.5.3 RLC Block Scheduler

In the third step, the RLC entity which has been polled for data by the MAC scheduler, checks if there are any data blocks available in the transmit buffer.

In case of *RLC acknowledged mode* the elements in V(B) indicate the acknowledgement status of related RLC data blocks. There are three possible states for each RLC data block:

- `NACK` indicates an RLC block which has not been transmitted yet, which has been negatively acknowledged or which has an expired timer
- `PENDING_ACK` indicates an RLC block which has been sent, but no acknowledgement has been received for this block yet
- `ACK` indicates data which has been sent and has already been acknowledged

The RLC block scheduling algorithm determines the order of transmission of the RLC blocks inside the RLC send buffer of a regarded TBF. The RLC data blocks in the RLC transmit window with the acknowledge state `NACK`, are forwarded to the MAC starting with the oldest one. If no `NACK` data block exists, the oldest RLC data block with the acknowledge state `PENDING_ACK` is retransmitted.

The priority of `NACK` blocks to `PENDING_ACK` blocks inside one RLC entity is specified in the standard [3GPP TSG GERAN (2002d)]. It is also specified that `PENDING_ACK` blocks should be transmitted if a radio block period is scheduled for the regarded TBF and if no `NACK` block exists for this TBF. A decoupled implementation of RLC and MAC leads to the transmission of `PENDING_ACK` blocks, while other TBFs still could have `NACK` blocks to transmit that are more urgent. This gives the motivation to implement an RLC/MAC layer with a MAC TBF scheduler that serves TBFs with `NACK` RLC blocks ahead of TBFs with only `PENDING_ACK` blocks, which is consistent with the GPRS standard.

3.4.5.6 Channel Coding Schemes

The redundancy added to the RLC/MAC blocks can be adapted by the operator to the channel conditions in the network. The GPRS channel *Coding Schemes* (CSs) CS-1 to CS-4 enable code rates from 1/2 to 1 (see Table 3.3). The RLC data block size depends on the number of LLC PDUs concatenated in the radio block. In this table one LLC PDU per radio block is assumed. CS-1 has to be used for RLC/MAC control blocks after the GPRS specification. Although all CSs can be used for RLC/MAC data blocks, only CS-2 is used for RLC/MAC data blocks in operational GPRS networks.

In the first step of the coding procedure a BCS is added for error detection. The second step consists of precoding the USF (except for CS-1), adding four tail bits, convolutional coding for error correction and puncturing to realize the desired code rate. The result is a bit pattern with a length of 456 bit independent of the coding scheme used. For CS-4 no tail bit is added and no convolutional coding (and puncturing) is performed; only the

Table 3.3: Coding parameters

	CS-1	CS-2	CS-3	CS-4
Code rate	1/2	2/3	3/4	1
GPRS net data rate [kbit/s]	9.05	13.4	15.6	21.4
RLC data block size [bit]	176	263	307	423
MAC header size (ex. USF) [bit]	5	5	5	5
RLC/MAC block size (ex. USF) [bit]	181	268	312	428
USF [bit]	3	3	3	3
BCS [bit]	40	16	16	16
Precoded USF [bit]	3	6	6	12
Tail bits	4	4	4	0
Radio block [bit] (before conv. coding)	228	294	338	456
Convolutional code rate	1/2	1/2	1/2	1
Coded radio block size [bit]	456	588	676	456
Puncturing [bit]	0	132	220	0
Radio block size [bit]	456	456	456	456

USF is precoded [3GPP TSG GERAN (2002e)].

In GPRS a punctured convolutional code that was specified for GSM is used [Stuckmann (2002b)]. The decoding of the USF is simplified whereby for coding CS-2–4 a 12-bit long code word, which is not punctured, is generated for the USF. With CS-2 and CS-3 the USF is precoded with 6 bit before the frame is convolutionally coded, with the first 12 bit not being punctured. When CS-1 is used, the entire frame is coded and the USF has to be decoded as part of the information field. Figure 3.16 shows an example of the coding of an information frame for CS-3.

The GPRS *net data rate* R_{net} is defined by the data rate of RLC/MAC blocks considering the channel coding overhead and considering the idle frames. Since one 52 multiframe comprises 12 radio block periods, the mean transmission duration of a radio block D_{RB} can be calculated as:

$$D_{\text{RB}} = \frac{52\ frames \cdot 4.615\ ms}{12\ frames} = 20\ ms \tag{3.1}$$

With X as the number of bits in one RLC/MAC block from Table 3.3 (RLC/MAC block size) the net data rate can be calculated as:

$$R_{\text{net}} = \frac{X\ bit}{D_{\text{RB}}} \tag{3.2}$$

For example:

$$R_{\text{net(CS-1)}} = \frac{181\ bit}{20\ ms} = 9.05\ \frac{kbit}{s} \tag{3.3}$$

The net data rate in GPRS does not equal the data rate that the RLC/ MAC layer offers to the LLC layer. The latter can be calculated based on the number of RLC data block information bytes from Table 3.2.

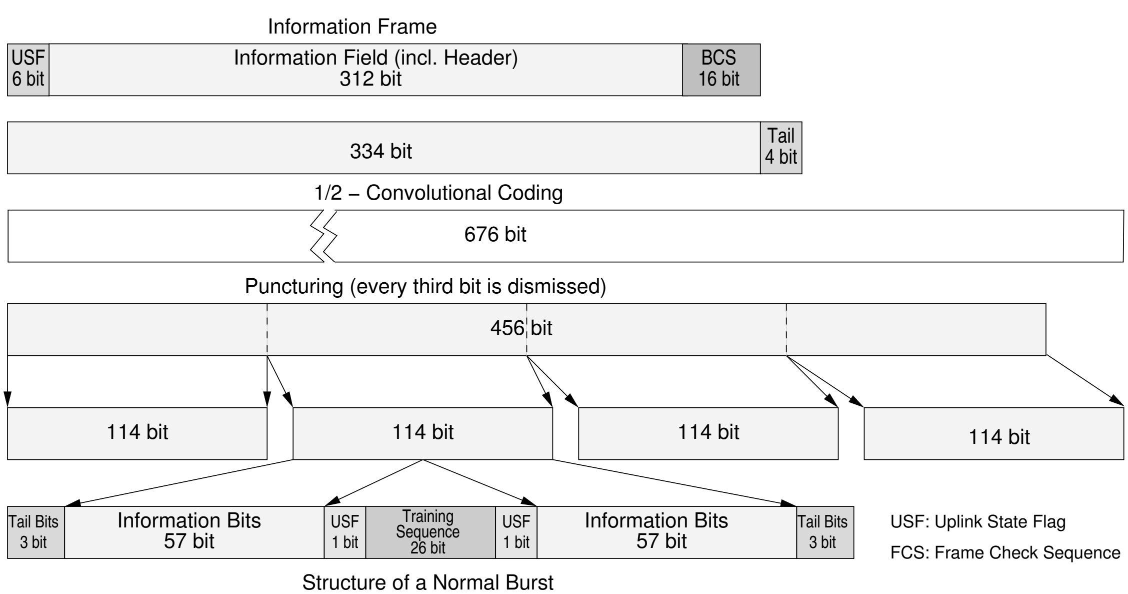


Figure 3.16: Coding of an information frame according to CS-3

3.4.6 Physical Layer (PL)

The *Physical Layer* (PL) at the air interface is divided into the *Physical Link Layer* (PLL) and the *Radio Frequency* (RF) layer. These sublayers are specified in [3GPP TSG GERAN (2002a,b)]. While in the RF sublayer mainly modulation and demodulation are carried out, the PLL sublayer provides services for data transmission over the radio interface. The PLL sublayer is responsible for the *Forward Error Correction* (FEC) coding, allowing the detection and correction of erroneous transmitted code words and the indication of uncorrectable code words. It is also responsible for the interleaving of one radio block over four bursts in consecutive TDMA frames and provides procedures for synchronization.

3.5 GPRS Control Plane

The GPRS protocol architecture is organized into two planes, the user plane, also called transmission plane, and the control plane, also called signaling plane. The user plane is responsible for data transmission, when packet data transfer is actually requested in uplink or downlink. To realize the transfer between the correct network nodes with adequate performance characteristics, the user plane protocols need certain information such as addresses of peer entities, the status of network elements or requested protocol options. This information has to be provided at the beginning of a session and has to be kept up-to-date during an ongoing session, while the mobile user might move around with his terminal or use different services in parallel or consecutively with idle periods in between.

The GPRS control plane realizes these management functions, namely *GPRS Mobility Management* (GMM), *Session Management* (SM) and *Quality of Service* (QoS) management. GMM keeps track of the MS's location and its state, updates databases with this information and supports cell change procedures. The SM manages the logical context between the MS and an external *Packet Data Network* (PDN) based on TCP/IP or X.25, when a new session is set up or changing its performance requirements. In GPRS networks different applications with different performance requirements are supported. QoS management in GPRS networks enables the network to establish QoS contracts with the MS differentiating applications and subscribers according to their QoS requirements. The GPRS QoS support is based on the SM procedures.

In Figure 3.17 the GPRS control plane between MS and SGSN is shown. The signaling layers SM and GMM reuse the layer-2 protocols and the GPRS channel interface of the user plane (see Section 3.4) to transmit signaling messages over the air interface.

3.5.1 Mobility Management

Mobility management in GSM and GPRS [3GPP TSG SSA (2002a)] includes functions for location registration with the PLMN and location updating to report the current location of an MS. Additionally it is responsible for identification and authentication of subscribers.

The GMM concept allows combined procedures for GSM and GPRS to reduce the signaling load. The handover procedure is split up into the cell update and the *Routing Area* (RA) update.

In contrast to the GSM system, where the handover decision is made by the network based on measurement reports sent by the MS [3GPP TSG CN (2001)], the cell change in GPRS is not necessarily prepared and triggered by the network, but is mostly decided

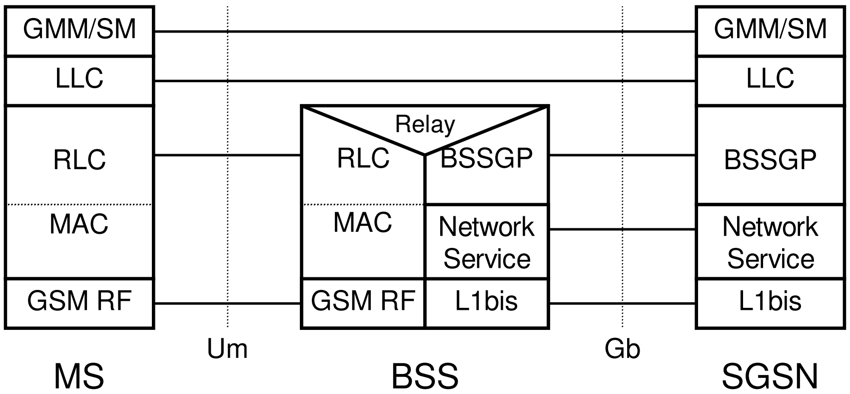


Figure 3.17: GPRS control plane (MS - SGSN)

by the MS, when it detects that it has entered a new cell or even a new RA by reading the system information on the broadcast channel.

3.5.1.1 Location Area and Routing Area

The *Location Area* (LA) and *Routing Area* (RA) are different areas comprising a certain number of cells (see Figure 3.18). For example one LA may comprise 30 cells. This hierarchical structure with LAs is chosen for GSM networks to manage the location of an MS. The network initially only knows the LA in which the MS is located. The MS provides this information periodically, or when the MS changes its LA. When it becomes necessary to determine the exact cell location of an MS, the network sends a paging command in all cells of the LA. The MS, which continuously monitors the broadcast channel, responds to this paging command so that a connection or a context can be established. With this procedure the signaling load on the air interface is minimized, since the network need not know about the exact cell location of the MS when it is not active. Additionally, the VLR does not have to be updated by the SGSN, as long as the MS is staying inside one LA.

The RA that does not exist in GSM networks without GPRS support was introduced to speed up the paging procedure and lower the signaling load for paging. The number of cells in an RA is generally less than or equal to that of an LA. For example, one LA with 30 cells may comprise three RAs with 10 cells each. An RA can not span more than one LA and an RA is served by only one SGSN. One SGSN can handle several RAs and the size of an RA can range from a part of a city to an entire province or even a small country. Figure 3.18 shows an example cell scenario, in which it is possible to see how cells could be grouped in a cell planning configuration. Cells 1, 2 and 3 together build one RA. Cell 4 alone builds another distinct RA. Together these two RAs form one LA. A set of LAs are served by one BSC, and many BSCs are served by an SGSN. The SGSNs are interconnected and represent the highest level in the GMM hierarchy. Cell 14 belongs to another SGSN in this example.

After an MS has changed from one cell to another it reads the system information of the new cell. Since the RA is broadcast as part of the system information, the MS can determine whether it has changed its RA. If this is the case, the MS sends an RA update to the SGSN to inform it about the new RA.

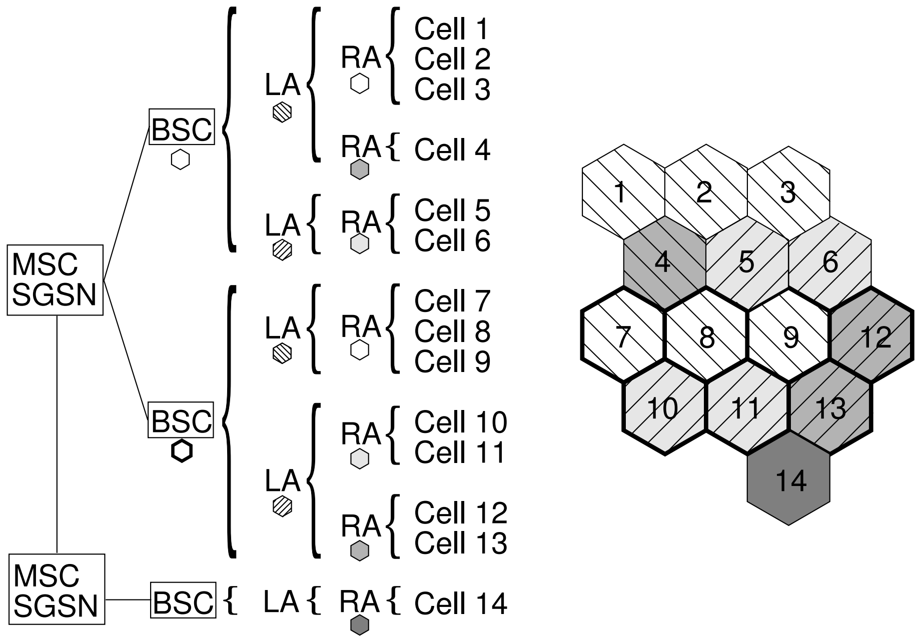


Figure 3.18: Example of a cell scenario comprising all cell change possibilities

3.5.1.2 Mobility Management Procedures

The *GPRS Mobility Management* (GMM) procedures comprise access control and authentication functions. To obtain access to GPRS services the MS has to initiate the GPRS attach procedure. During *Mobility Management* (MM) procedures user data can, in general, be transmitted while the signaling is going on. This can lead to loss of data during attach, authentication and RA update procedures. In order to minimize retransmissions the MS and SGSN should not transmit user data during these procedures.

3.5.1.3 States of the Mobility Management

The GMM functionalities are based on three different MM states defined at the MS and the SGSN states, in which each MS can be. These MM states are:

1. The `idle` state.
2. The `standby` state.
3. The `ready` state.

In each of these MM states the MS and the SGSN hold a different set of information about the GPRS terminal. These are denoted MM contexts. Figure 3.19 shows the different states and the related transitions. The MM context in the MS comprises among others the MM state MS identifiers, location and security data and zero or more PDP contexts, while the MM context at the SGSN additionally contains classmarks for the communication with the registers. The following states are defined.

In the idle state the MM context is empty, because the subscriber is not attached to the GPRS network. The MS and SGSN do not hold any valid information about the cell and RA in which the MS is located. Therefore no MM procedures are performed. The MS is seen as not reachable and can not be paged.

In the standby state the MS is attached to the network. Each of the MS and SGSN have established an MM context. Both the MS and the SGSN know the RA in which the

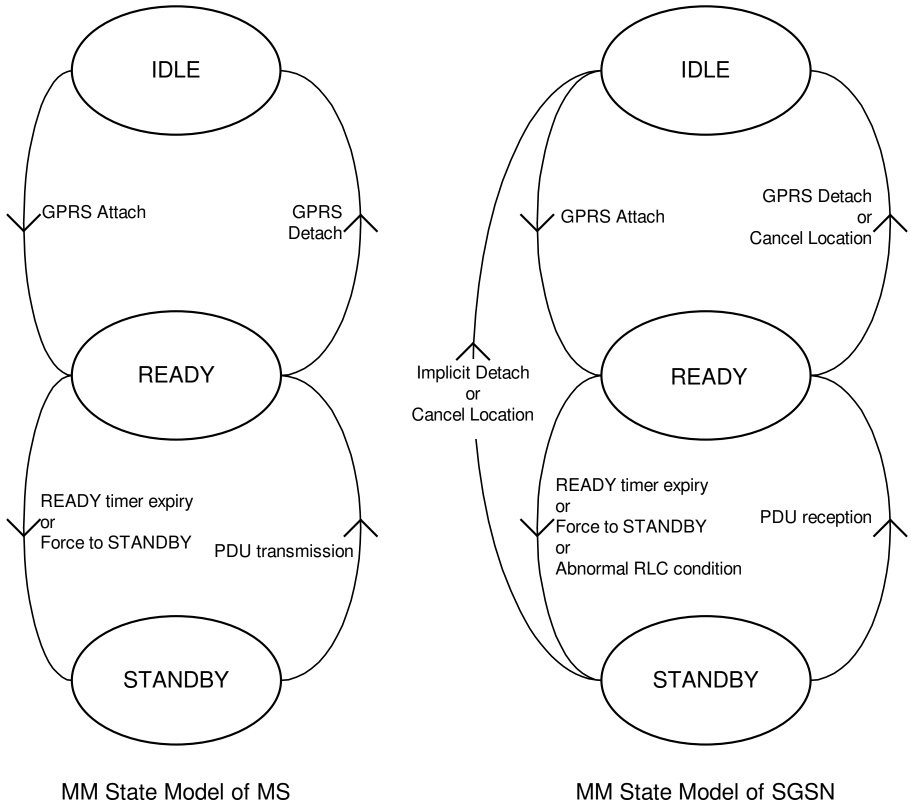


Figure 3.19: MM state transitions

MS is located. In this state the SGSN can send paging messages to all the cells of the RA to find the MS, when it messages have to be sent in the downlink direction. The MS may receive paging messages, but can not receive or transmit any data.

In the ready state the SGSN knows in which cell the MS is located and can therefore send a continuous stream of data packets on the downlink, even if the MS changes its cell supported by *Location Management* (LM) procedures (see Section 3.5.1.4). The MM context is practically the same as in the case of the `standby` state but extended by the information on the cell in which the MS is located. In this state the SGSN need not initially page the MS to send data. The MS can activate or deactivate PDP contexts. The MM context remains in the `ready` state even if there is no data to send. It returns to the `standby` state after the expiry of the `ready` timer.

3.5.1.4 Location Management Procedures

The *Location Management* (LM) function controls the cell and PLMN selection and provides a mechanism which enables the SGSN to know the RA of the MS in the `standby` and `ready` states [3GPP TSG SSA (2002a)]. If an MS is in the `ready` state, the SGSN knows the location of the MS on cell level.

When an MS enters a new cell, three different procedures can be performed:

1. A cell update procedure
2. An RA update procedure (intra- or inter-SGSN)

In Figure 3.20 all the cell change possibilities are shown. These can be derived from Figure 3.18, in which a cell planning scenario is shown. The first possibility is the intra-

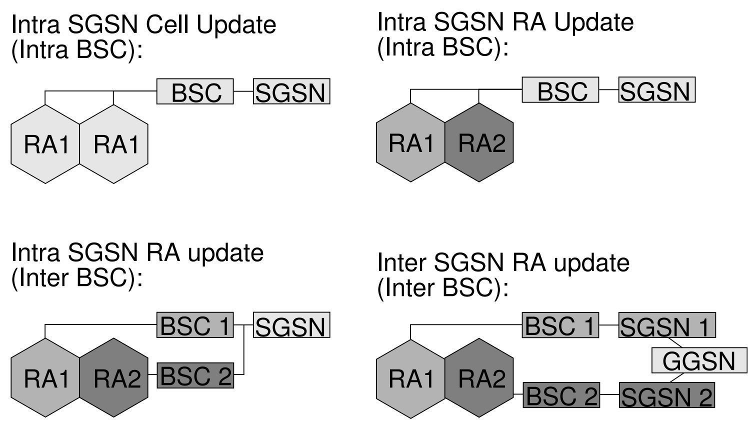


Figure 3.20: All cell change possibilities in GPRS

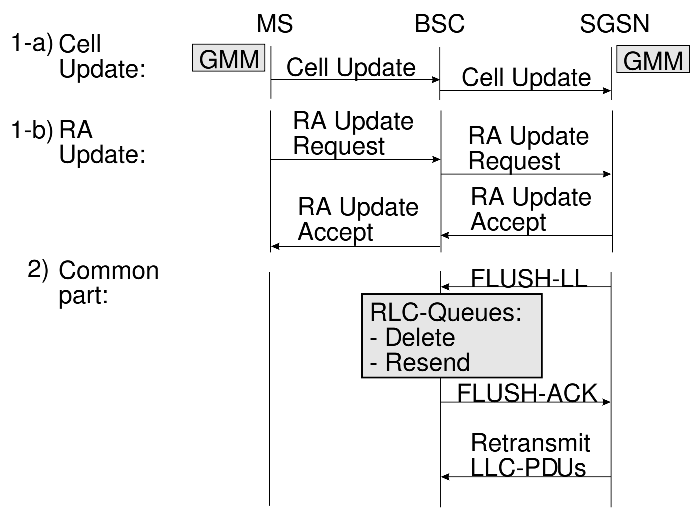


Figure 3.21: Message sequence chart of the cell change procedure

SGSN cell update intra-BSC where an MS changes between two cells belonging to the same RA served by one BSC. In this case, the MS performs a cell update procedure. The second possibility is the intra-SGSN RA update intra-BSC. Here the RA changes from one cell to the other. The MS has to perform an RA update procedure. In the third case the responsible BSC changes (intra-SGSN RA update inter-BSC). The fourth and last case is the inter-SGSN RA update where also the responsible SGSN changes.

3.5.1.5 Cell Change Signaling

Figure 3.21 shows a simplified overview of the cell change procedures in GPRS. Basically, two cell change possibilities are foreseen. The first one (1-a) is the cell update, where the MS notifies the SGSN GMM about a cell change by sending an uplink PDU in the new cell it enters. The second one (1-b) is the RA update, where the MS sends an `RA Update Request` message in the new cell it enters and must wait for a response from the GMM. The response shown in Figure 3.21 is an `RA Update Accept` message, but it also might be an `RA Update Reject` message.

Then, there is a common part (2) for both the cell update and the RA update where the GMM decides with the `FLUSH-LL` message to delete or to forward RLC PDUs queued in the old BSC [3GPP TSG GERAN (2002c)]. This message is acknowledged with the message `FLUSH-ACK`. After this response the SGSN can retransmit the LLC-PDUs immediately, if they have been deleted at the BSC.

3.5.2 Session Management

Session Management (SM) in GPRS comprises all signaling functions necessary for the access to external *Packet Data Networks* (PDNs). For interworking with external IP networks the IP interworking model is defined in [3GPP TSG CN (2002d)].

3.5.2.1 The IP Interworking Model

Network interworking is required whenever a PLMN is involved in communications with another network to provide end-to-end transport. The PLMN interconnects in a manner consistent with that of a *Packet Data Network* (PDN) defined by the requirements for interworking with *Public Switched Data Networks* (PSDNs) X.75 [ITU-T (1996a)].

GPRS supports interworking with networks based on the Internet Protocol. These interworked networks may be either intranets or the Internet [3GPP TSG CN (2002d)]. When interworking with IP networks, GPRS can operate IPv4 or IPv6. The interworking point with IP networks is at the G_i reference point as shown in Figure 3.22. The GGSN is the access point of the GPRS network. In this case the GPRS network will look like any other IP network or subnetwork.

The G_i reference point is between the GGSN and the external IP network. From the external IP network's point of view, the GGSN is seen as an ordinary IP router. The access to the Internet, an intranet or *Internet Service Provider* (ISP) may involve specific functions such as user authentication, user's authorization, end-to-end encryption between MS and Intranet/ISP and the allocation of a dynamic address belonging to the PLMN/Intranet/ISP addressing space. For this purpose the GPRS PLMN may offer:

- either direct transparent access to the Internet
- or a non-transparent access to an intranet/ISP.

In the transparent case the mobile station is given an address belonging to the operator's addressing space (see Figure 3.22). The address is a public IP address given either at subscription in which case it is a static address or at PDP context activation in which case

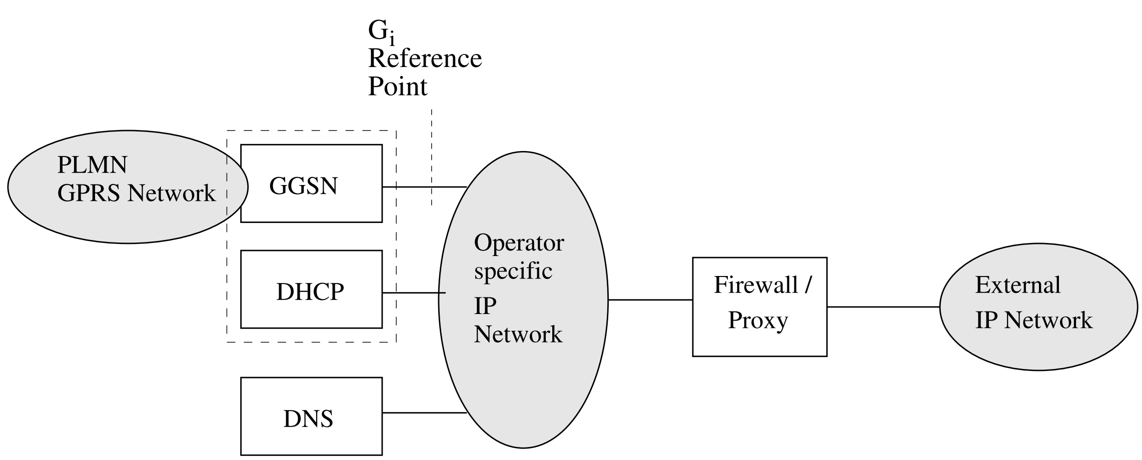


Figure 3.22: IP interworking

it is a dynamic address. This address is used for packet forwarding between the Internet and the GGSN and within the GGSN to tunnel user data through the core network.

The transparent case provides at least a basic ISP service. As a consequence of this it may therefore provide a bearer service for a tunnel to a private Intranet. In the non-transparent case the mobile station is given an address belonging to the Intranet/ISP private addressing space, static or dynamic. Packet forwarding within the GGSN and on the Intranet/ISP requires a link between the GGSN and an address allocation server, like *Remote Authentication Dial In User Service* (RADIUS) and *Dynamic Host Configuration Protocol* (DHCP) belonging to the Intranet/ISP. This server responses to authentication requests at a PDP context activation.

3.5.2.2 PDP Context Handling

The main function of *Session Management* (SM) [3GPP TSG CN (2002e)] is to support PDP context handling of the MS.

QoS profile negotiation between MS and the network is performed within the procedures for the activation or modification of PDP contexts. Therefore, SM procedures play an important role in the GPRS mechanisms for QoS support.

3.5.2.3 Session Management Procedures

The purpose of the PDP context activation procedure is to establish a PDP context between the MS and the network for a specific QoS on a specific N-SAPI (see Section 3.4.3). The PDP context activation may be initiated by the MS or the initiation may be requested by the network.

3.5.2.3.1 PDP Context and Traffic Flow Template

In GPRS Release 97/98, there is only one PDP context possible per MS. In GPRS Release 99, each PDP address may be described by one or more PDP contexts in the MS or the network. The first PDP context activated for a PDP address is called the *primary PDP context*, whereas all additional contexts associated with the same PDP address are called *secondary PDP contexts*. When more than one PDP context is associated with a PDP address, there will be a *Traffic Flow Template* (TFT) for each additional context. The TFT will be sent transparently via the SGSN to the GGSN to enable packet classification and policing for downlink data transfer [3GPP TSG SSA (2002a)]. A TFT consists of from one up to eight packet filters, each identified by a unique packet filter identifier. A packet filter has an evaluation precedence index that is unique within all TFTs associated with the PDP contexts that share the same PDP address. The MS manages packet filter identifiers and their evaluation precedence indexes, and creates the packet filter contents. The packet filter specifies attributes for incoming and outgoing IP packets. The IP datagram attributes have to match at least for the packet filter of one evaluation precedence index of the TFT to pass the GGSN.

3.5.2.3.2 PDP Context Information Element

All PDP context information is stored within an appropriate *Information Element* (IE) [3GPP TSG CN (2002c)]. In Figure 3.23 the fields corresponding to QoS Subscribed (QoS Sub), QoS Requested (QoS Req), and QoS Negotiated (QoS Neg), as well as the PDP Address, are emphasized. They play an important role in the implementation of the mechanisms to support QoS.

Octet \ Bit	8	7	6	5	4	3	2	1
1	Type = 130 (Decimal)							
2–3	Length							
4	Res–erved	AA	Res–erved	Order	NSAPI			
5	X	X	X	X	SAPI			
6–8	QoS Sub							
9–11	QoS Req							
12–14	QoS Neg							
15–16	Sequence Number Down (SND)							
17–18	Sequence Number Up (SNU)							
19	Send N–PDU Number							
20	Receive N–PDU Number							
21–22	Uplink Flow Label Signalling							
23	Spare 1 1 1 1				PDP Type Organization			
24	PDP Type Number							
25	PDP Address Length							
26–m	PDP Address [1..63]							
m+1	GGSN Address for signalling Length							
(m+2)–n	GGSN Address for signalling [4..16]							
n+1	APN length							
(n+2)–o	APN							
o+1	Spare (sent as 0 0 0 0)				Transaction Identifier			

Figure 3.23: PDP context information element

All QoS-related information, whether subscribed, requested, or negotiated, is included in a QoS profile. The associated IE is described in Section 3.5.3.2.

3.5.3 Quality of Service Management

As mentioned in Section 3.4.2, Section 3.4.5.5 and Section 3.5.2, there are several logical entities and protocols involved in GPRS QoS management.

PDP contexts as well as QoS profiles are negotiated between MS and SGSN. The BSS is provided with a *Packet Flow Context* (PFC) containing the *Aggregate BSS QoS Profile* (ABQP) (see Section 3.4.2) and is responsible for resource allocation on a TBF base and scheduling of packet data traffic with respect to the relevant QoS profiles negotiated. Moreover, it has to regularly inform the SGSN about the current load conditions in the radio cell. The tasks of the GGSN comprise mapping PDP addresses as well as classification of incoming traffic from external networks on behalf of downlink TFTs. The *GPRS Registers* (GRs) hold the QoS-related subscriber information and deliver it, on demand, to the *GPRS Support Nodes* (GSNs).

In Figure 3.24, part of a GPRS session is schematically outlined, depicting the instances involved, messages exchanged, and parameters negotiated for PDP context, PFC and TFT

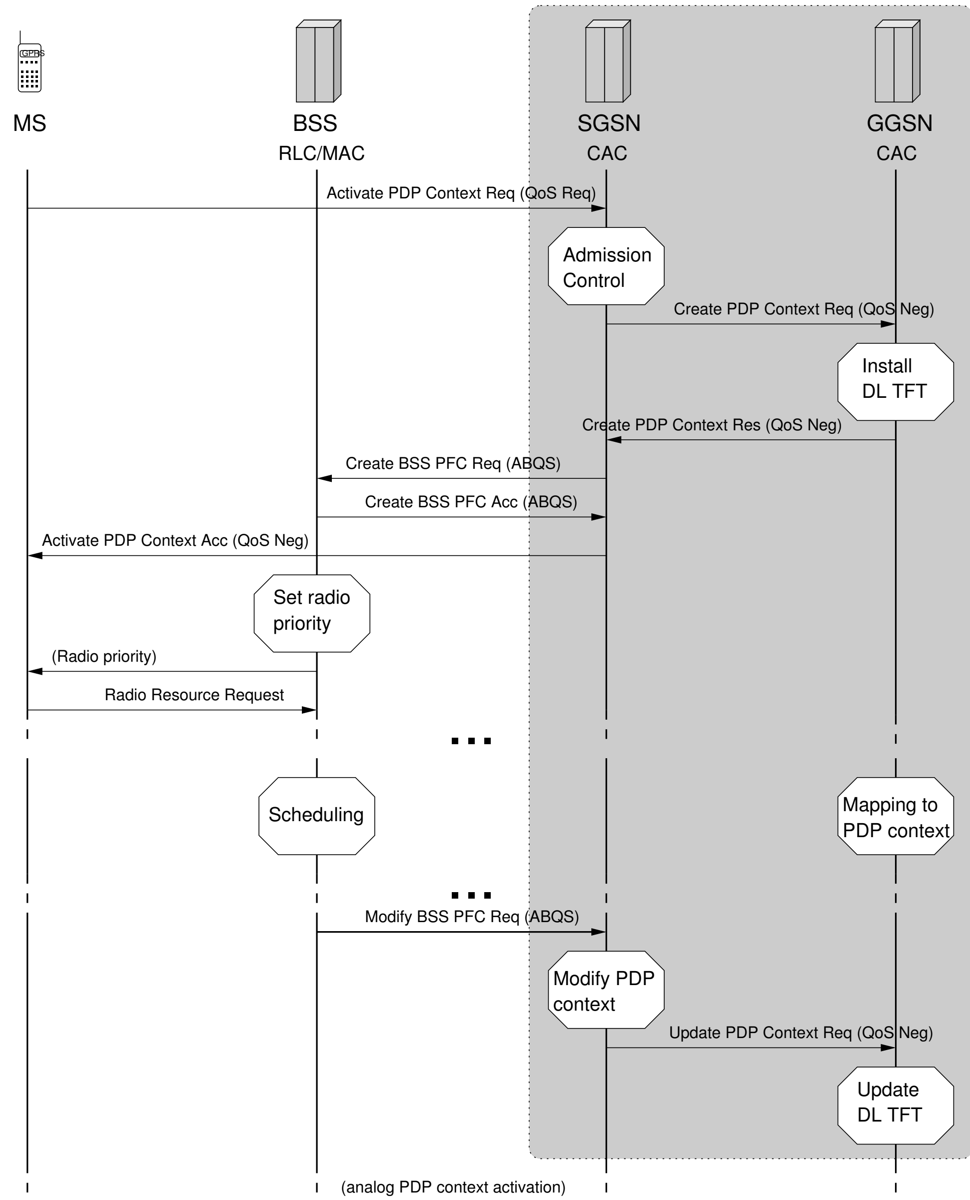


Figure 3.24: QoS negotiation and renegotiation procedures (example)

installation and renegotiation.

From a time-scale point of view, the mechanisms for QoS management within the GPRS might be regarded as a three-stage model (see Figure 3.25). On PDP context activation, the QoS parameters are negotiated. As long as the PDP context remains active, these parameters should be guaranteed, unless there is a QoS renegotiation. The QoS profile is considered both for each TBF and for each radio block period. At TBF setup, radio resources, like the number of PDCHs, the associated USFs and TFIs, are assigned according to the negotiated QoS parameters. During the TBF, radio blocks are scheduled at the BSS in competition with other existing TBFs in the radio cell. This scheduling function has to be done considering the QoS profiles of the PDP contexts associated with the TBFs [Baig et al. (2001); Stuckmann (2002a); Stuckmann and Müller (2001)].

A QoS profile can be considered as a single parameter value that is defined by a unique combination of attributes. There are numerous QoS profiles available based on

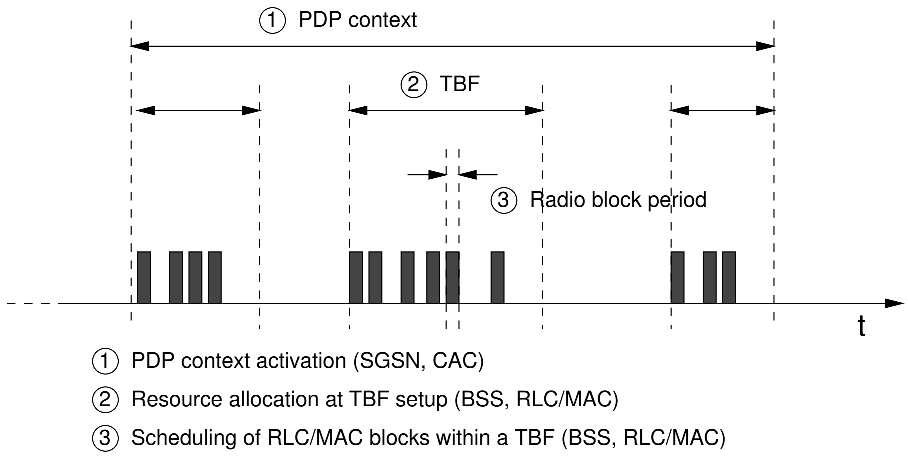


Figure 3.25: Three-stage model of QoS management

permutations of these different attributes, but each mobile network operator must choose to support only a limited subset, reflecting their planned range of GPRS subscriptions. In the following subsections, the QoS attributes defined in GPRS Release 97/98, as well as the changes made for Release 99, will be explained.

3.5.3.1 QoS Attributes According to GPRS Release 97/98

A QoS profile defines the QoS within the range of the following service classes [3GPP TSG SSA (2001a, 2002a)].

3.5.3.1.1 Precedence Classes

Under normal circumstances the network should try to meet all profiles' QoS agreements. The precedence specifies the relative importance to keep the conditions even under critical circumstances, e.g., momentarily high network load. The various precedence classes are presented in Table 3.4.

3.5.3.1.2 Delay Classes

The packet delay is defined by the time needed for transmission from one GPRS SAP to another. Delays outside the system, e.g., in transit networks, are not considered. [3GPP TSG SSA (2001a)] determines four delay classes (see Table 3.5). The network operator has to provide for convenient resources on the air interface to be able to serve the number of participants with a certain delay class expected within each cell. Although there is no need for all delay classes to be available, at least *best effort* has to be offered.

Table 3.4: Precedence classes

Precedence class	Identifier	To be served
1	High priority	preferably before classes 2 and 3
2	Normal priority	preferably before class 3
3	Low priority	without preference

Table 3.5: Delay classes

Delay class	128 byte packet		1 024 byte packet	
	Mean delay [s]	95 % [s]	Mean delay [s]	95 % [s]
1 (predictive)	0.5	1.5	2	7
2 (predictive)	5	25	15	75
3 (predictive)	50	250	75	375
4 (best effort)	unspecified			

Table 3.6: Reliability class

Reliability classes	GTP mode	LLC frame mode	LLC data mode	RLC block mode	Traffic type security
1	ACK	ACK	PR	ACK	NRT traffic, error sensitive, loss sensitive
2	UACK	ACK	PR	ACK	NRT traffic, error sensitive, slightly loss sensitive
3	UACK	UACK	PR	ACK	NRT traffic, error sensitive, not loss sensitive
4	UACK	UACK	PR	UACK	RT traffic, error sensitive, not loss sensitive
5	UACK	UACK	UPR	UACK	RT traffic not error sensitive, not loss sensitive
(U)ACK	(Un)acknowledged			NRT	Non-Realtime
PR/UPR	Protected/Unprotected			RT	Realtime

3.5.3.1.3 Reliability Classes

Data services generally require a low residual *Bit Error Ratio* (BER). Erroneous data is usually useless, while incorrectly received speech only leads to a worse perception. Reliability of data transmission is defined within the scope of the following cases:

- probability of loss of data
- probability of out-of-sequence data delivery
- probability of multiple delivery of data and
- probability of erroneous data.

The reliability classes specify the requirements for each layer's services. The combination of different modes of operation of the GPRS specific protocols GTP, LLC, and RLC, explained in Section 3.4, support the reliability requirements of various applications, e.g., *Real-Time* (RT) or *Non Real-Time* (NRT). The reliability classes are summarized in Table 3.6.

3.5.3.1.4 Peak Throughput Classes

User data throughput is specified within the scope of a set of throughput classes that characterize the expected bandwidth for a requested PDP context. It is defined by the choice of peak and mean throughput class. Peak throughput is measured in byte/s at the reference point G_i and in the terminal (see Figure 3.1). Peak throughput specifies the

Table 3.7: Peak throughput classes

Peak throughput class	Peak throughput [byte/s]	Peak throughput [kbit/s]
1	up to 1 000	8
2	up to 2 000	16
3	up to 4 000	32
4	up to 8 000	64
5	up to 16 000	128
6	up to 32 000	256
7	up to 64 000	512
8	up to 128 000	1 024
9	up to 256 000	2 048

Table 3.8: Mean throughput classes

Mean throughput class	Mean throughput [byte/h]	Mean throughput ≈ [bit/s]
1	100	0.22
2	200	0.44
3	500	1.11
4	1 000	2.2
5	2 000	4.4
6	5 000	11.1
7	10 000	22
8	20 000	44
9	50 000	111
10	100 000	220
11	200 000	440
12	500 000	1 110
13	1 000 000	2 200
14	2 000 000	4 400
15	5 000 000	11 100
16	10 000 000	22 000
17	20 000 000	44 000
18	50 000 000	111 000
31	Best effort	

maximum rate at which data is transmitted within a certain PDP context. No guarantee is given that this data rate is actually achieved at any time during transmission. Rather, this depends on the resources available and the capabilities of the MS. The operator may limit the user data rate to the peak data rate agreed on, even if there is capacity left for disposal. The peak throughput classes are presented in Table 3.7.

3.5.3.1.5 Mean Throughput Classes

Like peak throughput, mean throughput is also measured in byte/s at the reference points G_i and R. It specifies the average rate data transmitted within the time remaining for a certain PDP context. The operator may limit the user data rate to the mean data rate negotiated, even if excessive capacity is available. If *best effort* has been agreed on as the throughput class, throughput is made available to a MS whenever there are resources needed and at its disposal. Table 3.8 summarizes the classes of mean throughput.

3.5.3.2 QoS Profile Information Element

All QoS-related information to be exchanged between MS and SGSN is stored in a QoS profile. The appropriate IE is shown in Figure 3.26. It consists of an *Information Element Identifier* (IEI), a length field, five fields that contain the values of the service classes (see Section 3.5.3.1), and three fields filled with spare bits [3GPP TSG CN (2002e)]. IEI and length fields are not part of a PDP context IE (see Section 3.5.2.3.2).

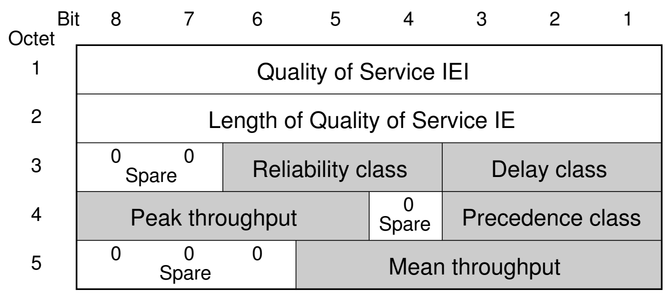


Figure 3.26: QoS profile information element

3.5.3.3 QoS in GPRS Release 99

It is evident from the above description that the QoS architecture defined in GPRS Release 97/98 shows some major drawbacks (see also [Gudding (2000); Stuckmann and Müller (2001)]):

1. The BSS is not aware of the negotiated QoS profile. This restricts the ability of the BSS to perform scheduling and resource management on the radio interface.
2. Neither MS nor GGSN can influence the QoS profile, even if they detect congestion in external networks as well as changing radio conditions or varying application requirements.
3. It is only possible to have one QoS profile for every PDP context that is associated with a specific service access. Thus, there can only be one QoS profile utilized by all applications for one PDP address.

GPRS Release 99 defines several further QoS parameters with finer grained properties to meet requirements on different levels of service for applications [3GPP TSG SSA (2002b)]:

- maximum bitrate
- delivery order
- SDU format information
- residual bit error ratio
- transfer delay
- allocation/retention priority
- guaranteed bitrate
- maximum SDU size
- SDU error ratio
- delivery of erroneous SDUs
- traffic handling priority
- source statistics descriptor ('speech'/'unknown')

Additionally, there are four distinct *traffic classes* introduced, with different parameters specifying their QoS requirements (see Table 3.9):

Table 3.9: End-user performance expectations for selected services belonging to different traffic classes [3GPP TSG SSA (2001b)]

Traffic class	Medium	Application	Data rate (kbit/s)	One-way delay
Conversational	Audio	Telephony	4–25	$< 150\,\mathrm{ms}$
	Data	Telnet	< 8	$< 250\,\mathrm{ms}$
Streaming	Audio	Streaming (HQ)	32–128	$< 10\,\mathrm{s}$
	Video	One-way	32–384	$< 10\,\mathrm{s}$
	Data	FTP	—	$< 10\,\mathrm{s}$
Interactive	Audio	Voice messaging	4–13	$< 1\,\mathrm{s}$
	Data	Web-browsing	—	$< 4\,\mathrm{s/page}$

- *conversational*
- *streaming*
- *interactive*
- *background*

Delay-sensitive services belonging to the *conversational* class, for example, do need absolute guarantees in terms of *guaranteed bitrate* and *transfer delay* attributes, while for *background* traffic only bit integrity is necessary.

In GPRS Release 97/98, the BSS cannot use QoS profile information to schedule resources on a continuous data flow, neither on the downlink, since there is no mechanism provided to download the QoS profile from the SGSN, nor on the uplink, because there is no QoS information available from the MS. With the introduction of Release 99, the BSS is not only provided with QoS profiles on a PFC base, but also with the ability to modify the QoS profile associated with a data flow in case of changing load conditions. Likewise, MS and GGSN may initiate QoS profile renegotiation, either because of changing application requirements, or due to congestion or a change in radio link quality.

Release 99 also solves the Release 97/98 problem of having only one PDP context installed per PDP address; thus all different applications running on top of this PDP address with one single QoS profile, independent of their specific requirements on, e.g., delay or reliability. GPRS Release 99 provides the possibility to install multiple PDP contexts per PDP address. Each PDP context is uniquely associated with a TFT which identifies the traffic flow. This makes it possible to assign different QoS profiles to simultaneous traffic flows, so that each application may receive the appropriate QoS requirement.

CHAPTER 4

Enhanced Data Rates for GSM Evolution

Enhanced Data rates for GSM Evolution (EDGE) is a further development of the GSM data services *High-Speed Circuit-Switched Data* (HSCSD) and GPRS and is suitable for circuit- and packet-switched services. The circuit-oriented part is the *Enhanced Circuit-Switched Data* (ECSD). The packet-oriented part is the *Enhanced General Packet Radio Service* (EGPRS). Applying modified modulation and coding schemes EDGE reaches very high raw bit rates of up to 69 kbit/s per GSM physical channel. If a user utilizes all 8 time slots in parallel, the theoretical maximum raw bit rate rises to 554 kbit/s. The maximum bearer bit rate achievable rises to about 384 kbit/s [Furuskär et al. (1999b)]. EDGE was introduced to the ETSI for the first time in 1997 for the evolution of GSM. After a successful feasibility study of the ETSI the standardization process for EDGE was initiated. Although EDGE was introduced for the evolution of GSM, this concept can be applied to increase the data rate in other systems.

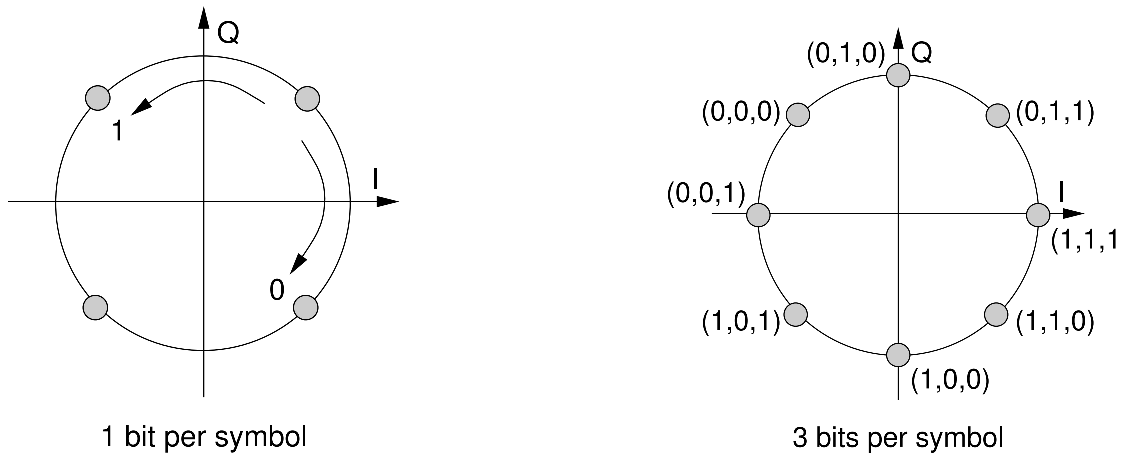
Since the network architecture of the GSM will remain similar for EDGE, the modifications at the air interface are depicted in this chapter. To support higher data rates, a modulation scheme called *8-Phase-Shift-Keying* (8-PSK) is introduced which will not replace the current *Gaussian Minimum Shift Keying* (GMSK) but coexist with it. With 8-PSK it is possible to provide a higher data rate, which is necessary to support bandwidth extensive data applications.

The modifications mostly concern the RLC/MAC layer and the physical layer. Since these protocols are implemented in the MS and the *Base Station* (BS), both have to be modified. In reality, the changes that have to be made comprise a new EDGE transceiver unit and software upgrades to the *Base Station Controller* (BSC), which then can handle standard GSM or GPRS traffic and will automatically switch to EDGE mode when needed.

The core of EDGE is the *Link Quality Control* (LQC) mechanism that allows the adaptation of the *Modulation and Coding Schemes* (MCSs) to a changing radio link quality. Although *Link Adaptation* (LA) is already possible within the GPRS standard, higher CSs are not supported by the actual equipment and will probably only be introduced together with EDGE functionality.

Additionally, a type-II-hybrid ARQ (soft ARQ) scheme is introduced. Soft information is stored during retransmissions to enable *Incremental Redundancy* (IR). The RLC/MAC protocol structure and retransmission mechanism proposed for EGPRS are based on the GPRS standard. An LLC frame is divided into a number of RLC blocks. The average duration of an RLC block is 20 ms, whereas its information content ranges from 176 to 1184 bit, depending on the MCS used. For error detection, the RLC blocks contain a CRC field. Upon reception of an RLC block, the receiver checks the CRC and determines if retransmission of the RLC block is necessary. To give a more detailed introduction, in the following the modulation scheme 8-PSK used for EDGE is described and compared with GMSK currently used for GSM. Additionally the new MCSs are presented. Then the main characteristics of the LQC mechanisms are given in Section 4.3, comprising LA and IR.

Finally the RLC/MAC protocol is adapted to the higher data rates that have to be supported. The RLC parameters *Sequence Number Space* (SNS), *Window Size* (WS) and



(a) GMSK modulation (b) 8-PSK modulation

Figure 4.1: I/Q-diagram of GMSK and 8-PSK modulation

consequentially the *Block Sequence Number* (BSN) field and the *Received Block Bitmap* (RBB) structure are adapted (see Section 3.4.5.3).

4.1 8-PSK Modulation versus GMSK Modulation

The modulation scheme that is used in GSM is called *Gaussian Minimum Shift Keying* (GMSK). In GPRS the same modulation type will be used and high bit rates are achieved through multislot operation. To allow higher bit rates within EGPRS, the *8-Phase-Shift-Keying* (8-PSK) modulation scheme is introduced in addition to GMSK [Furuskär et al. (1999b); Sollenberger et al. (1999)]. The differences between GMSK and 8-PSK are illustrated in an I/Q diagram as shown in Figure 4.1.

Within GMSK transmitting a 0-bit or a 1-bit is represented by incrementing the phase with $\pm 1/2\,\pi$. Every symbol that is transmitted represents one bit, which means that symbol rate and bit rate are equal. The symbol rate of 271 kbit/s is divided over eight time slots. After burst formatting (see Section 3.4.6) a bit rate of 22.8 kbit/s per time slot is achieved. Via different channel coding schemes this results in different net bit rates.

Within 8-PSK three consecutive bits are mapped onto one symbol in the I/Q-plane. With the same symbol rate as in GMSK of 271 kbit/s, bit rates of up to 813 kbit/s can be achieved. With burst formatting, this results in a bit rate of 69.2 kbit/s per time slot. 8-PSK modulation still has the GSM spectrum mask and leaves the burst duration unchanged, which offers the possibility of using both modulation schemes next to each other. However, 8-PSK modulated data is less robust if the channel conditions are bad. The advantage of the new modulation scheme is support for higher data rates under good channel conditions, and to reuse at the same time the channel structure of the GPRS system. Since the *Bit Error Ratio* (BER) depends significantly on the channel conditions, *Link Quality Control* (LQC) is of major importance for a high throughput in EGPRS.

LA as part of LQC gives the possibility of changing to another modulation scheme or channel coding, when the channel quality changes. Through the adaptation to the radio channel conditions it is possible to offer the highest data rates in good propagation conditions close to the site of the BSs, whereas under lower quality conditions a more

Table 4.1: EGPRS modulation and coding schemes and coding parameters

MCS	CR	HCR	Data per radio block [bit]	Fam.	RLC info [byte]	Data rate [kbit/s]
9 (8-PSK)	1.0	0.36	2×592	A	2×74	59.2
8 (8-PSK)	0.92	0.36	2×544	A	2×68	54.4
7 (8-PSK)	0.76	0.36	2×448	B	2×56	44.8
6 (8-PSK)	0.49	1/3	592 *(544+48)*	A	74	29.6 (27.2)
5 (8-PSK)	0.37	1/3	448	B	56	22.4
4 (GMSK)	1.0	0.53	352	C	44	17.6
3 (GMSK)	0.80	0.53	296 *(272+24)*	A	37	14.8 (13.6)
2 (GMSK)	0.66	0.53	224	B	28	11.2
1 (GMSK)	0.53	0.53	176	C	22	8.8

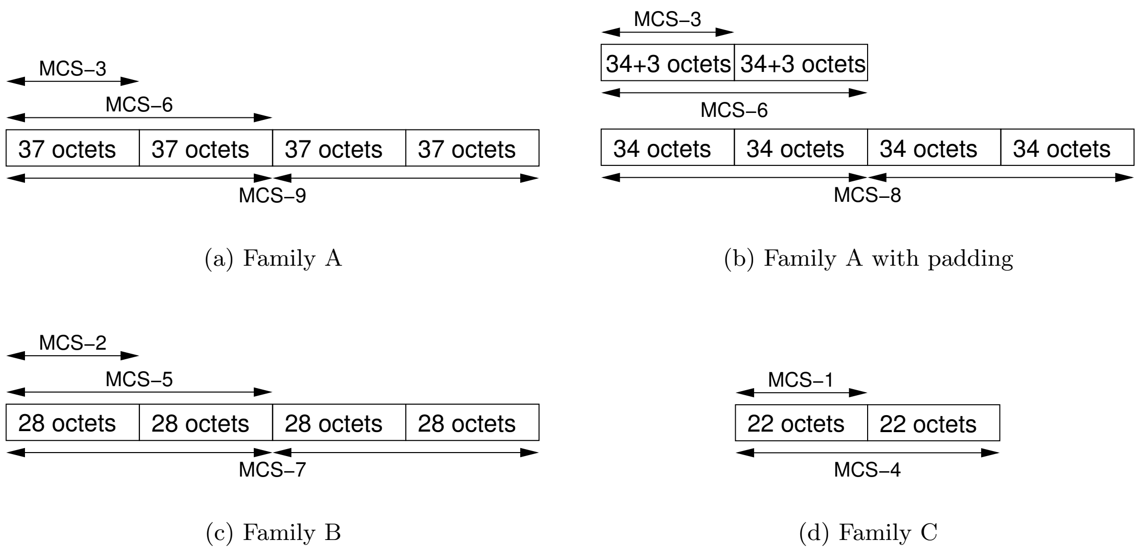


Figure 4.2: Basic block sizes and MCS families

robust coding scheme is selected.

4.2 Modulation and Coding Schemes

The additional use of 8-PSK modulation enables the introduction of new *Modulation and Coding Schemes* (MCSs). The possible MCSs are depicted in Table 4.1. MCS-1 to MCS-4 use GMSK as the modulation scheme and nearly equal the coding schemes used in GPRS, providing data rates up to 17.6 kbit/s. MCS-5 to MCS-9 use the previously described 8-PSK modulation scheme enhancing the data rate up to 59.2 kbit/s per time slot. The MCSs are classified in three different MCS families consisting of different basic payload sizes. Figure 4.2 gives an overview of how an RLC block is created depending on the MCS family. For instance, MCS-9 comprises four basic units of 37 octets each, whereas MCS-6 combines two basic units and MCS-3 only one. Through transmitting a different number of payload units within a 20 ms radio block duration, different code rates are achieved, resulting in bit rates of 8.8 kbit/s up to 59.2 kbit/s per time slot. According to the link quality, an initial MCS is selected for each RLC block. For retransmissions, it is only

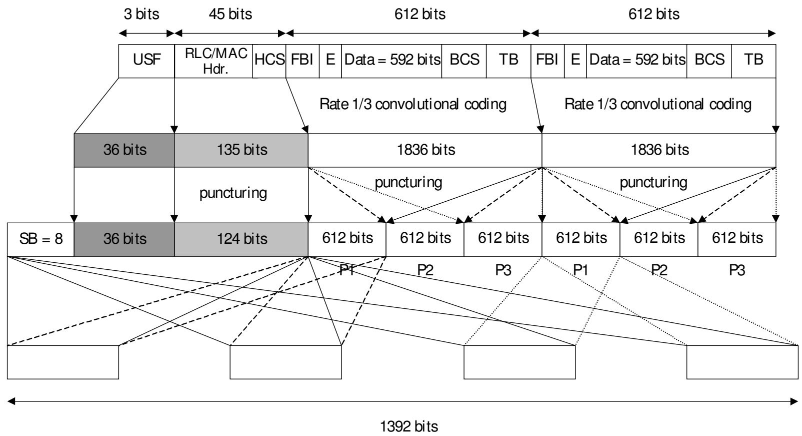


Figure 4.3: Coding and puncturing for MCS-9, uncoded 8-PSK, two RLC blocks

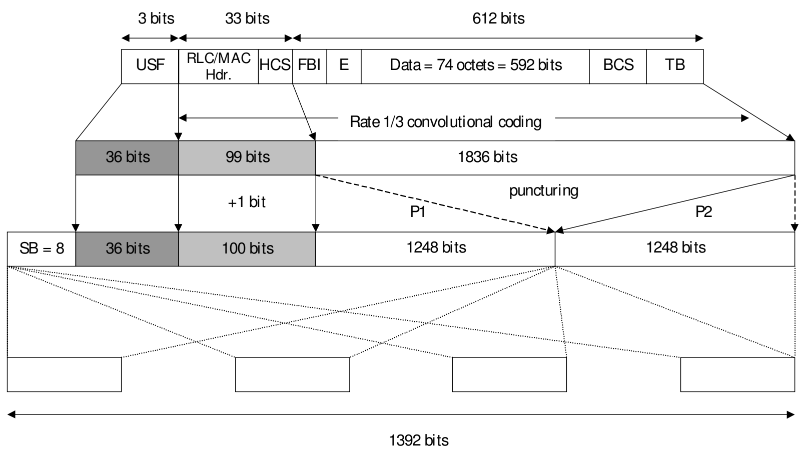


Figure 4.4: Coding and puncturing for MCS-6, rate 0.49 8-PSK, one RLC block

allowed to use the same MCS or to switch to another MCS of the same MCS family.

The following figures show the coding and puncturing used in Family A (MCS-9: Figure 4.3, MCS-6: Figure 4.4 and MCS-3: Figure 4.5). They differ in the amount of redundancy used for error correction, which determines accordingly the payload size.

Note that using MCS-7, MCS-8 and MCS-9, two RLC PDUs are transmitted within 20 ms. Both PDUs are interleaved over only two bursts, whereas their common header is interleaved over all four. Using a lower MCS, the RLC PDU (data and header) is interleaved over all four bursts. Figure 4.3 shows a MCS-9 RLC/MAC block consisting of two RLC PDUs of 592 bit (74 octets) payload and separate header containing the FBI field, extension field and the BSN (not pictured) to identify the RLC PDU within a TBF. The two RLC PDUs do not have to be in sequence; for example the first PDU can be a new RLC PDU and the second a retransmission of a previously sent RLC PDU. If there are no further RLC blocks to send, no retransmissions or pending retransmissions requested, the second payload is padded.

Initially, the RLC data block will be encoded with Puncturing Scheme PS-1, whereas PS-2 and PS-3 are used for retransmissions, then enabling the use of IR by using the different punctured information in PS-1, PS-2 and PS-3. The EDGE standard foresees an LA algorithm for the dynamic selection of MCSs to adapt to the quality of the radio link. This includes measurement and reporting the quality of the downlink and demands the

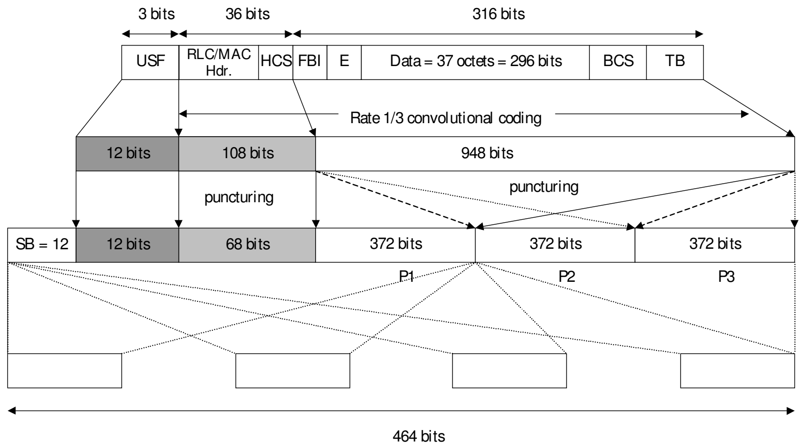


Figure 4.5: Coding and puncturing for MCS-3, rate 0.80 GMSK, one RLC block

indication of new modulation schemes (see Figure 4.6).

4.3 Link Quality Control

Link Quality Control (LQC) is a term used for techniques to adapt the channel coding of the radio link to the varying channel quality. Different modulation and coding schemes are optimal during different situations, depending on the link quality. The LQC used for EDGE is performed through the techniques of

- *Link Adaptation* (LA)
- *Incremental Redundancy* (IR)

LA provides a dynamic switching between coding and modulation schemes, so that the highest throughput (e.g., maximum user bit rate) according to the time-varying link quality (e.g. C/I) is achieved. With IR, information is first sent with very little coding. If decoding is successful, this will yield a very high user bit rate or throughput. However, if decoding is unsuccessful, then additional coded bits are sent until decoding is successful. The more coding bits that have to be sent, the less the resulting bit rate and the higher the delay. As a result of the link quality control, a low radio link quality will not cause a dropped data transfer, but only give a reduced bit rate for the user. Furthermore, a tight frequency plan (with a small cluster size, say 3) can be introduced while still providing high data rates for packet data services.

4.3.1 Link Adaptation

An LA scheme is proposed to maximize the user bit rate on the link through estimation of the time varying link quality in EGPRS and respectively adapt the most appropriate modulation and coding scheme. Through measurements of the link quality, the predictive algorithm estimates the performance of the currently used scheme and decides whether another MCS would perform better than the currently used one. The algorithm chooses the MCS with the best performance for the currently measured radio link quality and therefore for the expected link quality during the next bursts (it is assumed that the link quality can vary with time and location). During active TBF the most appropriate MCS is chosen on the basis of the channel quality measurements concerning the used MCS family. For compatibility reasons it is not allowed to switch the chosen MCS family during retransmissions, otherwise too much padding would decrease efficiency.

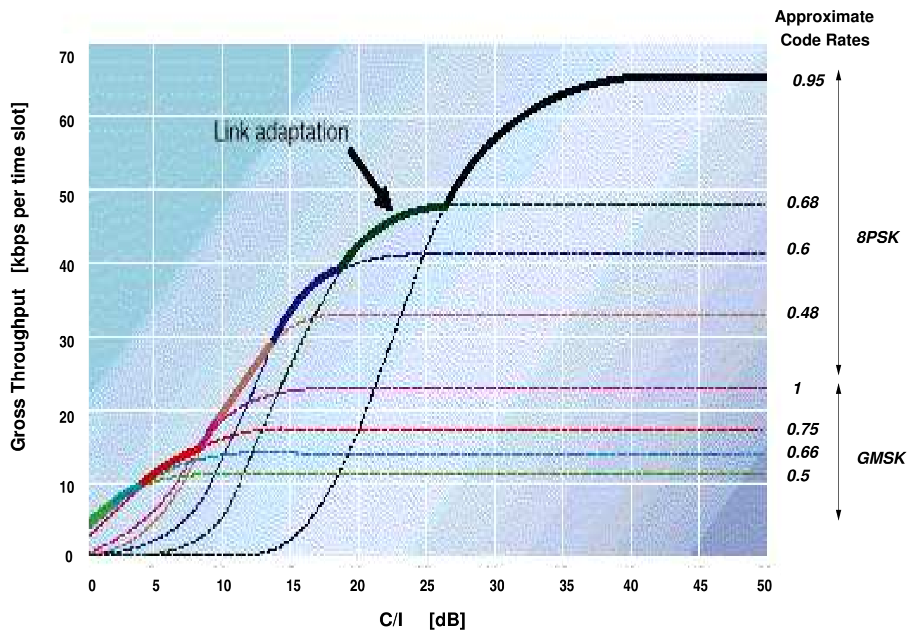


Figure 4.6: Link adaptation algorithm [Furuskär and Olofsson (1999)]

In order to perform the most efficient LA, measurements of the quality of the radio channel are the basis of the choice of the MCS. The operation of link quality control relies on measurements of the link quality, depending also on C/I values, frequency errors, time dispersion, interleaving gain due to frequency hopping, velocities and bit error probability.

4.3.2 Incremental Redundancy

Operating in RLC acknowledged mode, retransmissions are possible, when requested, or when capacity is available for pending retransmissions (RLC blocks not yet acknowledged). An ARQ scheme ensures an error-free data transmission by retransmitting erroneous blocks. When an error is detected in a block, all information about the block is usually discarded. The basic idea of IR is soft combining, i.e., not to forget what was already received in the erroneous transmission. Instead, the soft bit information from the erroneous blocks is saved and combined with the information received with the next (erroneous) retransmission. That way the *Block Error Ratio* (BLER) is reduced.

With IR mode PS-1 is used at first, and if decoding is unsuccessful, PS-2 will be used, etc. The IR functionality saves soft information of PS-1 and performs a joint decoding with PS-2 after the received retransmission with PS-2. A *Synchronization downlink Burst* (SB) is used to indicate which code rate and which puncturing scheme is used. Information about link quality control is given in [3GPP TSG GERAN (2002d)], [3GPP TSG GERAN (2002e)], and [3GPP TSG GERAN (2002f)]. Through measurements of the link quality every 60 ms (example inter-arrival time of the measurement report), the most suitable MCS family is chosen, based on the mean and variance of the C/I. For the following explanation, MCS family A is assumed. All assumptions apply also for the other families. For retransmissions three possible scenarios are drawn. The first one assumes that the link quality remains stable during the retransmissions. Then the RLC data blocks can be retransmitted with the originally used MCS. If the link quality worsens, an MCS downgrading becomes necessary, otherwise if the link quality improves, an MCS upgrading is possible. In the following, proposals for upgrading and downgrading actions are provided.

Upgrading When performing an MCS upgrade, retransmissions can be executed using the old MCS, allowing the use of IR by using different puncturing schemes (PS-1, PS-2, PS-3). That way, the BLER is reduced by using information from the different puncturing schemes. Otherwise no IR is possible.

Downgrading If the channel quality worsens, an MCS downgrading is necessary to provide a low BLER through higher protection of the bits. Thus, the retransmissions are carried out using a lower MCS, rejecting the possibility of the use of IR because of the incompatible puncturing of the different MCSs. For instance, when MCS-6 is chosen, when previous RLC blocks were sent with MCS-9, no coding problem occurs since MCS-9 comprises two individual RLC blocks consisting of two basic blocks each. MCS-6 also comprises two basic blocks, therefore only the coding, puncturing and interleaving will change and not the number of RLC blocks distinguished by separate BSNs. Retransmission of a previous MCS-9 RLC block results in a retransmission of two separate RLC blocks in MCS-6.

Assuming MCS-6 and downgrading to MCS-3, the payload data has to be deblocked into two subblocks (also called split blocks) and the header duplicated, creating two MCS-3 RLC blocks with the same BSN. Both MCS-3 blocks are retransmitted consecutively with the split bit set to 1 for the first and 2 for the second MCS-3 block to allow accurate assembly at the receiver side requiring the error-free arrival of both the first and second subblocks. Performing downgrading, all consecutive RLC blocks will be sent using the lower MCS, remaining in this MCS even if changed channel quality allows upgrading. This applies only for retransmissions, whereas the newly segmented LLC frame will be sent with the actual MCS, taking advantage of the better channel quality and the transmission capacity of the higher MCS.

It is not possible to acknowledge the subblocks separately, therefore the receiver has to wait until both subblocks have arrived before they enter the receiver window. The RLC block, consisting of two subblocks, is acknowledged positively if both subblocks have arrived error-free, otherwise both subblocks have to be requested by using the BSN of the former MCS-6 PDU.

4.4 Flow Control Modifications for EDGE

To support higher data rates in an efficient way the flow control function in the RLC/MAC protocol has to be modified. The *Window Size* (WS) is increased from 64 in GPRS to the range from 64 to 1024 in EGPRS. Consequentially the *Sequence Number Space* (SNS) is adapted from 128 to 2048 and the *Block Sequence Number* (BSN) is increased from 7 bit to 11 bit. To be able to maintain the selective acknowledgement function for this large SNS the *Received Block Bitmap* (RBB) is segmented into a *First Partial Bitmap* (FPB) and *Next Partial Bitmap* (NPB) controlled by a *Extended Polling Function* in EGPRS.

Traffic Models for Mobile Applications

In parallel with GSM evolution, the data applications performed by mobile users will evolve. In the first phase of the GSM evolution, where the data services *Circuit Switched Data* (CSD) and GPRS are available, applications based on the *Wireless Application Protocol* (WAP) [Wireless Application Protocol Forum (1999a)] running on smart phones and *Personal Digital Assistants* (PDAs) besides conventional Internet applications running on laptop computers or enhanced PDAs will dominate. Then Video and Audio Streaming applications [Elsen et al. (2001)] and *Large Data Transfer* (LDT) applications including the *Multimedia Messaging Service* (MMS) based on WAP version 2.0 [Wireless Application Protocol Forum (2001c)] are felt to become more popular with the optimization of GPRS and with the introduction of EDGE and the related packet data service *Enhanced General Packet Radio Service* (EGPRS). With the evolution of EDGE to GERAN and the convergence with UTRAN towards the IP-Multimedia Subsystem defined in 3GPP Release 5, *Voice over IP* (VoIP) technology can be introduced in the radio network and VoIP applications are expected to smoothly replace circuit-switched voice services.

This chapter gives a general survey of the different emerging mobile data applications and will present related traffic models. The traffic models are based on well-known models in the literature modified for the use for cellular packet radio networks traffic engineering. Frequently used applications like e-mail, WWW, WAP or FTP are outlined as well as emerging applications like Video and Audio Streaming and Voice over IP.

Traffic traces and packet size histograms are provided, gained by stochastic simulation of the traffic models over the TCP/IP stack and a narrowband channel model for the transmission of the generated IP packets defined by a fixed data rate, a certain Internet packet delay and a certain packet loss probability.

5.1 WWW Applications

The Internet represents the largest existing *Wide Area Network* (WAN) offering a broad spectrum of information resources in form of technical reports, product information, software, images, audio and video sources and electronic commerce services. The popularity among larger user groups to use these resources was driven by the introduction of the *World Wide Web* (WWW). In the WWW, global naming conventions, protocols and object formats are used. An object can be a text file, an audio source, an image file or even the code of a program that can be interpreted and executed on the user's terminal (Java applet). *Graphical User Interfaces* (GUIs), so-called *WWW browsers*, enable the user to navigate within the WWW and to transfer and display WWW objects. Besides the use of the WWW as the public Internet, the WWW concept is increasingly introduced in company networks. By the introduction of such *intranets* a higher transparency of company processes and the simple usage of company databases as well as the optimized management of customer relationship (CRM) is expected and realized.

A WWW browser displays WWW pages that are composed of one or more WWW objects. Figure 5.1 shows a WWW page displayed by a WWW browser. It contains several different objects such as text and images. Within one page *links* can be defined

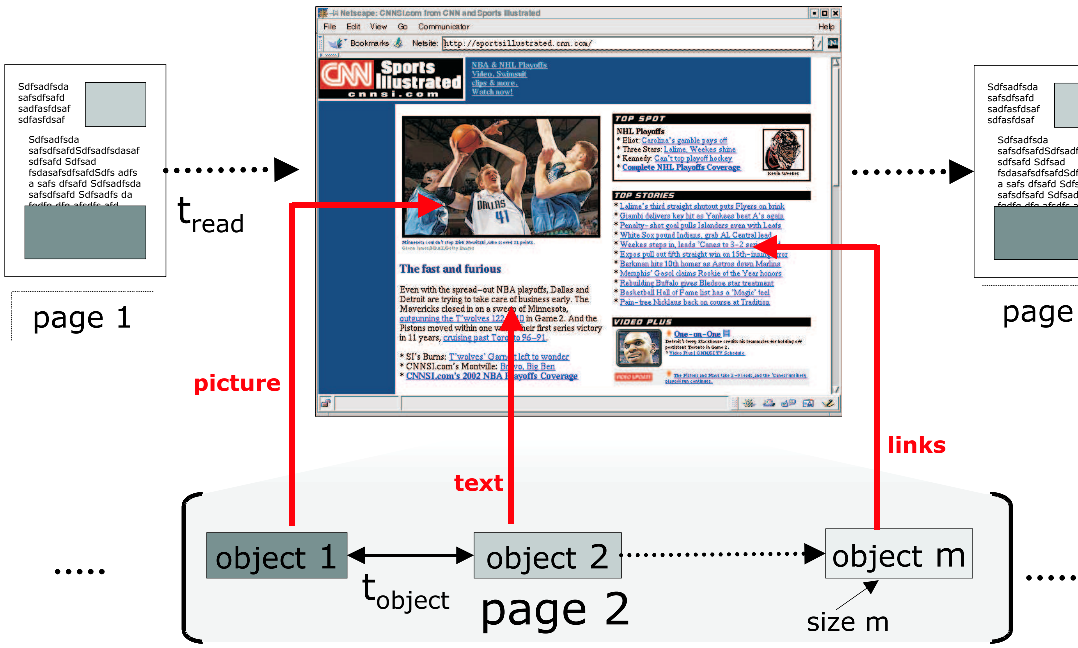


Figure 5.1: Structure of a WWW page

that are connected to an identifier of a WWW resource, a *Unified Resource Locator* (URL). When the user chooses a link, a new object or page is displayed. By the direct specification of a URL in the text field of the WWW browser, a specific page or object can be chosen by the user (see Figure 5.1). Every object of the WWW has a unique URL.

The presentation of a WWW page is described by the *Hypertext Markup Language* (HTML). In an HTML file the appearance and the URLs of the contained objects are specified. An HTML page can also contain text that is displayed at the related position of the page. An HTML page itself is also a WWW object that is analyzed by the browser. Further WWW objects are specified in the HTML file as parts of the page that are requested and transmitted.

The WWW is characterized by a client-server architecture. A WWW browser represents a client that poses requests to a WWW server concerning the transmission of WWW objects that are stored on the server. The communication between client and server is controlled by the *Hypertext Transfer Protocol* (HTTP) [Fielding et al. (1997)]. HTTP is an application-oriented protocol that uses the *Transport Control Protocol* (TCP) [Postel (1981)] for end-to-end transport. In the initial version of HTTP a new TCP connection is set up for each requested object. The client transmits a request for a certain object to the server. The request contains the URL of the object, data format parameters and parameters for access control. The server processes this request and transmits the requested object to the client. Then the TCP connection is released. In newer HTTP versions, e.g., HTTP version 1.1, TCP connections can be reused by the following objects (persistent TCP connections) and several TCP connections can be set up in parallel to be used for the pipelined transmission of several different objects (pipelined TCP connections). Figure 5.2 depicts the sequence of a WWW session.

5.1.1 Adapted Mosaic WWW Model

As mentioned above WWW sessions consist of requests for a number of pages. These pages consist of a number of objects with a certain object size. Another characteristic

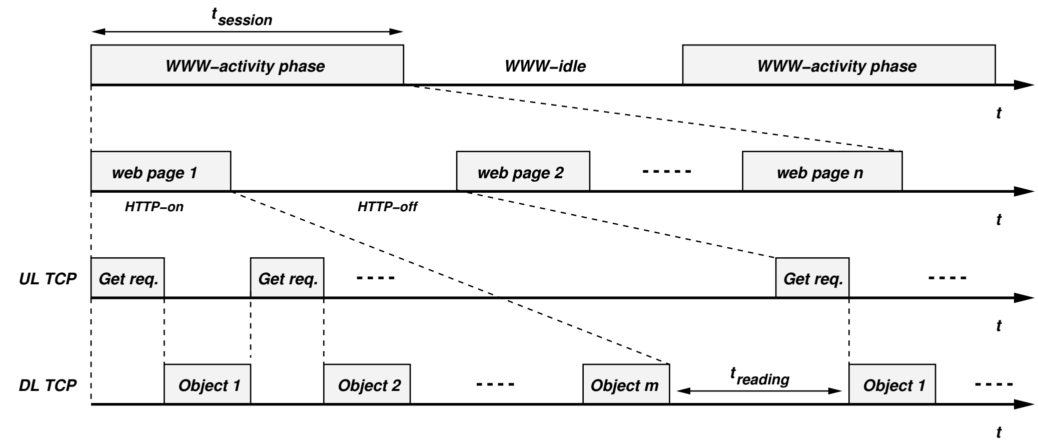
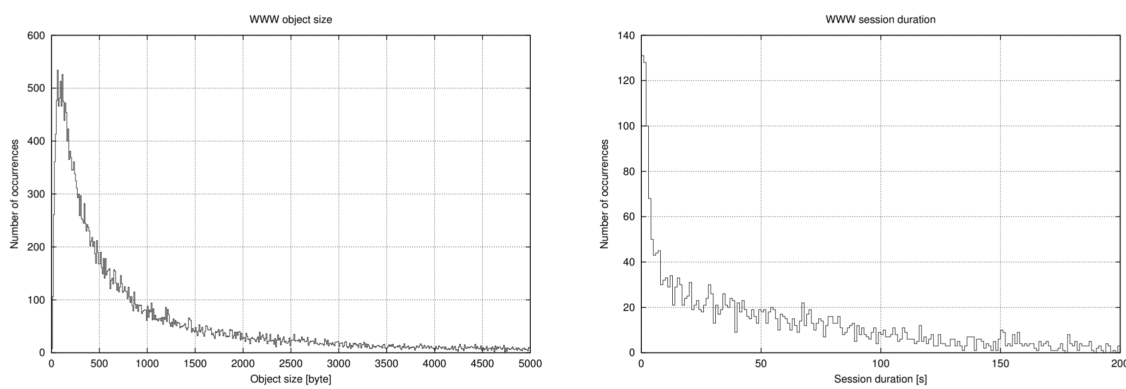


Figure 5.2: Sequence and parameters for a WWW session

Table 5.1: Adapted Mosaic WWW traffic model parameters

WWW Parameter	Distribution	Mean	Variance
Pages per session	geometric	5.0	20.0
Intervals between pages [s]	negative exponential	12.0	144.0
Objects per page	geometric	2.5	3.75
Object size [byte]	$\log_2$-Erlang-k ($k = 17$)	3700	$4.67 \cdot 10^9$
	transformed Erlang	9.4	5.2



(a) Object size

(b) Session duration

Figure 5.3: WWW traffic characteristics

parameter is the delay between two pages depending on the user's behavior to surf around the Web (see [Arlitt and Williamson (1995); ETSI 3GPP (1998)]). Table 5.1 gives an overview of the Mosaic WWW traffic parameters. The small number of objects per page (2.5 objects), and the small object size (3700 byte) were adapted based on [ETSI 3GPP (1998)], since Web pages with a large number of objects or large objects are not suitable for thin clients such as PDAs or smart phones served by (E)GPRS. The traffic characteristics of the WWW model can be seen in the distribution function of the object size.

In Figure 5.3 histograms for the object size and the session duration are shown. They were gained during stochastic simulation of the traffic models over the TCP/IP stack and a narrowband channel model for the transmission of the generated IP packets with a fixed data rate of 89.6 kbit/s (typical user data rate for EDGE/UMTS), an IP packet delay of 110 ms and without packet loss in the Internet.

5.1.2 Choi's Behavioral Model of Web Traffic

Choi's behavioral model is newer than the Mosaic model and therefore is representative for WWW browsing performed on *Personal Computers* (PCs) connected to the fixed Internet. Therefore the model is characterized by higher object sizes and larger WWW pages than in the Mosaic model. Since the performance of mobile networks in the first years of the GSM evolution is lower than in fixed access networks, this model is less suitable for GPRS traffic engineering. Once the performance of evolved GSM networks will be comparable to the performance of today's fixed access networks (e.g. with EGPRS or UMTS), the importance of this model will increase for cellular packet radio networks traffic engineering problems.

Choi's traffic model is an on/off-model with alternating phases of packet generation and silence (see Figure 5.4) [Choi and Limb (1999)]. An *ON* phase starts after the arrival and acceptance of a web request. During this phase, the objects of a WWW page are requested. The *OFF* phase represents a silence period after all objects have been retrieved. Thus, the ON and OFF phases equal the page loading times and page viewing times, respectively.

During the ON phase the page's objects are downloaded. We distinguish two types of objects: the main object containing the document's HTML code and inline objects, such as linked objects, images or Java applets. To fetch all those inline objects modern browsers open several TCP connections in parallel after the successful retrieval of the main object.

Table 5.2: Choi WWW traffic model parameters

WWW Parameter	Distribution	Mean	Variance
Viewing time [s]	Weibull	39.5	$8.57 \cdot 10^3$
No. of inline objects per page	Gamma	5.55	130.0
Size of main object [kbyte]	Log-Normal	10	625.0
Size of inline objects [kbyte]	Log-Normal	7.7	$1.59 \cdot 10^4$

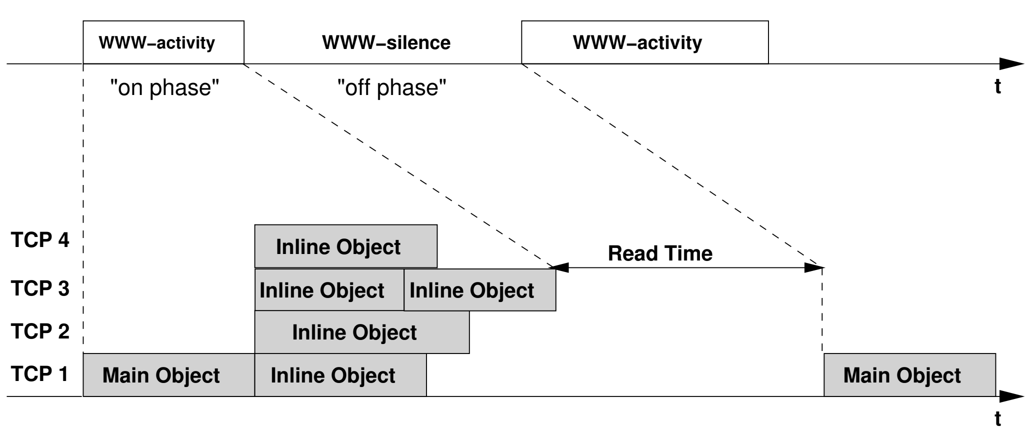


Figure 5.4: Choi's behavioral model

In Table 5.2 the random variables used by Choi to describe the object sizes, the number of inline objects and the length of the viewing time are listed.

The amount of data S that is generated during the ON phase is given by

$$S = s_{Main} + \sum_{i=0}^{n_{Inline}} s_{Inline,i} \tag{5.1}$$

which is a combination of three random variables. The random variable S represents the payload to be transmitted by the HTTP protocol.

5.2 E-mail Applications

E-mails are transmitted in the Internet by using the *Simple Mail Transfer Protocol* (SMTP) [Postel (1982)] or the protocols *Post Office Protocol version 3* (POP3) [Myers and Rose (1996)] and *Internet Message Access Protocol* (IMAP) [Crispin (1996)] for e-mail download.

5.2.1 SMTP, POP3 and IMAP

While local mail delivery systems, e.g., in *Local Area Networks* (LANs), are often based on the SMTP, it is often not possible to keep an e-mail client interconnected to the Internet for a long amount of time. Therefore the protocols POP3 and IMAP are widely used to download e-mail from a so-called mail-drop service (see Figure 5.5). These protocols are deployed to permit a terminal to retrieve mail that the server is holding for it, called the *store-and-forward* concept. TCP is used for e-mail transfer as the end-to-end transport protocol.

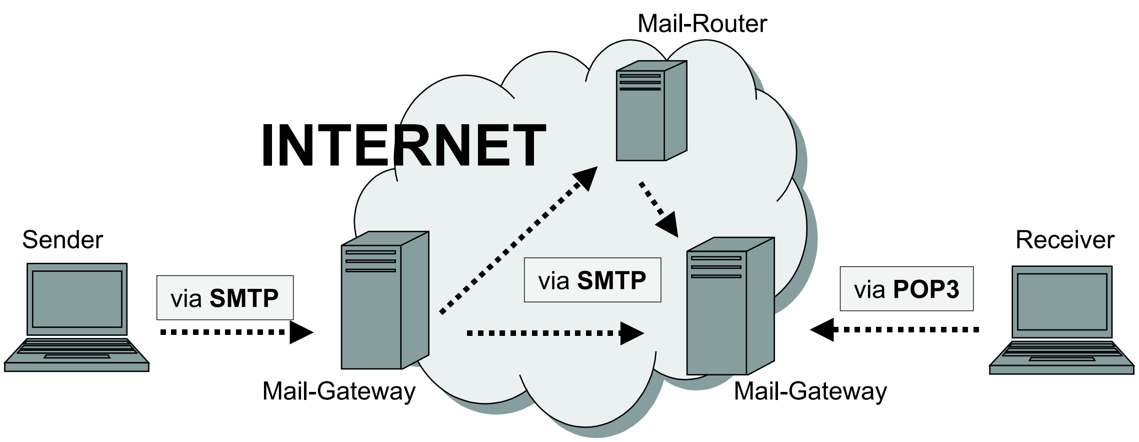


Figure 5.5: SMTP and POP3 for e-mail transfer

Table 5.3: E-mail traffic model parameters

E-mail Parameter	Distribution	Mean	Variance
E-mail size (lower 80%) [byte]	$\log_2$-normal	1700	$5.5 \cdot 10^6$
	transformed normal	10.0	2.13
E-mail size (upper 20%) [byte]	$\log_2$-normal	16000	$71.3 \cdot 10^9$
	transformed normal	9.5	12.8
Base quota [byte]	constant	300	0

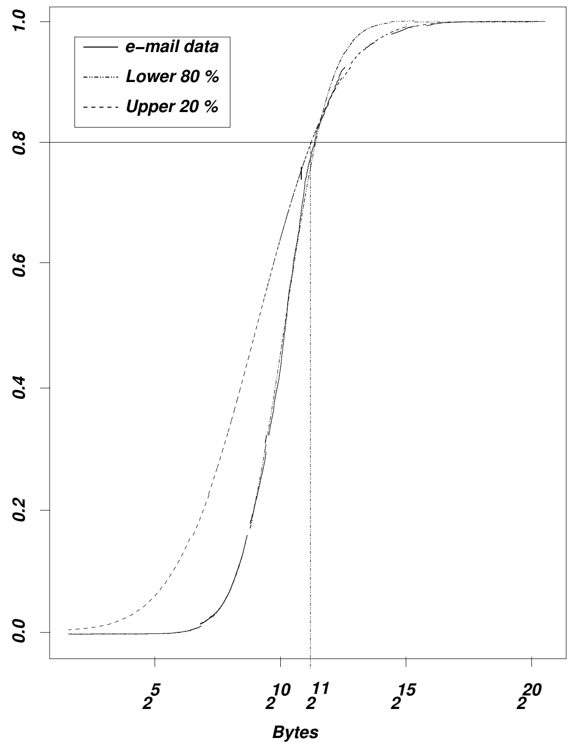
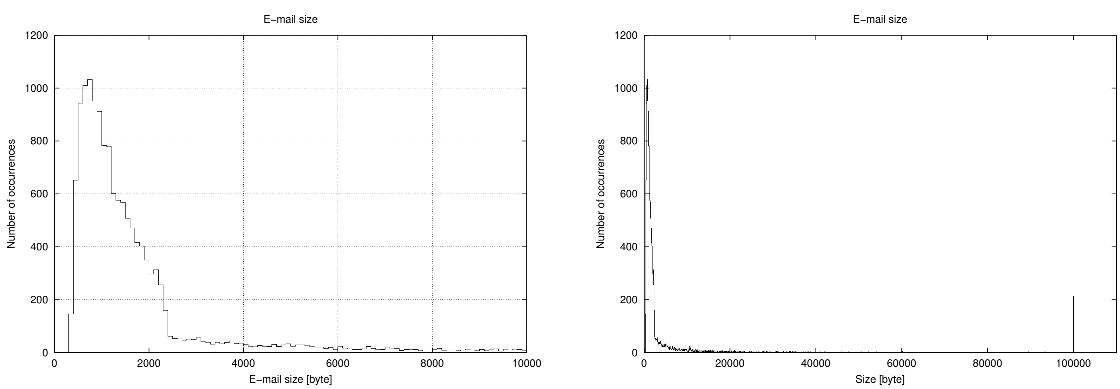


Figure 5.6: Bimodal $\log_2$ normal distribution function



(a) Small sizes

(b) Total range

Figure 5.7: E-mail size

5.2.2 E-mail Traffic Model

Since the size of an e-mail download on a mobile device is the crucial parameter for cellular packet radio networks traffic engineering, a traffic model defining e-mail sizes is suitable. The introduced e-mail model based on [PAXSON (1994)] describes the load arising with the transfer of messages performed by a user. The only parameter is the e-mail size that is characterized by two $\log_2$-normal distributions plus an additional fixed quota of 300 byte (see Table 5.3). The base quota was assumed to be a fixed overhead. Subtracting the overhead, a bimodal distribution remains. The lower 80% can be interpreted as text-based mails, while the upper 20% could represent mails with attached files, which can be rather large. The transition between these two distributions is 2 kbyte (see Figure 5.6). The maximum e-mail size is set to 100 kbyte.

In Figure 5.7 the histograms for the lower 80% of small e-mails and for the total range of the e-mail size are presented. Two different classes of e-mails can be observed. The size of the text-based e-mails begins with 300 byte, which is the default size needed for header information. The borderline between the distribution functions of these two e-mail classes is 2 kbyte according to [Paxson (1994)].

5.3 File Transfer Applications

The *File Transfer Protocol* (FTP) [Postel and Reynolds (1985)] was developed to enable the safe, secure and efficient file transfer between two computers. System-specific features like the structure of file systems of the computers involved in the transfer are overcome with this standard. The term FTP is commonly used both for the protocol and for the file transfer application program.

5.3.1 FTP Protocol

FTP is an application-oriented protocol that uses TCP as the transport layer. The systems communicating over FTP are building a client-server relationship. A system is called an FTP client when it poses requests to another system, the FTP server, concerning the transfer of files. The specified file is then transferred between client and server.

FTP uses two connections between client and server, which are called *FTP-Control* and *FTP-Data* (see Figure 5.8). An FTP control connection is used for communication control and command transfer managed by the *Protocol Interpreter* (PI). Examples of

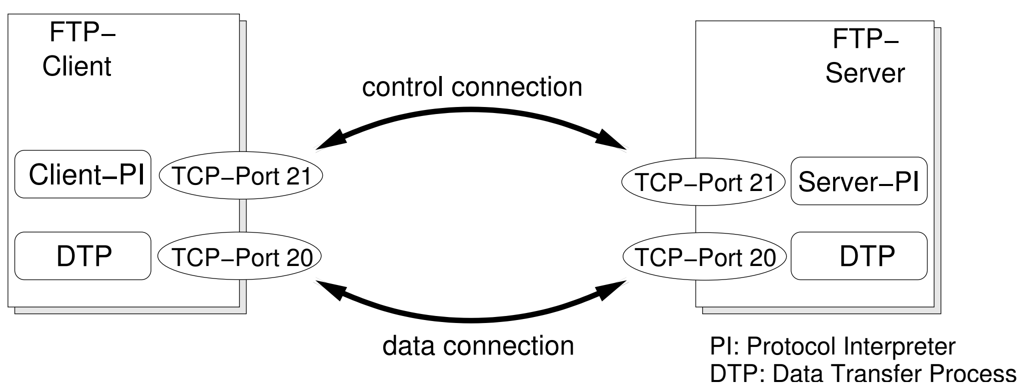


Figure 5.8: FTP data and control connection

Table 5.4: FTP traffic model parameters

FTP Parameter	Distribution	Mean	Variance
Total amount of data [byte]	$\log_2$-normal	32542	9661
	transformed normal	15.0	$1.9 \cdot 10^{-5}$
Amount of data per file [byte]	$\log_2$-normal	3000	1000
	transformed normal	11.55	$2.31 \cdot 10^{-4}$
Interval between connections [s]	$\log_{10}$-normal	4	2.54
	transformed normal	0.57	0.028

commands are `dir` for the listing of a directory or the commands `put` and `get` for the transmission of files between client and server. Only one control connection is used during an FTP session, while for each transmitted object one data connection is used managed by the *Data Transfer Process* (DTP). One object can be a file or, e.g., a directory listing.

The importance of FTP is decreasing with the wide usage of HTTP downloads and e-mail attachments. Particularly in wireless networks the usage of FTP is expected to diminish.

5.3.2 FTP Traffic Model

The FTP model introduced in this section represents a unidirectional data transfer. It describes a file transmission from an FTP server to an FTP client. The examinations are based on extensive WAN measurements documented in [Paxson (1994)]. Over 95% of the measured FTP data connections are performed by a `get` command. Since the downlink represents the bottleneck in cellular networks supporting Internet applications, which are asymmetric in nature, it is sufficient to regard only data transfer from server to client. Traffic originated by FTP control connections is not considered in the following model. Parameters describing an FTP session are:

- the total bulk of data per session
- the size of each transferred object
- the interval between two object transmissions

The bulk of data per session characterizes the duration of a session. A *log_2-normal distribution* is proposed for this total data per session and for the object size likewise. The given number of objects includes both the file transfer and, e.g., listing of directories. To describe the interval between two object transmissions a model created in [Paxson (1994)] can be used. Intervals between two objects should follow a *log_{10}-normal distribution*. All parameters, mean and variance of the distribution functions as well as the transformed mean and variance values are summarized in Table 5.4.

5.4 WAP-based Applications

The WAP specifications, which are implemented in today's mobile terminals, including WAP 1.2.1, aim at optimizing operation in 2G networks. Therefore WAP 1.2.1 defines a distinct technology comprising protocols and content representation. WAP is a suite of specifications that defines an architecture framework containing optimized protocols (e.g., WDP, WTP, WSP), a compact *Extensible Markup Language* (XML)-based content representation (*Wireless Markup Language* (WML), *WAP Binary Extensible Markup Language* (WBXML)) and other mobile-specific features like *Wireless Telephony Application* (WTA) [Wireless Application Protocol Forum (1999a,b)].

5.4.1 WAP Release 1.x

In addition to the goal of optimized operation in 2G networks, WAP had been developed because graphics-enhanced web services could not be brought to and displayed on thin clients, e.g., GSM mobile phones, and IP was not applicable in some environments, e.g., WAP over *Short Message Service* (SMS) or *Unstructured Supplementary Service Data* (USSD). Because of the optimization and the different protocols it was not possible to run WAP end-to-end to regular Internet sites. Instead, a WAP Gateway must be used. The main service a WAP Gateway provides is a protocol relay function between WAP stack

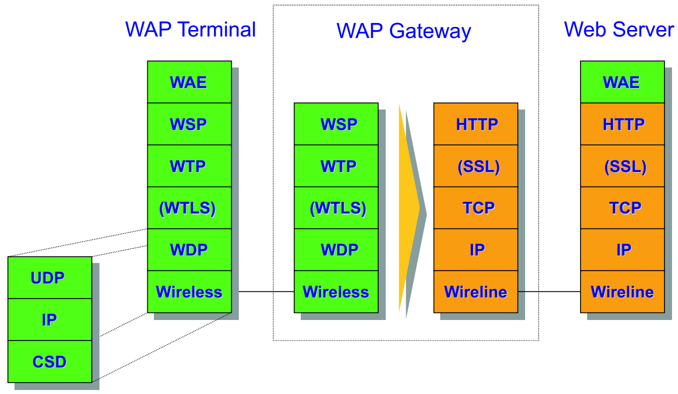


Figure 5.9: The WAP protocol architecture

and Internet stack. In addition to this standardized functionality, many gateway vendors provide a variety of value-added services that allow for personalization, for example.

The WAP protocol architecture (see Figure 5.9) comprises the following components:

Wireless Application Environment (WAE) With WAE an environment for development and execution of services is provided. The core component of WAE is the micro browser that renders the respective WML content.

Wireless Session Protocol (WSP) The session layer protocol WSP contributes methods to handle the session context. Thus communication between client and server is enabled.

Wireless Transaction Protocol (WTP) The transaction layer protocol WTP layer provides connectionless as well as connection-oriented communication. WTP class 2 represents the acknowledged client-to-server communication, which is commonly used for most of the WAP traffic.

Wireless Transport Layer Security (WTLS) WTLS is an optional function, which works in a similar way to *Secure Socket Layer* (SSL) or *Transport Layer Security* (TLS) in the Internet. WTLS provides authentication and privacy for ongoing communication.

Wireless Datagram Protocol (WDP) The transport layer WDP is an adaptation layer between WAP and the underlying bearer. In case of IP as the network layer, UDP is used instead of WDP.

Compared to the well-known Internet protocols like TCP and HTTP, WAP protocols differ in the following aspects:

- WDP or UDP provides an unacknowledged datagram transport only. This approach avoids overhead and delay for services that operate on a connectionless basis. One example is a WTP class 0 transaction.
- In-order delivery and retransmission are handled by a separate transport layer. It provides these functions only if they are requested by a higher-layer service (WTP class 1 or 2).
- WAP works in a transaction-oriented way. The amount of data that can be transferred by one transaction defaults to 1400 byte. It can be negotiated at WSP session setup time. If WTP segmentation and reassembly is used the amount of data per transaction is limited by the negotiated SDU size.
- WSP implements all features of HTTP 1.1 and enhances these by a binary encoding aiming at greater efficiency over the air.

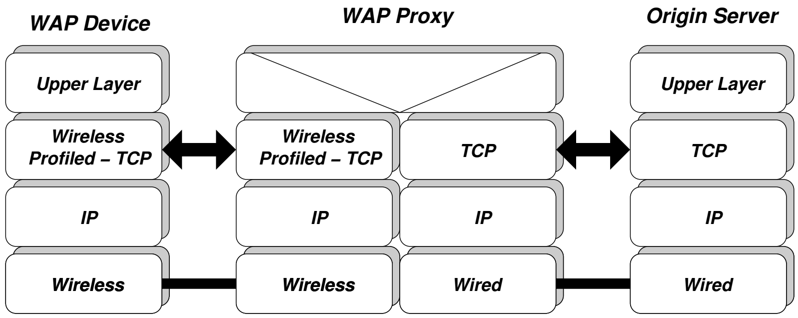


Figure 5.10: Wireless profiled TCP with WAP proxy

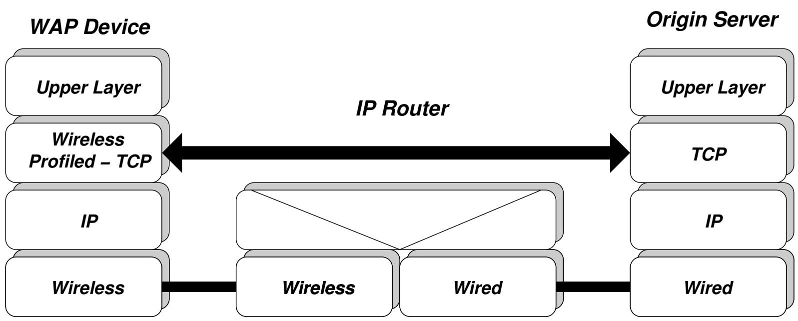


Figure 5.11: Wireless profiled TCP without WAP proxy

5.4.2 WAP Release 2.0

In the specification WAP 2.0 [Wireless Application Protocol Forum (2001b,c)] some existing WAP protocols have been extended by new capabilities. WAP 2.0 converges with widely used Internet protocols like TCP and HTTP. *Internet Engineering Task Force* (IETF) work in the *Performance Implications of Link Characteristics* (PILC) Working Group has been leveraged to develop a mobile profile of TCP for wireless links. This profile is fully interoperable with the common TCP that operates over the Internet today.

The inclusion of *Wireless Profiled TCP* (WP-TCP) into the WAP architecture of release 2.0 has been motivated by the introduction of wireless networks with higher data rates (e.g. GPRS, EGPRS and UMTS). The advantage of TCP is its ability to transfer large data, end-to-end security (using *Transport Layer Security* (TLS)) and convergence with IETF protocols. WP-TCP is optimized for wireless communication systems, but can interoperate with standard TCP implementations.

Further, WAP 2.0 does not require a WAP proxy, since the communication between the client and the server can be conducted using HTTP 1.1. However, deploying a WAP proxy can still optimize the communication process and may offer mobile service enhancements, such as location, privacy, and presence based services. In addition, a WAP proxy remains necessary to offer Push functionality.

Wireless Profiled TCP (WP-TCP) implementations can also be used for end-to-end connectivity without intermediate nodes. Thus WP-TCP must support both modes of operation, split- and end-to-end TCP. Which mode of TCP is chosen depends on factors such as the current provisioning, the application or the network access point. The requirements for this WP-TCP implementation are described in detail in [Wireless Application Protocol Forum (2001d)]. It is generally based on standard TCP with additional features concerning TCP optimization for wireless links. An example list of additional options is as follows:

- Large Window Size based on Bandwidth Delay Product
- Window Scale Option
- Timestamp Option
- Large Initial Window
- Selective Acknowledgement Option (SACK)
- Path *Maximum Transfer Unit* (MTU) larger than default IP MTU
- Explicit Congestion Notification

In addition to protocol work, the WAP Forum has continued its work on service-enabling features for the mobile environment, like the push service or synchronization issues.

Although WAP 2.0 has been finished in 2001, WAP 1.x protocol stacks will still be used in the mobile terminals in the next years. For the WAP traffic model presented in the following, only WAP 1.2.1 is regarded.

5.4.3 Multimedia Messaging Service (MMS)

The *Multimedia Messaging Service* (MMS) is intended to provide a rich set of content, e.g., images, photographs in combination with voice or text, to subscribers in a messaging context. It supports both sending and receiving of such messages by properly enabled client devices and is comparable to many messaging systems in use today (see Section 5.2). Examples include traditional e-mail available on the Internet and wireless messaging systems such as paging or SMS. These services provide a store-and-forward usage paradigm and it is expected that the MMS will be able to interoperate with such systems [Wireless Application Protocol Forum (2001a)].

MMS notifications inform the MS about a received message. The MMS notification message consists only of MMS headers. The purpose of the notification is to allow the client to directly fetch a multimedia message from the location indicated in the notification. A client retrieves messages by sending a WSP/HTTP get request (see Section 5.4.1) to the MMS server containing the URL of the received message. When successful, the response to the retrieve request contains headers and the body of the incoming message.

MMS presentation is done by means of ordering, layout, sequencing and timing of multimedia objects on the terminal screen and other devices such as a speaker. With MMS presentation, the sender of the multimedia message has the possibility of organizing the multimedia content in a meaningful order and instructing how the multimedia objects are rendered at the receiving terminal.

There are various alternatives for the presentation language, most notably WML (see Section 5.4.1) and *Synchronized Multimedia Integration Language* (SMIL) [W3C (2001)]. While the WML presentation for multimedia messaging offers the same sequencing and layout capabilities as for browsing applications, SMIL provides extended capabilities, such as timing of multimedia objects as well as animation. SMIL is a simple XML-based language that consists of a set of modules that define the semantics and syntax for certain areas of functionality. Examples of these modules are the layout module, the timing and synchronization module and the animation module.

The MMS presentation language information is transferred in the same message that is carrying the multimedia objects. Thus, a multimedia message is a compact package of multimedia objects and optional presentation information.

Since several components of an MMS correspond to different applications like e-mail or WWW browsing, existing traffic models could be applied for an MMS traffic model.

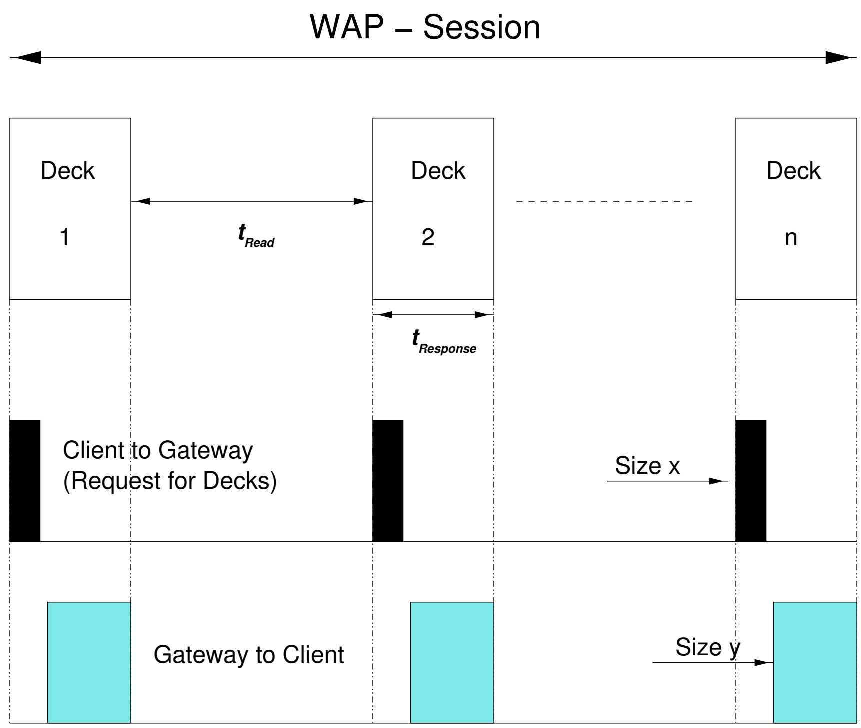


Figure 5.12: Sequence chart and important parameters for a WAP session

Table 5.5: WAP 1.x traffic model parameters

WAP Parameter	Distribution	Mean	Variance
Decks per session	geometric	20.0	3800
Intervals between decks [s]	negative exponential	14.1	198.8
Size of 'Get Request' packet [byte]	$\log_2$-normal	108.2	$4.1 \cdot 10^3$
	transformed normal	6.34	0.71
Size of 'Content' packet [byte]	$\log_2$-normal	511.0	$3.63 \cdot 10^5$
	transformed normal	8.6	1.55

Typical sizes of first multimedia messages are predicted to 20 kbyte and the maximum size because of memory restrictions is expected to be 100 kbyte.

5.4.4 WAP Traffic Model

A WAP 1.2.1 traffic model was developed and applied in [Stuckmann et al. (2001)]. The parameters were derived by measurements performed at a WAP gateway in test operation. The main characteristics of the model are very small packet sizes (511 byte) approximately following a *$\log_2$-normal distribution* and a limited value of the packet size (1400 byte).

A WAP session consists of several requests for a deck performed by the user. The sequence chart of a session is depicted in Figure 5.12.

WAP sessions are described by

- requests for a number of decks, n
- packet size in up- and downlink, *size x* and *size y*
- a reading time the user takes before requesting the next deck, t_{Read}
- a response time of the network, $t_{Response}$

The latter is not determined by the user. This parameter depends totally on the underlying network.

Subsequent to [Stuckmann et al. (2001)], the number of decks is modeled by a *geometric distribution*, the reading time by a *negative exponential distribution* and the packet size by a truncated *log_2-normal distribution*. While the model of the session can be seen as general for WAP sessions, the parameters depend on the content. The parameters taken here are typical for WAP applications directly after introduction of WAP. Since the applications will change over the next years, the parameters of the model are subject to future updates. Newer WAP deck sizes are in the order of 1 kbyte for monochrome decks and 3 kbyte for colored decks.

5.5 Streaming Applications

Many Internet portal sites are offering video services for accessing news and entertainment content from a Personal Computer (PC). Beside *Moving Pictures Expert Group* (MPEG), H.263 is the currently most accepted video coding standard for Video Streaming applications. In the area of Audio Streaming applications based on *MPEG Audio Layer-3* (MP3) are seen as potential applications. In the near future, mobile communication systems are expected to extend the scope of today's Internet Streaming solutions by introducing standardized Streaming applications as described in [Elsen et al. (2001)]. For the transport of Video frames over the Internet the *Real-time Transport Protocol* (RTP) [Schulzrinne et al. (2001)] is widely used.

5.5.1 Video Streaming

5.5.1.1 H.263 Video Coder for Low Bandwidth

To transfer real-time video data generally high rates are necessary. Looking at an uncompressed video stream, even high capacity wireline networks hardly fit the needs of such a stream. Table 5.6 gives an impression of needed bit rates.

Without compression no manageable video transmission is possible. More or less all video codecs work on the same principles. They compress the original pictures by the reduction of redundant information and the removal of irrelevant data.

Redundancy reduction: the loss-free reduction of redundant information depends on signal properties and statistics. Therefore algorithms like predictive coding, coding with variable word length and run-level-coding are used.

Irrelevance removal: this lossy removal exploits the characteristics of human perception (e.g. details in an image can be omitted without visible distortion) by using different quantization levels.

Although video transmission at 64 kbit/s was already introduced in 1979 reasonable picture quality at bit-rates below 64 kbit/s is still difficult. However, in 1996 the ITU-T Study Group XV specified a standard suitable for video transmission below 64 kbit/s. The

Table 5.6: Bit rates of common uncompressed video formats

Video format	Resolution	Bit rate
D1, digital studio standard	720×576 Pixel	166 Mbit/s
VHS quality (CIF)	352×288 Pixel	36.5 Mbit/s
Videophone (QCIF)	176×144 Pixel	3 Mbit/s

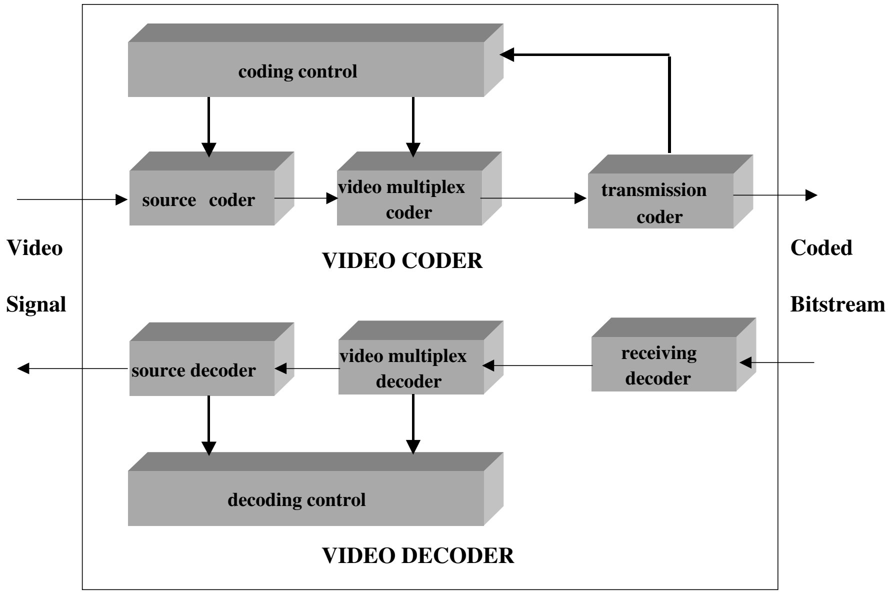


Figure 5.13: Outline block diagram of the H.263 video codec

international standard ITU-T Rec. H.263 [ITU-T (1997)] specifies a coded representation that can be used for compressing moving pictures for audio-visual services at low bit rates.

Another well-known and propagated video standard is the MPEG standard. H.263 offers great similarity to MPEG, since the experience with its use was included. It has MPEG-like blocks and macroblocks with prediction and motion compensation. The zigzagged quantized coefficients are coded using the MPEG run-level methods with different tables.

The basic configuration of the H.263 video source coding algorithm is based on ITU-T Rec. H.261 [ITU-T (1993)] and is a hybrid approach applying inter-picture prediction to utilize *temporal dependency* and transform coding of the remaining signal to remove *spatial redundancy*. Half pixel precision is used for the motion compensation, in contrast to Rec. H.261, where full pixel precision and a loop filter are used. Variable length coding is utilized for symbols to be transmitted.

In addition to the basic video source coding algorithm, negotiable coding options are included for improved performance. The most important ones are:

- *Unrestricted Motion Vector Mode* allows references which are outside the picture. For that purpose edge samples are duplicated to create the missing samples (H.263 annex D)
- *Syntax-based Arithmetic Coding* replaces the variable length coding for the same picture quality with fewer bits needed for coding (H.263 annex E)
- *Advanced Prediction Mode* sends a motion vector for each 8×8 luminance block for P frames (H.263 annex F)
- *Bidirectional Predicted Frames* enable the generation of B-frames in between two successive P-frames, using both P-frames for prediction. Thus the remaining difference which has to be coded is minimized (H.263 annex G)

A block diagram of the H.263 codec is shown in Figure 5.13. The video coder consists of the source coder, the video multiplex coder and the transmission coder.

- The *source coder* reduces the spatial and temporal redundancies of the input video signal by employing a DCT (Discrete Cosine Transform) algorithm

- The *video multiplex coder* defines the hierarchical structure of the coded video data suitable for transmission
- Finally, the *transmission coder* controls the output bitstream to comply with the channel bandwidth requirements. It merges the data from other services (e.g. audio and data) and produces the coded bitstream, which is directed to the transmission channel

5.5.1.2 Video Streaming Traffic Model

In the scope of modeling video sources, a lot of attention has been paid to long range dependent or self-similar models of traffic streams in telecommunication networks [Willinger et al. (1997)]. Many such models have been used to investigate *Variable Bit Rate* (VBR) video sources with a statistical analysis of empirical sequences and estimation of the grade of self-similarity [Rose and Frater (1994)]. Since MPEG and H.263 video traffic consists of a highly correlated sequence of images due to its encoding, the correct modeling of the correlation structure of the video streams is essential [Zaddach and Heidtmann (2001)].

For the proposed traffic model in this work no stochastic models of video streams with self-similar or highly correlated traffic characteristics are applied. Real video sequences coded by an H.263 coder are used to generate the Streaming traffic.

The Video Streaming traffic model used within the scope of this book is based on three video sequences in the *Quarter Common Intermediate Format* (QCIF) with the resolution of 176×144 pixels. The sequences are proposed by the *Video Quality Expert Group* (VQEG) and are for this reason commonly used for performance analysis. Each sequence represents a particular group of videos with different intensities of motion.

- `Claire` stands for videos with a very low motion intensity and can be seen as a characteristic video conferencing sequence or inactive visual telephony.
- `Carphone` includes both periods with rather high motion and periods of low motion intensity. It represents many kinds of vivid or active video-conferences or even visual telephony.
- The third video, `Foreman`, is a sequence with permanently high motion intensity of both the actor and the background. This permanent motion is characteristic for sport events or movie trailers.

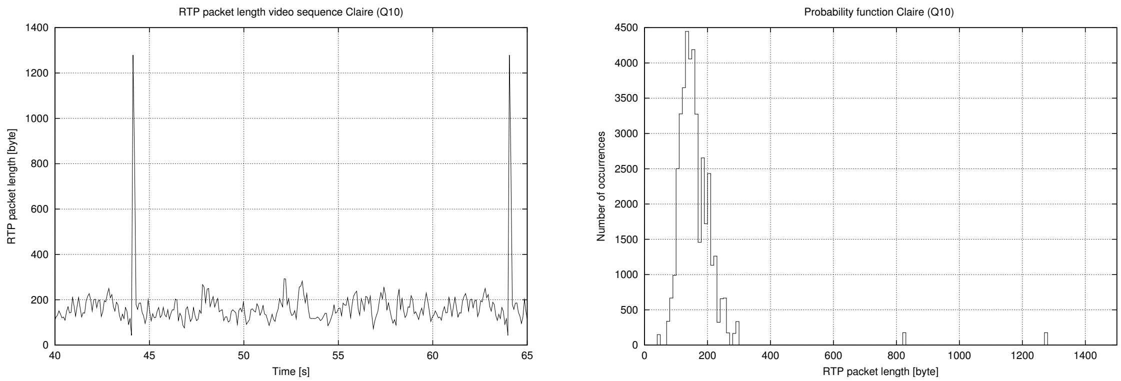
The H.263 coder was used with a skip factor of 2, which means that every second frame of the original sequence was skipped, so that the frame rate of the coded sequences was reduced from 25 to only 12.5 frames/s.

Q20 was adjusted for Intra (I)- and Predictive (P)-frames. The resulting video quality is low, but it is acceptable for mobile devices with their limited visual output capacities.

A conservative mix of sequences including 80% `Claire`, 10% `Carphone` and 10% `Foreman` has been selected for the proposed traffic model. The mix represents video streams with low motion and only a few streams with higher motion intensity.

Table 5.7: Offered IP traffic of video sequences

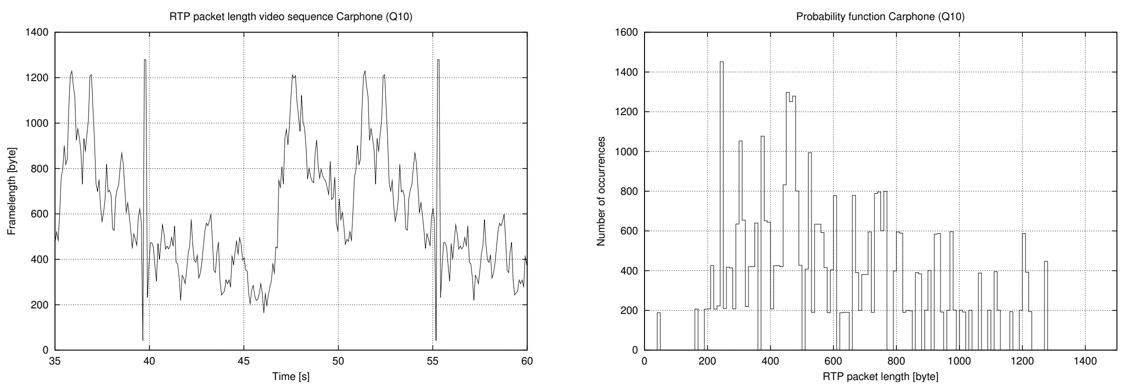
Sequences	Offered IP traffic	
	Q20	80-10-10 Mix
`Claire`	10.9 kbit/s	**14.39 kbit/s**
`Carphone`	26.7 kbit/s	
`Foreman`	31.7 kbit/s	



(a) Video trace

(b) PDF of RTP packet length

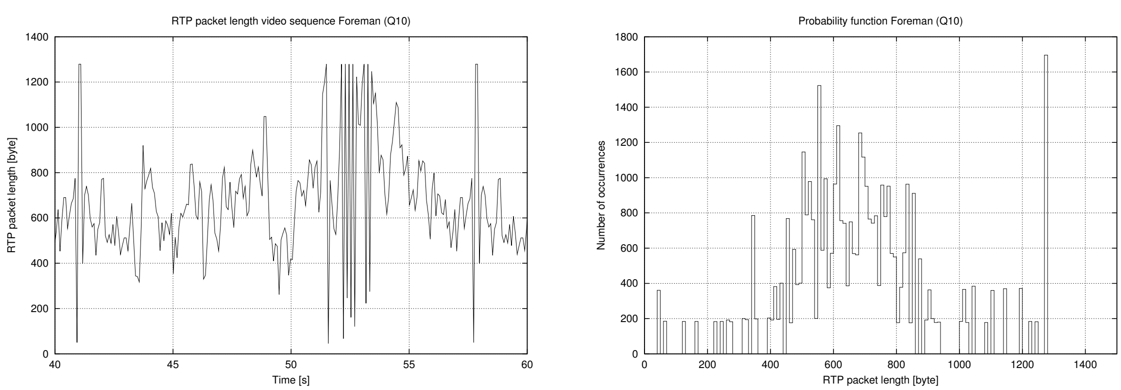
Figure 5.14: Video sequence `Claire` (*Quantization Level 10* (Q10))



(a) Video trace

(b) PDF of RTP packet length

Figure 5.15: Video sequence `Carphone` (Q10)



(a) Video trace

(b) PDF of RTP packet length

Figure 5.16: Video sequence `Foreman` (Q10)

Due to the negligible size of *Real-Time Streaming Protocol* (RTSP) and RTP control messages in comparison to the size of real-time data, they are proposed to be neglected. The resulting average IP traffic offered by this particular mix is 14.39 kbit/s (see Table 5.7).

All of the new emerging applications are predicted to be relatively short in duration. So-called *heavy users*, generating long streams with huge amounts of data, have not been taken into account. The duration of video sessions is modeled by a negative-exponential distribution with an average value of 120 s. This is an assumption with regards to the prognosis for 3G networks in [ETSI 3GPP (1998)] where the duration of real-time calls is proposed to be modeled by a negative-exponential distribution.

In Figure 5.14, Figure 5.15 and Figure 5.16 the RTP packet length trace and histogram of the three video sequences `Claire`, `Carphone` and `Foreman` are presented.

5.5.2 Audio Streaming

5.5.2.1 MPEG-1/MPEG-2 Audio Layer-3 (MP3)

For a digital audio signal from a CD, the bit rate is 1 411.2 kbit/s. With MP3 compression, CD-like sound quality is achieved at 128 kbit/s. Sound quality similar to FM radio for stereo signals can be achieved even at a bitstream of 64 kbit/s, and for a mono signal at only 32 kbit/s. For the use of low bit rate audio coding schemes in broadcast applications at bit rates of 60 kbit/s per audio channel, the International Telecommunication Union (ITU)-R recommends MPEG Layer-3.

Highest coding efficiency is achieved by algorithms exploiting signal redundancies and irrelevancies in frequency domain based on a model of human auditory systems.

All coders use the same basic structure (see Figure 5.17). The coding scheme can be described as perceptual noise shaping or perceptual subband/transform coding. The encoder analyzes the spectral components of the audio signal by calculating a filterbank (transform) and applies a psychoacoustics model to estimate the just noticeable noise-level. In its quantization and coding stage, the encoder tries to allocate the available number of data bits in a way to meet both the bit rate and masking requirements.

It consists of the following building blocks:

- Filter bank: A filter bank is used to decompose the input signal into subsampled spectral components (time/frequency domain). Together with the corresponding filter bank in the decoder it forms an analysis/synthesis system.
- Perceptual model: Using either the time domain input signal and/or the output of the analysis filter bank, an estimate of the actual (time and frequency dependent) masking threshold is computed using rules known from psychoacoustics.
- Quantization and coding: The spectral components are quantized and coded with the aim of keeping the noise, which is introduced by quantizing, below the masking threshold. Depending on the algorithm, this step is done in very different ways, from simple block companding to analysis-by-synthesis systems using additional noiseless compression.
- Encoding of bitstream: A bitstream formatter is used to assemble the bitstream, which typically consists of the quantized and coded spectral coefficients and some side information, e.g. bit allocation information.

5.5.2.2 Audio Streaming Traffic Model

Quite contrary to Video Streaming, the MP3 streams mainly consist of a *Constant Bit Rate* (CBR). In principle MPEG audio does not work just at a fixed compression ratio.

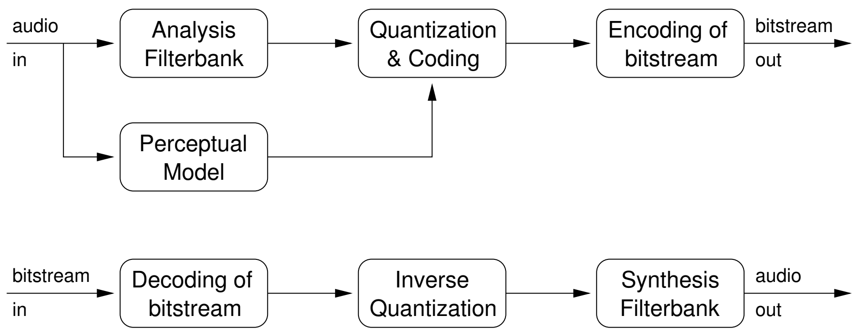


Figure 5.17: Block diagram of a perceptual encoding/decoding system

The selection of the bit rate of the compressed audio is, within some limits, completely left to the implementer or operator of an MPEG audio coder. The standard defines a range of bit rates from 32 kbit/s (in the case of MPEG-1) or 8 kbit/s (in the case of MPEG-2 Low Sampling Frequencies extension (LSF)) up to 320 kbit/s (MPEG-1) or 160 kbit/s (LSF). In the case of MPEG-1/2 Layer-3, the switching of bit rates from audio frame to audio frame has to be supported by decoders. This, together with the bit reservoir technology, enables both variable bit rate coding and coding at any constant bit rate in the range set by the standard. The Real-Audio (from RealNetworks Inc.) uses adaptive streaming for alternating bandwidth. From most operators the CBR is chosen. For example, QuickTime streams (from Apple Inc.) and MP3 streams (offered by many Radio Stations) are realized using the *Real-time Transport Protocol* (RTP).

Hence, the proposed traffic model for Audio Streaming over (E)GPRS is a CBR source with 10-64 kbit/s. For example the MPEG-2 *Advanced Audio Coding* (AAC) codec [ISO/IEC JTC1/SC29 (1997)] generates 256 byte every 64 ms for 32 kbit/s Stereo and generating 128 bytes every 64 ms for 16 kbit/s Mono.

5.6 Voice over IP

The transfer of voice traffic over packet networks, and especially VoIP, is rapidly gaining acceptance. While many corporations have been using voice over *Circuit-Switched* (CS) networks to save money, the dominance of IP has increasingly shifted the attention from CS to VoIP. VoIP transfer can significantly reduce the per-minute cost, resulting in reduced long-distance bills and a better utilization. In fact, many dial-around-calling services available today already rely on VoIP backbones to transfer voice, passing some of the cost savings to the customer.

One approach to providing a network architecture that allows new innovative services to be deployed quickly, efficiently and by a number of independent parties is to structure the network into three distinct functional areas: a Service Layer, a Call Layer and a Switching and Routing Layer. This is the vision that the *Session Initiation Protocol* (SIP) [Handley et al. (1999)] provides for *Next Generation Networks* (NGNs) independent of whether fixed or mobile operators are involved. In fact, in some cases, it could be seen that when a SIP enabled, converged services core network is in place with an operator, the difference between a fixed and a mobile network simply becomes how a user accesses the service and applications delivered from the core. VoIP services need to be able to connect to traditional CS voice networks. The ITU-T has addressed this goal by defining H.323, the set of standards for packet-based multimedia networks [ITU-T (2000)], and the IETF by defining the SIP architecture. For controlling gateways from external call control elements called media gateway controllers or call agents, the *Media Gateway Control Protocol* (MGCP) [Arango et al. (1999)] can be applied.

5.6.1 Audio Codecs

Voice channels occupy 64 kbit/s using *Pulse Code Modulation* (PCM) coding when carried over CS links. Compression techniques (Codecs) were developed allowing a reduction in the required bandwidth while preserving voice quality.

5.6.1.1 Speech Codec G.723

The G.723 [ITU-T (1996b)] is used for speech transmission in videoconferencing systems and is used in the video conference standard H.323.

The coder has two possible bit rates:

- 5.3 kbit/s
- 6.3 kbit/s

It encodes speech or other audio signals in frames using linear predictive analysis by synthesis coding. The excitation signal for the high rate coder is *Multi-Pulse Maximum Likelihood Quantization* (MP-MLQ) and for the low rate coder *Algebraic Codebook Excited Linear Predictive* (ACELP). The recommendation foresees a frame size of 30 ms in which the speech or other audio signals are encoded. Both rates are a mandatory part of the encoder and decoder. It is possible to switch between the two rates at any 30 ms frame boundary. An option for variable rate operation using discontinuous transmission and noise fill during non-speech intervals is also possible. In addition, there is a look ahead of 7.5 ms which results in a coding delay of 37.5 ms. All additional delays in the implementation and operation of this coder are due to:

- time spent processing the data in the encoder and decoder
- transmission time on the communication link
- additional buffering delay for the multiplexing protocol

The output of the speech coder during talkspurt and silence periods is depicted in Figure 5.18, generating speech packets of 24 byte at high rate and 20 byte when transmitting at low rate. Silence compression techniques reduce the transmitted data volume during silence intervals. Those techniques will be described in the following as *Voice Activity Detection* (VAD) and *Comfort Noise Generator* (CNG).

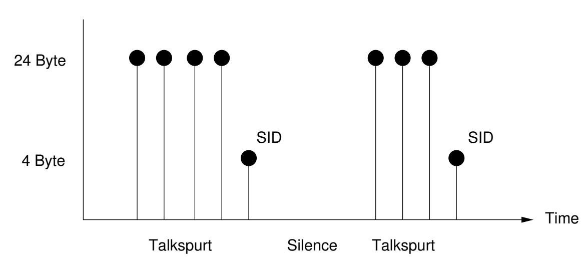


Figure 5.18: Output of speech coder G.723 (6.3 kbit/s)

5.6.1.2 Silence Compression Scheme

Silence compression techniques are used to reduce the transmitted bit rate during silent intervals of speech. Systems allowing discontinuous transmission are based on a VAD algorithm and a CNG algorithm that allow the insertion of an artificial noise during silence periods. This feature is necessary to avoid noise modulation introduced when the transmission is switched off. If the background acoustic noise that was present during active periods abruptly disappears, this very unpleasant noise modulation may even reduce the intelligibility of the speech.

5.6.1.2.1 Voice Activity Detection

The purpose of the VAD is to reliably detect the presence or absence of speech and to convey this information to the CNG algorithm. Typically, VAD algorithms base their decisions on several successive frames of information in order to make them more reliable and to avoid producing intermittent decisions. The VAD is constrained to operate on the same 30 ms speech frames which will subsequently be either encoded by the speech coder or filled with comfort noise by the comfort noise generator. The output of the VAD algorithm is passed to the CNG algorithm.

The largest difficulty in the detection of speech is the presence of any of a diverse range of background noise conditions. The VAD must be able to detect speech even in very low signal-to-noise ratio conditions. It is impossible to distinguish between speech and noise using simple level detection techniques when parts of the speech utterance are buried below the noise. The distinction between these conditions can only be made by taking into consideration the spectral characteristics of the input signal. In order to do this, the VAD incorporates an inverse filter, the coefficients of which are derived during noise-only periods by the CNG.

5.6.1.2.2 Comfort Noise Generator

The purpose of the CNG algorithm is to create a noise that matches the actual background noise with a global transmission cost as low as possible. At the transmitting end, the CNG algorithm uses the activity information given by the VAD for each frame and computes the encoded parameters needed to synthesize the artificial noise at the receiving side. These encoded parameters compose the *Silence Insertion Descriptor* (SID) frames, which require fewer bits than the active speech frames and are transmitted during inactive periods.

The main feature of this CNG algorithm is that the transmission of SID frames is not periodic. For each inactive frame, the algorithm makes the decision of sending an SID frame or not, based on a comparison between the current inactive frame and the preceding SID frame. In this way, the transmission of the SID frames is limited to the frames where the power spectrum of the noise has changed.

During inactive frames, the comfort noise is synthesized at the decoder by introducing a pseudo-white excitation into the short-term synthesis filter. The parameters used to characterize the comfort noise are the *Linear Predictive Coding* (LPC) synthesis filter coefficients and the energy of the excitation signal. At the encoder, for each SID frame the algorithm computes a set of LPC parameters and quantizes the corresponding LSPs using the coder *Linear Spectral Pair* (LSP) quantizer on 24 bit. It also evaluates the excitation energy and quantizes it with 6 bit. This yields encoded SID frames of 4 byte including the 2 bit for bit rate and *Discontinuous Transmission* (DTX) information.

A notable feature of this CNG algorithm is the method used to evaluate the spectrum of the ambient noise for each SID frame. It takes into account the local stationarity or non-stationarity of the input signal.

5.6.2 Performance Issues

5.6.2.1 Delay

Delay causes two problems - echo and talker overlap. Echo is caused by the signal reflections of the speaker's voice from far-end telephone equipment back into the speaker's ear. Echo becomes a significant problem when the round-trip delay becomes greater than 50 ms. Since echo is perceived as a significant quality problem, VoIP systems must address the need for echo control and implement some means of echo cancellation.

Talker overlap (or the problem of one talker stepping on the other talker's speech) becomes significant if the one-way delay becomes greater than 250 ms. The end-to-end delay budget is therefore the major constraint and driving requirement for reducing delay through a packet network.

Accumulation Delay: This delay (sometimes called algorithmic delay) is caused by the need to collect a frame of voice samples to be processed by the voice coder. It is related to the type of voice coder used and varies from a single sample time (0.125 ms) to many milliseconds, e.g., 30 ms for the G.723 codec (see Section 5.6.1.1).

Processing Delay: This delay is caused by the actual process of encoding and collecting the encoded samples into a packet for transmission over the packet network. The encoding delay is a function of both the processor execution time and the type of algorithm used. Often, multiple voice coder frames will be collected in a single packet to reduce the packet overhead.

Network Delay: This delay is caused by the physical medium and protocols used to transmit the voice data, and by the buffers used to remove packet jitter on the receive side. Network delay is a function of the capacity of the links in the network and the processing that occurs as the packets transit the network. The jitter buffers add delay which is used to remove the packet delay variation that each packet is subjected to as it transits the packet network. This delay can be a significant part of the overall delay since packet delay variations can be as high as 70-100 ms in IP networks.

5.6.2.2 Jitter

The delay problem is compounded by the need to remove jitter, a variable inter-packet timing caused by the network a packet traverses. Removing jitter requires collecting packets and holding them long enough to allow the slowest packets to arrive in time to be played in the correct sequence. This causes additional delay.

The two conflicting goals of minimizing delay and removing jitter have engendered various schemes to adapt the jitter buffer size to match the time varying requirements of network jitter removal. This adaptation has the explicit goal of minimizing the size and delay of the jitter buffer, while at the same time preventing buffer underflow caused by jitter.

5.6.2.3 Header Compression

One of the problems with IP over wireless is the large overhead introduced by IP and other above protocol headers such as UDP and RTP (see Figure 5.19).

These headers are used to transmit each packet to the correct host with the proper application in the correct order and at the right time. In case of real-time IP services, the

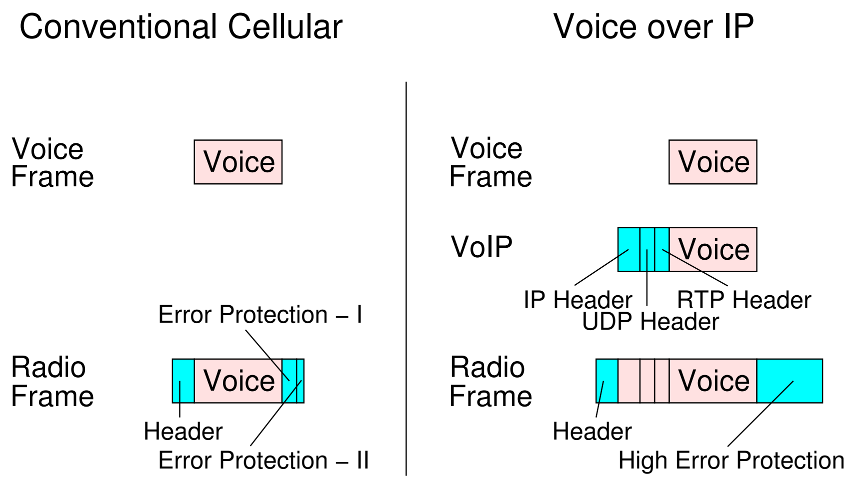


Figure 5.19: Voice frames

RTP/UDP/IP protocol stack is used to carry the media frames. When transporting packet voice frames, the length of the RTP/UDP/IP header (40 byte for IPv4 and 60 byte for IPv6) is larger than the payload. Namely, GSM codecs use 20 ms frames. This corresponds to 32.5 bytes every 20 ms in the case of the GSM Full Rate codec. The overhead problem is addressed by header compression to enable VoIP to be transmitted in a spectrum efficient way. Header compression algorithms must be efficient and robust against errors to be used on the air interface. The proposed *Compressed RTP* (CRTP) algorithm, which is used for wireline VoIP, is not robust enough for wireless links. More relevant algorithms have been proposed in the *Robust Header Compression* (ROHC) specification [Bormann et al. (2001)].

5.6.3 Voice Traffic Characteristics

Voice activities can be considered as alternating between two states: talkspurt and silent. Data are generated during talkspurt only, and no data is transmitted during silence, thereby making a statistical multiplexing gain possible. In [Brady (1969)] it was presented that both talkspurt and silence periods of digitized voice are exponentially distributed.

5.6.3.1 Six-state Model

Figure 5.20 shows the model for two speakers developed in [Brady (1969)]. Allowable state transitions are indicated. There is no attempt to control *B's* behavior; he starts and stops talking in his own manner. His state changes cause horizontal transitions.

Vertical transitions are determined by A. If he is talking, he stops when an *all silent pulse* occurs, and if silent he starts when a *start talking pulse* occurs. These are called β- and α-pulses, respectively. These pulses are a result of Poisson processes, so that, for example, if A is talking and B is silent (A solitary talk state), he stops talking in the next dt seconds with probability $\beta_{sol}^{A} \cdot dt$. For notation, the subscript on β, the fall silent parameter, describes the present state, while the subscript on α, the start talking parameter, denotes the event that will occur if the pulse occurs. The superscript refers to A or B. The six values for β and α are denoted $\beta_{sol}, \beta_{ted}, \beta_{tor}, \alpha_{pse}, \alpha_{alt}$, and α_{int} (see Figure 5.20) in which the abbreviations mean *sol*itary, interrup*ted*, interrup*tor*, *p*au*se*, *alt*ernate, and *int*errupt.

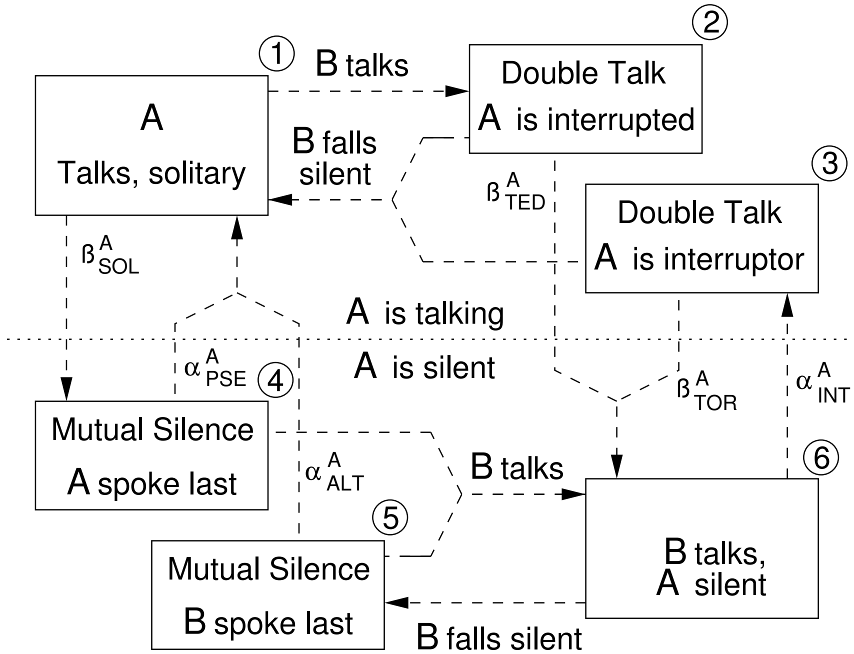


Figure 5.20: Six-state model

5.6.3.2 Single-Speaker Model

The commonly accepted model for a speaker in a voice call is a continuous-time, discrete-state Markov chain. The holding time in each state is assumed to be exponentially distributed with mean $1/\lambda$ and $1/\mu$, respectively (see Figure 5.21).

The commonly used values are $1/\lambda = 0.65\,s$ and $1/\mu = 0.352\,s$. These values are often referred to as the *Traditional Packet Voice Model*. $1/\lambda = 0.7428\,s$ and $1/\mu = 0.4357\,s$ were also measured from actual telephone traffic with silence detector of high sensitivity [Yatsuzuka (1982)]. The fundamental basis of all packet voice performance studies for the past is the classic model of exponential distributions by Brady. This model, however, was based on digitized voice statistics. Its applicability to packet voice with different audio content has not been examined before.

Packetization may change the traffic characteristics of packet voice, since both the techniques and implementations are different. In addition, while digitized voice in the traditional model was mainly concerned with telephone conversations, packet voice will be used for a wider range of new applications that may affect the traffic model, too. For instance, a significant amount of applications will be for information or entertainment distribution. The effect of audio contents other than conversations is not well understood, yet. It is, therefore, necessary to study the traffic characteristics of packet voice, and to examine the suitability of Brady's model to today's packet voice applications.

5.6.3.3 Audio Contents

In [Deng (1995)] spontaneous conversations between members or visitors of GTE Laboratories were evaluated. In addition to diversifying application environments and configurations, the topics of conversations were varied. Compared to a social conversation, technical discussion may have more pauses, i.e., longer silences, caused by thinking or attending to other tasks such as retrieving a file, finding a reference and so on. It is also necessary

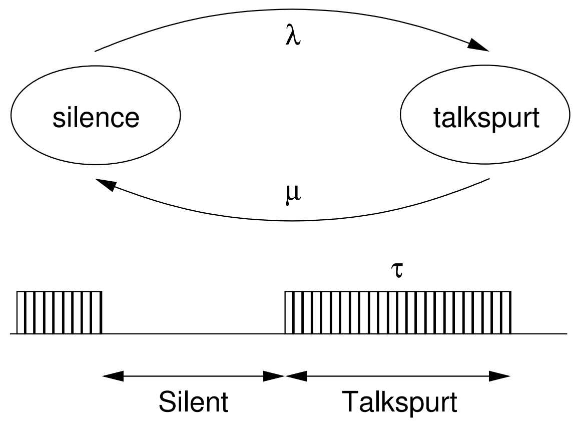


Figure 5.21: Single-Speaker model

to include the type of audio contents. The traditional traffic model was developed solely for telephone conversations. Packet voice, however, is serving a much wider range of applications. Prominent among them is information distribution. Examples may include delivery of news or entertainment programming, remote lecturing, retrieval of voice mail or e-mail in synthesized voice. They are referred to as *scripted speech*. Scripted speech may conceivably have different traffic characteristics from conversational applications.

5.6.4 Packet Voice Traffic Model

The proposed traffic model for VoIP is based on the Single-Speaker Model as described in Section 5.6.3.2 (see Figure 5.21).

In [Deng (1995)] the suitability of the six-state model for voice traffic was tested to verify the hypotheses: "Talkspurt period and silence period of a packet voice session are exponentially distributed". Deng's statistical analysis of the actual packet voice traffic indicates that only the average talkspurt and silence periods are much longer (see Table 5.8) than the values traditionally used, but the traditional model of exponential distribution is also statistically rejected in some cases.

Table 5.8: Voice traffic model parameters

	Single-Speaker model for digitized speech [Brady (1969)]	Traditional Packet Voice [Deng (1995)]	Packet Voice [Deng (1995)]
talkspurt	1.34 s	0.352 s	7.24 s
silence	1.67 s	0.65 s	5.69 s

During talkspurt periods, the G.723 codec generates a CBR of 24 byte every 30 ms (based on the high data rate of the G.723, Section 5.6.1.1), or 20 byte at the low rate. Directly after the transition to the silent state a four-byte frame for the CNG is transmitted.

CHAPTER 6

Simulation Tool for Performance Analysis

The (E)GPRS Simulator GPRSim is a software solution based on the programming language C++. For implementation of the simulation model the *Communication Networks Class Library* (CNCL) [Junius (1993)] has been used that is a predecessor to the *SDL Performance Evaluation Tool Class Library* (SPEETCL) [Steppler (1998)]. This allows an object-oriented structure of programs and is especially applicable for event-driven simulations. The complex protocols like SNDCP, LLC, RLC/MAC, the Internet traffic load generators and TCP/IP are specified formally with the *Specification and Description Language* (SDL) and are translated to C++ by means of the Code Generator SDL2CNCL [Olzem (1995)] and are finally integrated into the simulator.

Unlike the usual approaches to building a simulator, where abstractions of functions and protocols are being implemented, the approach of the GPRSim is based on the detailed implementation of the standardized protocols. This enables a realistic study of the behavior of EGPRS and GPRS. In fact, the real protocol stacks of (E)GPRS are used during system simulation and statistically analyzed under a well-defined traffic load.

The event control is performed by event handlers that are activated by arriving events and that send events to other event handlers after processing. The scheduling is done by a scheduler, which determines the order of processing of the events. Each event has a priority and a defined processing time. The simulation time advances in discrete steps, when all events, the processing time of which coincides with the current simulation time, are processed. Through this contemporaneity in the simulation process the simultaneous reaction of, e.g., several mobile stations to an event is represented realistically.

6.1 Structure of the (E)GPRS Simulator

The logical structure of the GPRSim and the information flow between the modules is shown in Figure 6.1. The simulator comprises the modules MS, BS and SGSN, the transmission links, the load generator, session control modules, a graphical user interface for presentation and a module for statistical evaluation. Multiple instances of MS and BS can be generated and studied in a multicellular environment.

The formal structure of MS, BS and SGSN is similar. They contain the implementations of the respective protocol stacks. The layers BSSGP and *Frame Relay* (FR) are not represented in the GPRSim because the focus was set on the radio interface. The respective classes do not provide any functionality and simply forward the service data units to the peer entity.

The transmission links are represented by error models. While the G_b interface is regarded as ideal, block errors on the radio interface U_m can be simulated based on look-up tables, which map the actual *Carrier-to-Interference* (C/I) to a *Block Error Probability* (BLEP) considering mobility.

The load generator comprises generators for both circuit- and packet-switched traffic and includes transport and network protocols implementations of the TCP/IP protocol stack. The module *Channel Management* supervises the physical GSM channels available in the respective cell and allocates channels for the GPRS resource management entity,

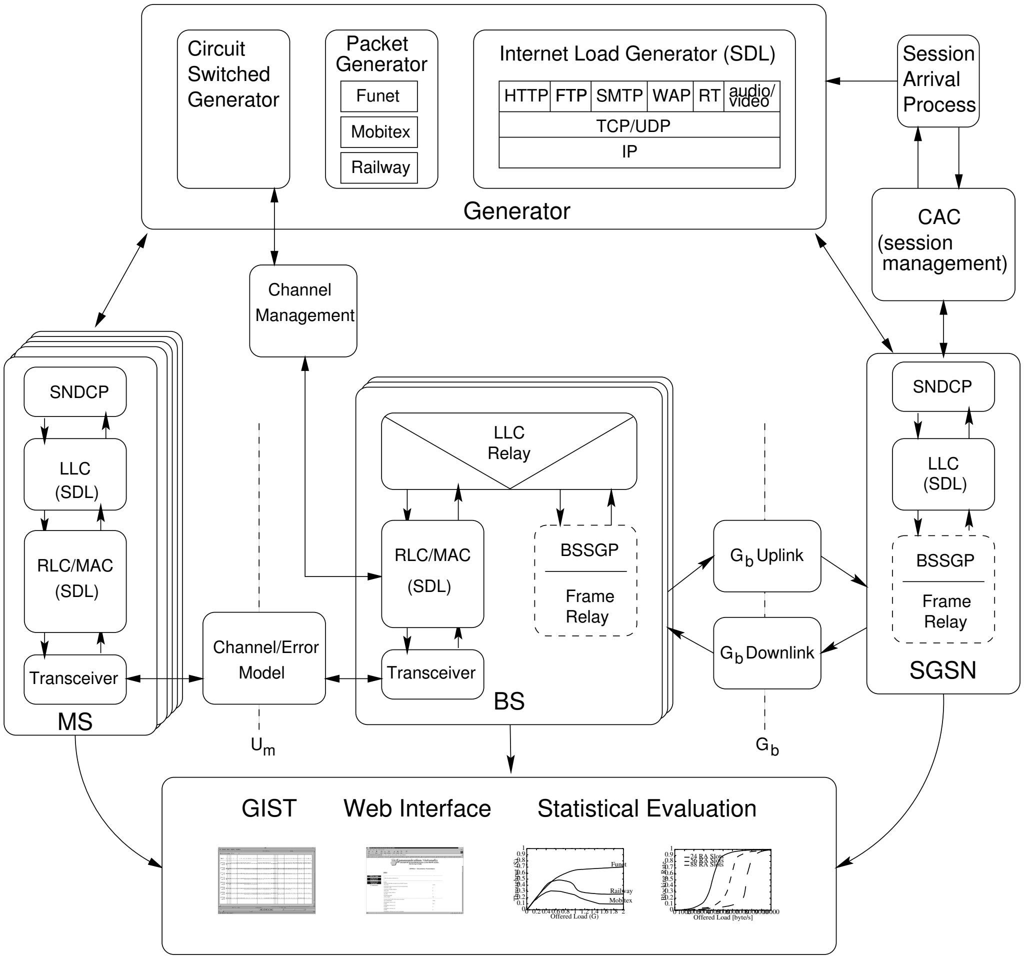


Figure 6.1: Structure of the (E)GPRS simulator

if the channels are not used by circuit-switched services or if they are not allocated as dedicated PDCHs for (E)GPRS.

The output of the simulator comprises a graphical presentation of the protocol cycles and the statistical evaluation results of the performance measures.

In the following sections the different modules are presented with their functionality and interactions.

6.2 Packet Traffic Generators

The GPRSIM comprises an Internet load generator based on the Internet traffic models for the applications WWW, FTP, e-mail, WAP, and Streaming (see Chapter 5). The transport and network protocols TCP, UDP, and IP are also implemented. The TCP implementation is comprising both slow start and congestion avoidance consistent with the specification *TCP Reno*.

Besides the Internet load generator a packet generator based on the models *Finnish University and Research Network* (FUNET) (electronic mail), *Mobitex* (fleet management) and *Railway* (railway application) as proposed by ETSI [BRASCHE and WALKE (1997)] are implemented. These traffic models were proposed during the GPRS standardization process and have lost their importance for GPRS traffic modeling.

The conception of the GPRSIM *Internet Load Generator* is the specification of a

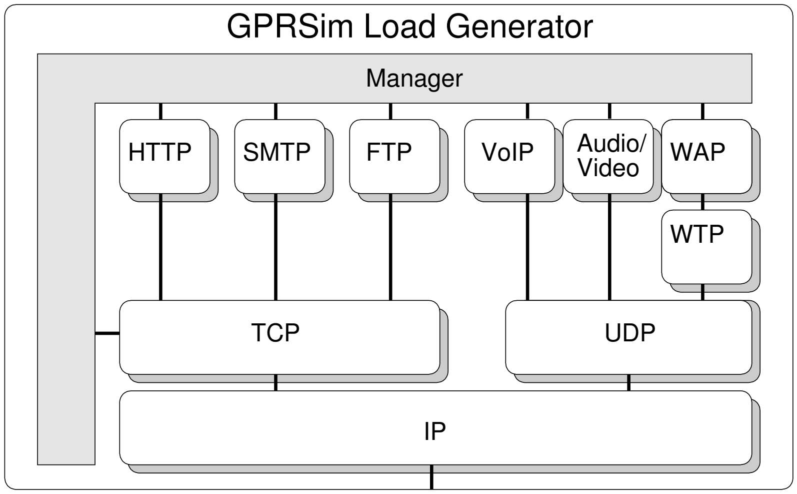


Figure 6.2: The GPRSim load generator

system *Client* that behaves like a user running an Internet session, and a system *Server* representing the peer entity for the Internet service. For every MS one client and one server instance of the load generator are created. Below the application model the protocols TCP, UDP, WTP and IP are included. IP packets are generated and transmitted over the GPRS protocol stack.

6.2.1 Internal Structure

The internal structure of the GPRSim Load Generator (see Figure 6.2) follows the ISO-/OSI reference model. The *Application Layer* comprises the most important applications we know today like surfing in the *World Wide Web* (WWW) via HTTP, sending and receiving e-mails over SMTP/POP3 or downloading large files via FTP. Additional to these TCP-based applications there are UDP-based services.

The connection-oriented (TCP) and the connectionless transport protocols (UDP, WTP) belong to the *Transport Layer*.

Finally every TCP segment is encapsulated by the *Internet Protocol* (IP) belonging to the *Network Layer*.

Starting a new session the Manager chooses randomly between the applications based on configurable session probabilities. Then the client instance sends a request to the corresponding server to initiate a data transfer and the server confirms the request. The size and number of requests and answers depend on the chosen application type.

Furthermore the Manager is transmitting session control signals, e.g. `session_start` or `session_error`, from the different layers to the underlying `SDL_Process_environment`. Conversely the environment sends signals from the session control module, e.g., `init` or `session_end_request` to the corresponding entities.

6.2.2 Implementation

6.2.2.1 System Client

The system `sClient` represents the client for Internet traffic. Six different blocks are present inside the system (see Figure 6.3). While five of them determine the overall system behavior, the block `bManager` manages administrative purposes.

Inside the block `bApplication` one process for each traffic model exists (see Chapter 5).

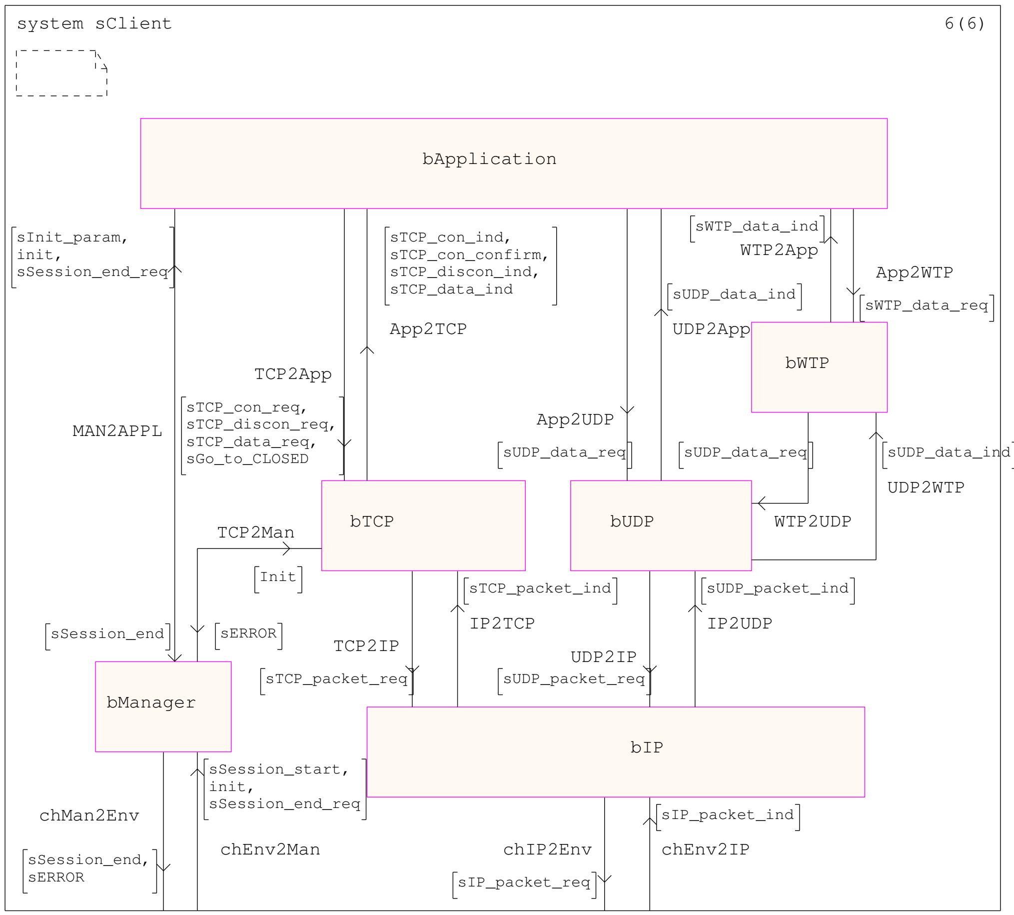


Figure 6.3: The system `sClient` in SDL

- At `sSession_start` the block `bManager` chooses randomly between the applications. The default probabilities of the different applications can be set via the configuration file.
- Receiving `sInit_param`, the block `bApplication` gets information about the protocol to be used. The specified SDL-process starts to initiate the application-related actions and requests objects from the Internet server instance according to the traffic model. Four parameters are generated by the `sClient`:
  - the number of Web pages or files to request,
  - the delay between each page/file,
  - the size of transmitted e-mails,
  - the duration of a video stream or VoIP session.
- The blocks `bTCP` and `bUDP` respectively receive the generated segments and encapsulate them according to the protocol with the corresponding header. In case of TCP different phases like *slow start* or *congestion avoidance* are performed.
- After the block `bIP` has fragmented packets bigger than the Ethernet *Maximum Transfer Unit* (MTU) (1500 byte) it adds the IP header to datagrams and sends them to the environment.

Conversely, packets received from the environment via `sIP_packet_ind` are forwarded to the block `bIP` and are passed through the protocol stack in the opposite direction.

6.2.2.2 System Server

The system `sServer` models the server of the Internet session. It is also composed of six different blocks and is similar in structure to the system `sClient`.

- Only the block `bApplication` differs from the client side, corresponding to the different functionality of load generation. The application-related processes have to generate random values for the size of requested objects, files, messages or e-mails.
- The process `pReal_Time_Server` inside `bApplication`, which is responsible for streaming video, is generating a whole sequence of RTP packets. The basis for these data is recorded traces of video sequences with different levels of motion, namely `Claire`, `Carphone` and `Foreman`.
- The different application protocols are distinguished by the protocol control received within the header as in the `sClient`.

6.3 Traffic Generator for Circuit-switched Services

The *Circuit-Switched* (CS) traffic generator generates events with an interarrival time determined by a negative-exponential distribution. These events correspond to calls initiated in a cell. When one event arrives, directly the next event is generated to realize the arrival process behavior. A channel assign request is sent to the channel management module. If there is a traffic channel available, the request is confirmed and the call duration, also negative-exponentially distributed, is determined. After the call duration an event is sent to the channel management module to release the assigned channel. The traffic value in Erlang is given by the two configurable mean values of the call interarrival and the call duration times.

6.4 Channel Management

The Channel Management module in the GPRSim applies dynamic channel allocation to control the pool of GSM traffic channels (TCHs) for GPRS and GSM applications as described in [3GPP TSG GERAN (2002e)]. A TCH can be used for either a circuit-switched speech connection or a PDCH. Channel Management communicates with the CS

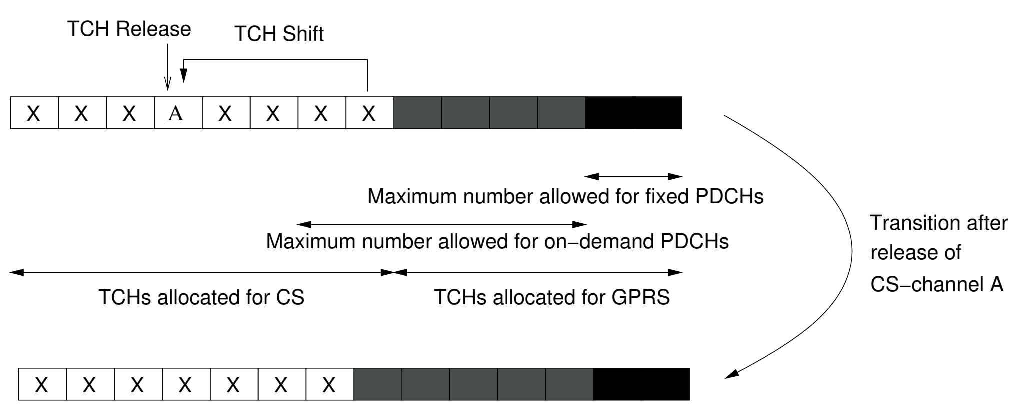


Figure 6.4: TCH and PDCH allocation table

traffic generator and the GSM/GPRS Radio Resource Management (RRM) instance that is located in the RLC/MAC layer. The state of this module is defined by the number of TCHs available to a radio cell, the maximum allowed number of fixed PDCHs, the maximum allowed number of on-demand PDCHs and the actual number of CS connections. Based on these parameters Channel Management determines the number of PDCHs currently available for GPRS. CS connections are prioritized with preemption, i.e., a new CS request interrupts and uses a PDCH used so far by GPRS immediately, if no other TCH is free in the cell. All TCHs that are not used by CS connections are available for GPRS, if the number of PDCHs does not exceed the sum of the maximum allowed number of fixed and on-demand PDCHs. It is assumed in the simulations presented that TCHs used for CS connections are placed adjacently beginning with the TDMA channels of carriers not used for GPRS. If a CS connection (say A) is released, the TCHs are shifted in a way so that TCHs and PDCHs are adjacent again after release (see Figure 6.4). In a real system this goal could be reached by intracell handoff of only one TCH. This mechanism is known as repacking. A transition, i.e., when the number of PDCHs available for GPRS changes, is immediately indicated to the RRM of RLC/MAC.

6.5 Quality of Service Management

In this section, a description of the concepts developed to model the *Quality of Service* (QoS) management functions presented in Section 3.5.2 and Section 3.5.3 is provided, as well as a brief description of the relevant implementation [Stuckmann and Müller (2001)].

6.5.1 QoS Profile and Aggregate BSS QoS Profile

A data structure is needed for the data defining the various aspects of QoS within the GPRS system regarded.

The class `QoSProfile` contains all parameters necessary to specify a mobile user's subscribed QoS, a mapping function of the application type onto an appropriate traffic class (see Section 3.5.3.3), and the traffic classes' definition in terms of the following service classes (see Section 3.5.3.1):

- precedence class
- reliability class
- delay class
- mean throughput class
- peak throughput class

The service classes correspond to the relevant values representing the values listed in Section 3.5.3.1.

`QoSProfile` offers an interface to initialize these values according to the simulation scenario requirements, to modify them during PDP context activation or modification procedures, and to set appropriate service class values for application traffic flows.

However, the service class values are not utilized for scheduling purposes—this is done solely on behalf of ABQPs, as described in Section 3.4.2. The ABQP, thus, is considered a single parameter with multiple data transfer attributes that is exchanged between SGSN and BSS during PFC context-related procedures. For implementation purposes, it is not necessary to combine subscription and traffic class into one parameter. They are instead encapsulated in a C++ `struct`, as shown in the following source code.

```
typedef struct aggregateBSSQoSProfileType : public CNObject
{
  int                tlli ;
  subscriptionType subscription;
  trafficClassType   trafficClass ;
};
```

For QoS differentiation both application classes (Conversational, Streaming, Interactive and Background) and subscriber classes (Premium, Standard and Best-Effort) are considered.

6.5.2 PDP Context

Since communication paths between the GPRSim modules involved in PDP-context related signaling are modeled only logically (see Section 6.5.4), there are few fields of the PDP context *Information Element* (IE) necessary to be modeled within the scope of the simulator. Hence, the class `PDPContext` simply contains the QoS Sub, QoS Req, QoS Neg fields, and the PDP address (see Section 3.5.2.3.2).

```
class PDPContext : public CNObject
{
  public:     //***** Constructors *****/
    PDPContext();
    PDPContext(int);
    PDPContext(const PDPContext&);
    PDPContext(CNParam *) {}

  public:     //***** Public interface *****/
    int           PDPAddress;

    QoSProfile* QoSSub;
    QoSProfile* QoSReq;
    QoSProfile* QoSNeg;

    PDPContext& operator=(const PDPContext&);

  private:    //***** Internal private members *****/

// {{{ Member functions required by CNCL...
};
```

6.5.3 GPRS Register (GR)

The class `GR`, representing the *GPRS Register* (GR) within the *Home Location Register* (HLR), is responsible for the storage of subscriber related QoS profile data. It provides an array of pointers to objects of type `QoSProfile`. The array size is equivalent to the number of MSs for the respective simulation scenario.

The tasks of `GR` comprise the initialization of the array of QoS profiles according to the relation of *premium* to *standard* to *best-effort* subscribers specified for the respective scenario, and the delivery of subscriber QoS profile data on request from an SGSN handling a PDP context activation. Moreover, the class `GR` has to provide an interface to the generator to allow traffic load generation in conformance with the subscriber QoS profiles valid for the simulation scenario under examination.

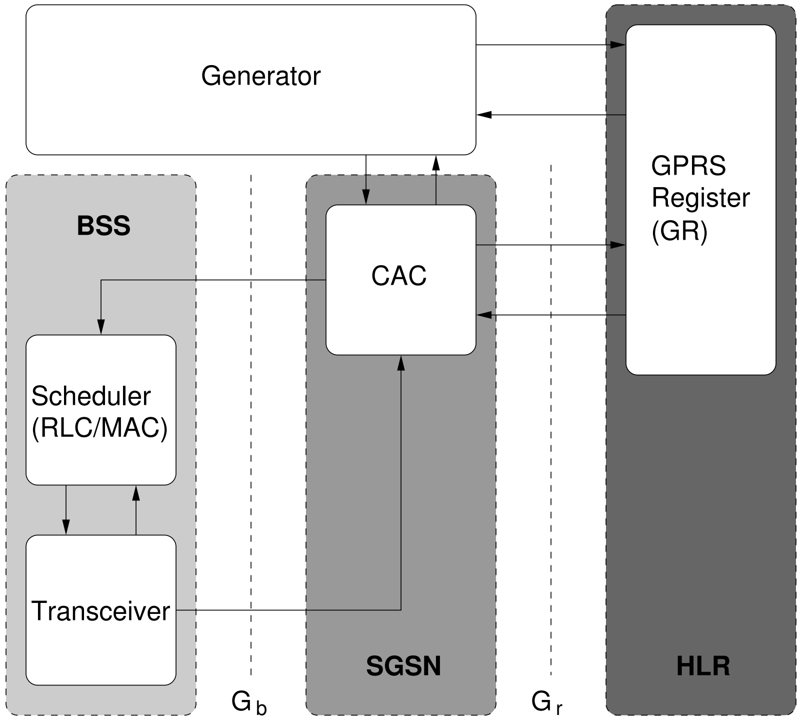


Figure 6.5: Traffic flow between the GPRSim modules involved in QoS management

6.5.4 Connection Admission Control (CAC)

As outlined in Section 3.5.3 the CAC function plays a decisive role in QoS-oriented traffic management in a GPRS environment. Within the scope of the GPRSIM, it has to fulfill various tasks, from *Admission Control* (AC) and QoS profile negotiation to radio resource status monitoring. Within the context of the GPRSIM the class CAC represents the QoS-management related functions of both the SGSN and the GGSN. Communication paths from MS via BSS to the CAC module within the GSNs are modeled only logically and do not utilize any physical signaling resources of the system.

As already shown in Figure 3.24, the module CAC is responsible for handling PDP context activation and deactivation requests, keeping track of traffic load and radio resource utilization status within the radio cell on behalf of BSSGP flow control information from the BSS, and performing PDP context renegotiation. Thus, there are interfaces necessary for communication with the modules GR, generator, BS transceiver, and the RLC/MAC modules via the SDL process environment.

The message flow between the module CAC and the other GPRSIM modules involved in the AC process is summarized in Figure 6.5.

The module CAC must provide methods and objects to store, update, and evaluate radio resource utilization data as well as traffic-flow-related QoS profile data (*Traffic Flow Template*, TFT). Further, it has to perform *Admission Control* on the basis of this data, and react to changing resource utilization in a radio cell by engaging PDP context renegotiation procedures.

To perform *Admission Control*, there is an AC policy to be implemented in a way that all active traffic flows can be served according to the QoS profiles negotiated. Preferably, there is only a limited number of privileged connections allowed simultaneously.

Furthermore, any *Standard* traffic should receive the resources necessary to meet its QoS requirements. Thus, it might be preferable to reject or downgrade a PDP context

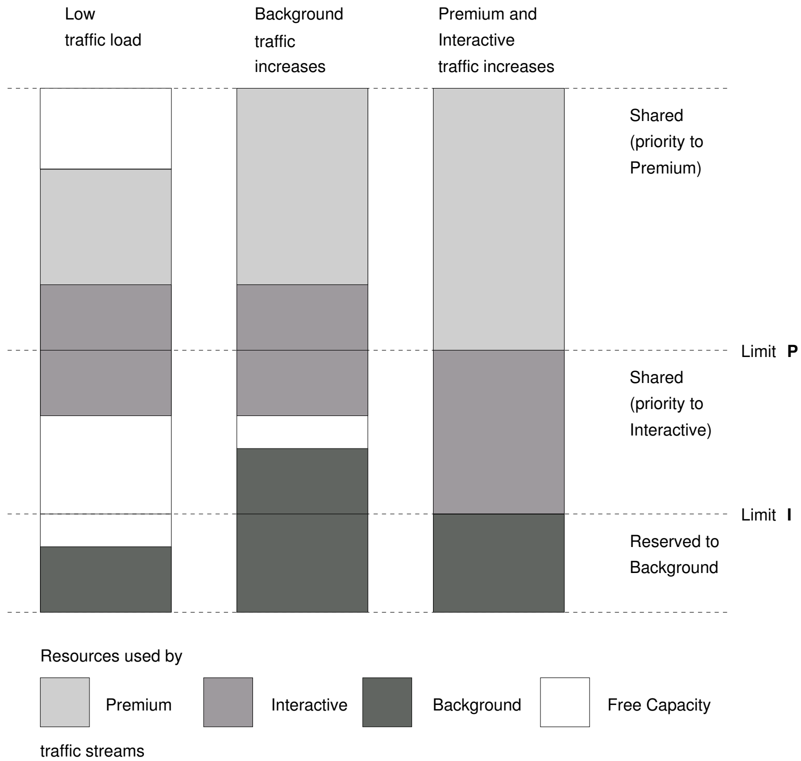


Figure 6.6: Admission control policy (example)

activation request rather than to endanger the quality of all sessions. On the other hand, it might be advantageous to displace background or even interactive traffic flows to allow for an additional *Conversational* traffic flow to be admitted.

In this CAC implementation, PDP context activation requests are differentiated according to their subscribed QoS profile and their traffic class. If additional PDP contexts are allowed for the requested QoS profile, the CAC decides on the basis of the radio resources available to admit the session or not.

The implementation for the traffic and subscriber classes Gold, Interactive and Background is shown in Figure 6.6.

To avoid a total withdrawal of resources from the *Standard* traffic classes with lower QoS requirements, e.g., other than *Streaming*, there is a share reserved for Interactive traffic from the pool of radio resources in the cell. In times of high load traffic flows with more demanding QoS requirements are allowed to displace flows belonging to applications with lower QoS requirements, but only up to a certain limit. The limits are specified by the maximum allowed number of active sessions for the regarded traffic class (thresholds P and I in Figure 6.6). When this limit is reached, the requested QoS is not accepted, but degraded to the next-lower-prioritized class.

6.5.5 Scheduler

On behalf of the QoS profile negotiated the BS MAC layer has to perform scheduling of the radio blocks (see Section 3.4.5.5). For evaluation purposes, only the downlink direction is regarded within this framework, since the traffic models used within the generator module of the GPRSim produce a considerably higher amount of data traffic for this direction, compared with the uplink. Thus, the downlink direction may be regarded as the bottleneck in respect of resource utilization. Traffic-class based uplink scheduling is performed during

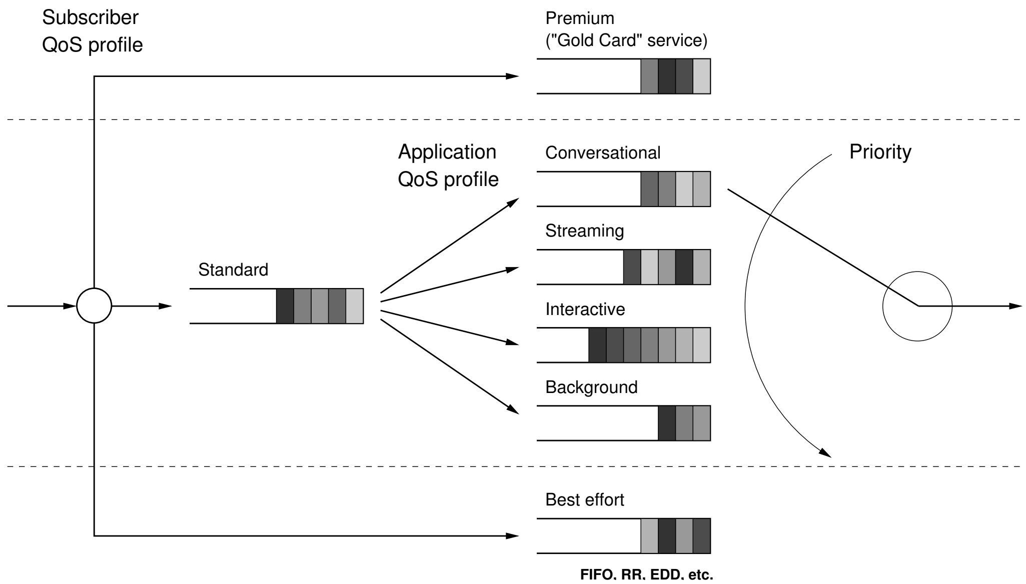


Figure 6.7: Principle of the scheduling function located in the BS MAC layer

USF assignment and channel request scheduling.

The scheduling mechanism implemented for both the downlink and uplink directions follows a three-stage principle (see Figure 6.7). First, incoming radio blocks are classified into one of three queues according to the QoS subscription associated with the respective traffic flow. It is differentiated between *Premium* (*Gold Card*) service, *standard* service and *best-effort* service.

The second stage is only valid for *standard* service traffic. Regarding a packet's application QoS profile, the appropriate traffic class queue is chosen from *Conversational*, *Streaming*, *Interactive*, or *Background*. Best-effort traffic from the first stage is put into a fifth queue. Within the traffic class queues, TBFs are scheduled using, e.g. a RR algorithm.

The third stage is built by, e.g. a priority mechanism, serving the traffic class queues in order from highest priority *(premium)* to lowest priority *(best-effort)*.

All three stages are part of the GPRSim RLC/MAC layer module and specified in SDL.

Enhanced scheduling algorithms were developed and implemented in the framework of this thesis and are presented in Section 10.2.

6.6 SNDCP Layer

The *Sub-Network Dependent Convergence Protocol* (SNDCP) class has two main functions. First TCP/IP header compression is performed (see Section 3.4.3). Then the IP packets coming from the IP layer are adapted to the LLC layer. As the *Maximum Transfer Unit* (MTU) in the LLC layer is 1520 byte, bigger IP packets have to be segmented and reassembled by SNDCP. In addition SNDCP has a queueing function in the GPRSim. With the function `put_ip_packet` the given parameter `ip_packet` is segmented into SNDCP PDUs and queued as LLC *Interface Data Unit* (IDU). When the LLC layer is ready to send new data, an LLC IDU can be dequeued.

6.7 LLC Layer

The LLC protocol is also specified in SDL. The class SDLProcess_SystemManagement represents the interface with the network layer. It prepares the signals received from the SDL process environment for the LLC layer and sends the associated signals to the environment. The classes SDLSystem_LLC_MS and SDLSystem_LLC_SGSN represent the LLC entities of the MS and the SGSN. They contain several processes that organize and control the transfer procedures. The class SDLProcess_Kernel is the kernel of the LLC layer. Here received frames and service primitives are processed and the related reactions are performed. In the kernel it is decided which frames can be sent and which services are offered to the network layer.

The kernel informs the class SDLProcess_FrameBuilder which frames have to be sent. This class composes the frames according to the frame type or command and transfers them with an RLC/MAC- or BSSGP-DATA-REQ to the underlying protocol instances.

The class SDLProcess_Decomposer receives an LLC PDU by an RLC/MAC- or BSSGP-DATA-IND. After processing the control field the information field is forwarded to the LLC kernel.

6.8 RLC/MAC Layer

This protocol implementation emulates the RLC/MAC standard [3GPP TSG GERAN (2002d)]. The full functionality of RLC and MAC is represented. The architecture of this layer in MS and BS is similar, but the functionality of the processes is different. In the following the structure and the functions of the BS MAC layer are described. After that the differences in the MS MAC implementation are outlined.

6.8.1 RLC/MAC in the Base Station

In Figure 6.8 the structure of the RLC/MAC layer in the base station and the interaction between the functional processes and the environment are shown.

After the start of a simulation the process prBS_Config obtains the configuration parameters from the environment and creates and initializes the instances prBS_RLC_entity, prBS_DL_FSM and prBS_UL_FSM for every mobile station associated with this base station. The other processes exist only once in the system. The data flow in the system is defined as follows. In the transmit direction LLC PDUs are requested from the higher layer always if the LLC queue is not empty and if there is enough space for new data to transmit in the send buffer of the RLC entity.

If necessary a new downlink TBF initiation is requested to the downlink *Finite State Machine* (FSM). The control blocks, which are part of the TBF establishment process, are given to the prBS_Transceiver and are transmitted as soon as radio resources are available. As soon as a TBF is established RLC blocks can be transmitted if downlink resources are available. These resources are managed by the BS transceiver, which requests new RLC blocks by asking the related RLC entity for new blocks to transmit. The order of RLC entities that are requested is defined by the scheduling function (see Section 3.4.5.5 and Section 6.5.5).

The coded radio blocks for each block period are forwarded to the prBS_PLL_Interface. This process forwards one block after the other to the environment at the correct time regarding the multiframe structure. In the receive direction radio blocks are received by the BS transceiver from the environment and are forwarded to the related uplink or downlink FSM. If these radio blocks contain RLC blocks, i.e., data or acknowledgement blocks, they

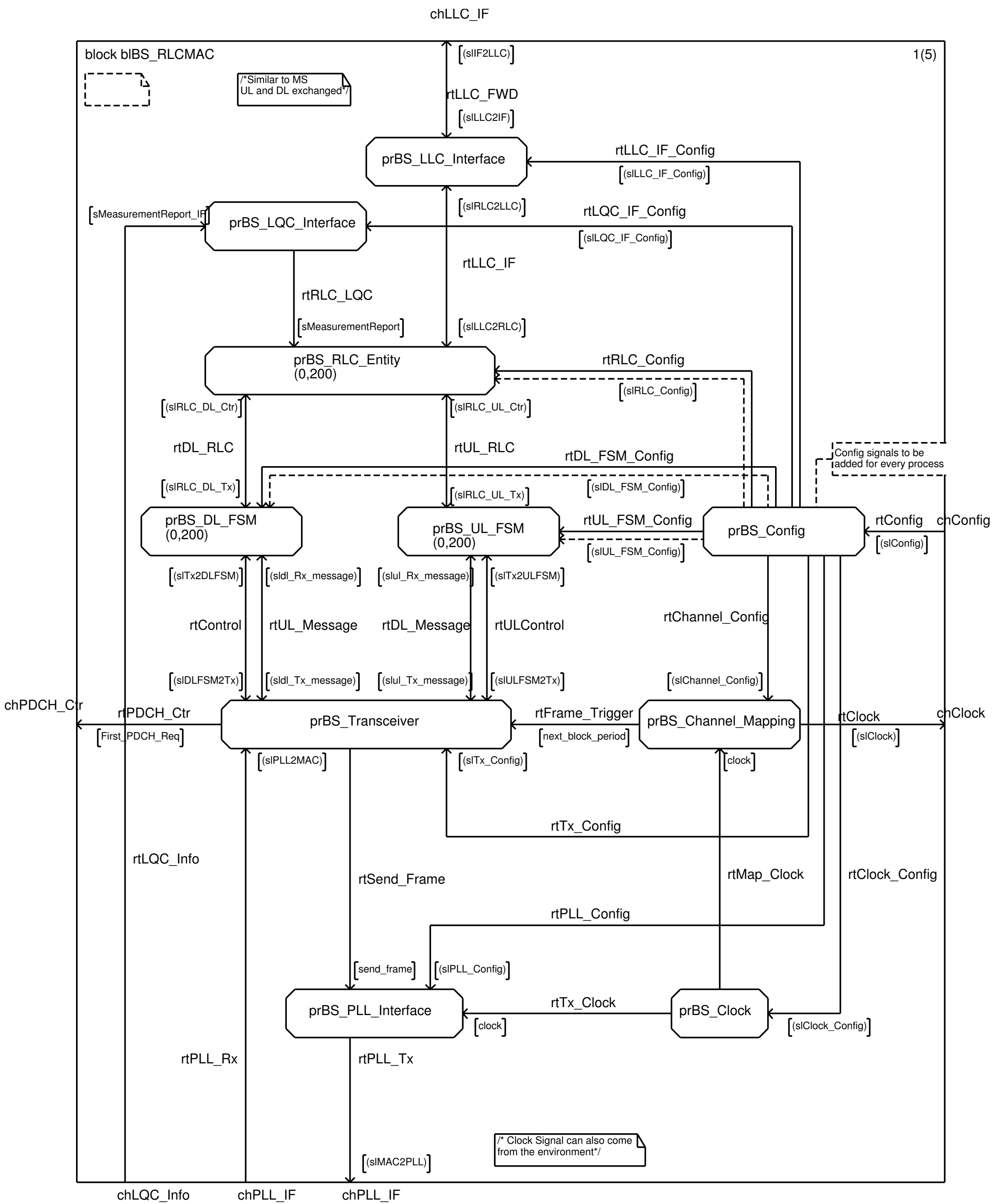


Figure 6.8: RLC/MAC base station block architecture

are forwarded by an FSM to the RLC entity. RLC data blocks are saved in the receive window and reassembled into LLC PDUs, which are finally given to the LLC layer.

6.8.2 RLC/MAC in the Mobile Station

In Figure 6.9 the structure of the RLC/MAC layer in the mobile station is shown.

The processes `prMS_PLL_Interface` and `prMS_Clock` are identical. Since there is only one logical link, one uplink TBF and one downlink TBF associated with the mobile station,

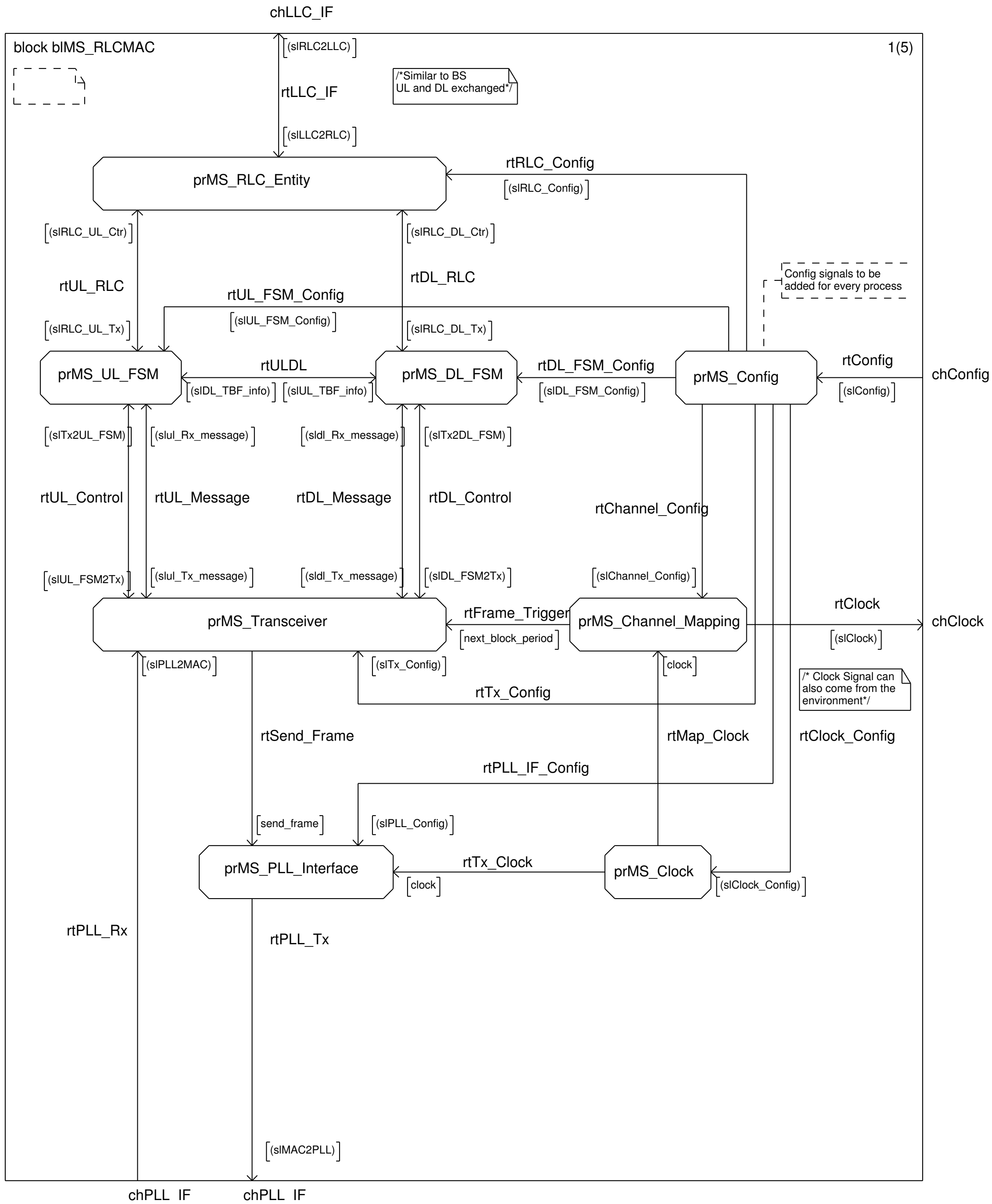


Figure 6.9: RLC/MAC mobile station block architecture

each process exists only once in the system initialized by the process `prMS_Config`. The LLC interface is not necessary since there is only one LLC entity and one RLC entity and thus no mapping function to be realized.

6.9 Air Interface Transmission Error Models

The transceiver modules have the task of forwarding the radio blocks which are given from the RLC/MAC layer to the peer entity or of forwarding the radio blocks which are received from the peer entity to the RLC/MAC layer. An error flag which has been added to the radio block data structure is set if a collision of packet channel request blocks of several MSs is detected or if the channel model generates a block error. Additionally the delay which corresponds to the duration of the block transmission on the channel interface is realized here.

6.9.1 Channel Models

The channel models define the state of the radio channel and the resulting block errors for uplink and downlink radio blocks. The GPRSIM provides three channel models.

6.9.1.1 Error-free Channel

This channel model assumes an error-free channel. It can be used for GPRS and EGPRS simulations. Only when a collision of packet channel request blocks of several MSs is detected the corresponding blocks are rejected by setting the error flag.

6.9.1.2 Channel Model with Uncorrelated Errors

This channel model is a model for pure GPRS simulations. Input of the error model is a fixed C/I ratio and a fixed *Coding Scheme* (CS) which can be selected before simulation starts. Control block errors are not regarded. The look-up function giving a *Block Error Probability* (BLEP) for every C/I ratio is derived from link level simulations [FURUSKÄR et al. (1999a); WIGARD and MOGENSEN (1996); WIGARD et al. (1998)]. A block error occurs if one or more bits in a coding block (1 block = 4 bursts) is erroneous after decoding. Perfect error detection is assumed. The TU3 (Typical Urban) channel model in GSM Rec. 05.05 [3GPP TSG GERAN (2002g)] is used in all link level simulations. Frequency hopping is not used. The channel fades according to 3 km/h on a 900 MHz channel. The results are obtained without antenna diversity.

In Figure 6.10 the BLER versus C/I that is the basis for the look-up function used in the uncorrelated channel model is shown. When radio blocks are exchanged between the transceiver modules, the channel model determines whether the arrived block is correctly received at the receiver or not.

6.9.1.3 Gilbert-Elliot Channel Model

This model is an enhanced channel model for pure GPRS simulations. This channel model is based on the *Gilbert model* [GILBERT (1960)], which is in turn based on a two-state *Markov chain*. This model offers the possibility of simulating burst errors. Figure 6.11 illustrates the model and the state transitions.

To simulate burst errors, the states B and G must tend to persist, which means that the transition probabilities p_{GB} and p_{BG} are small. The Gilbert model was extended by *Elliott* [ELLIOT (1963)], in so far as the probability for errors in the state *GOOD* is no longer zero. In this way there is the opportunity of simulating slight background noise in the good state.

6.9.1.4 Channel Model Considering Mobility

This channel model is usable for both EGPRS and GPRS. Figure 6.12 shows a block diagram of the model.

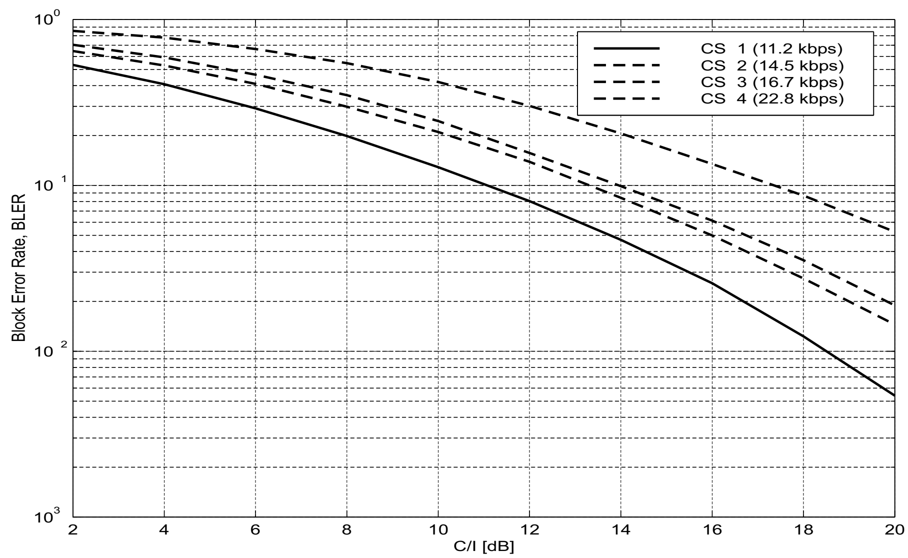


Figure 6.10: BLEP over C/I reference function used for channel model 1 (GPRS)

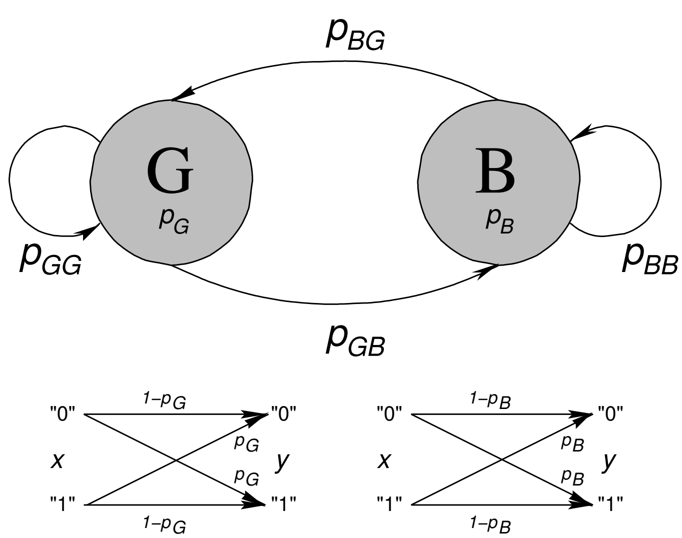


Figure 6.11: Gilbert-Elliott channel model

Table 6.1: EGPRS channel model parameters

Parameters	Default values
Velocity of the MS	100 km/h, 6 km/h
Cell radius	3 km, 300 m
Cluster size	3, 7, 9
C/I variance (shadowing)	7
Path loss parameter γ	3.7

To generate block errors a set of mapping functions is used. These are gained from link level simulations and allow the mapping of C/I values to the corresponding block error probabilities for every radio block as described in Section 6.9.1.2.

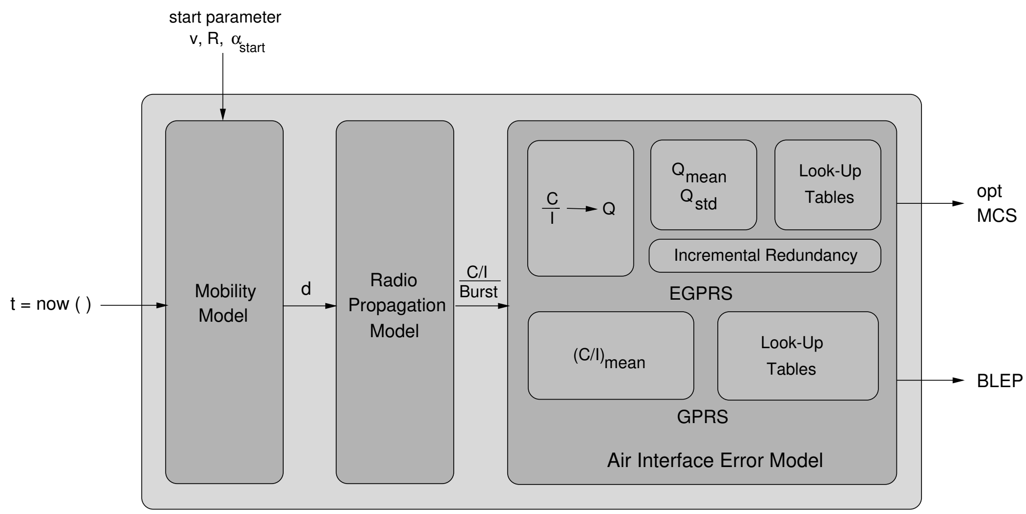


Figure 6.12: Block diagram of the EGPRS channel model

Beside the BLEP value the channel model with mobility determines the optimal MCS for the actual channel quality (see Figure 6.12).

The parameters for this channel model are listed in Table 6.1.

6.9.1.4.1 EGPRS

The value of the C/I ratio on burst basis is mapped onto a generic burst quality measure Q, which is directly related to the BLEP. This measure corresponds to a quality level that can be measured in a real radio receiver. It was introduced by equipment manufacturers to simplify the technical realization of a GPRS/EDGE receiver. The exact knowledge of the C/I is not necessary as long as a comparable measure is available which describes the quality of the radio link in a similar way. By measurement it is possible to map this measure applied in reality into the corresponding C/I ratio. A reference function mapping C/I to Q that was gained by measurement at a real receiver is shown in Figure 6.13 [Schieder (2003)].

The values Q_{mean} and Q_{std} are calculated for every data block. The number of bursts per data block depend on the selected MCS. If the MS operates with MCS-1–MCS-6, every data block is interleaved over four bursts. For MCS-7, MCS-8, and MCS-9 one data block is interleaved over two bursts. If IR is used, the Q values of the previous transmissions are considered. The Q_{mean} and Q_{std} values are then mapped to the BLEP using appropriate mapping functions, which are obtained from link level simulations [3GPP TSG GERAN (2002f); Furuskär et al. (1999a); Schieder (2003)]. For each MCS one mapping table is implemented in the simulator. As an example, two of these functions are shown in Figure 6.14.

6.9.1.4.2 GPRS

For GPRS the values of the C/I ratio on burst basis are used to calculate an average C/I. The average C/I is mapped directly to a BLEP. As described in Section 6.9.1.2 the reference function of BLEP versus C/I, which is taken from link level simulations, is used.

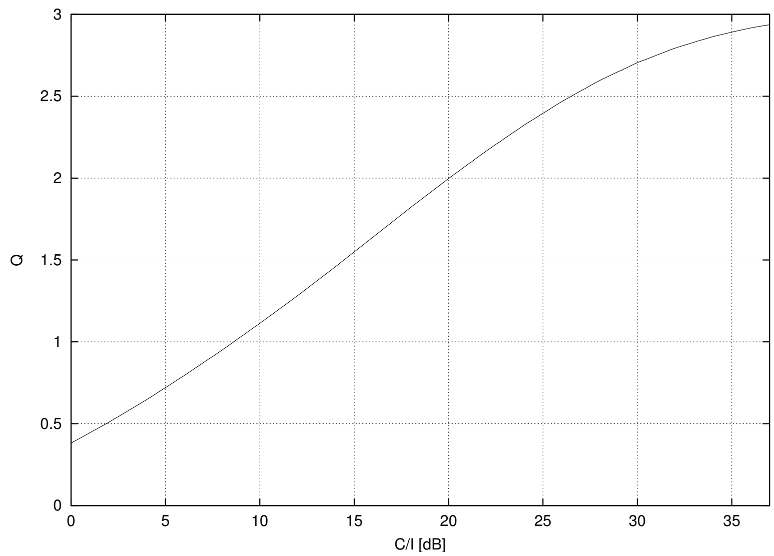
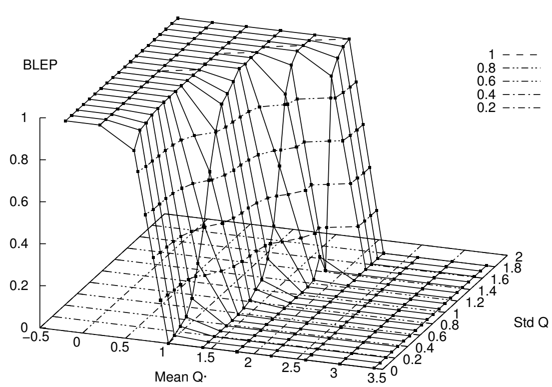
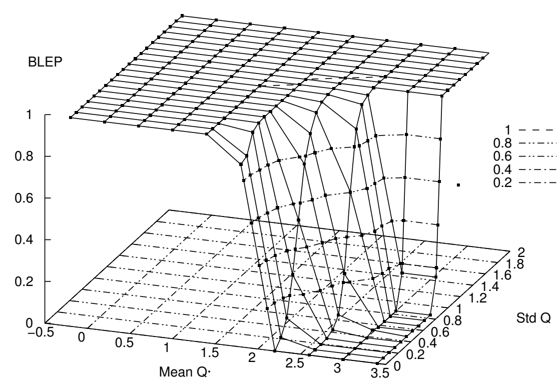


Figure 6.13: C/I to Q mapping used for the EGPRS channel model



(a) MCS-3



(b) MCS-9

Figure 6.14: Mapping of Q_{mean} and Q_{std} to BLEP

6.9.2 Mobility Models

The mobility model determines the actual distance between MS and BS within every time step of the simulation. The distance is used by the radio propagation model, together with statistical behavior assumptions of the channel and the transmission activities of MS and BS, to calculate the signal-to-interference ratio at the receiver. Within the channel model the BLEP is derived and the radio block is set to either error-free or defective. Two different mobility models have been implemented.

6.9.2.1 Secant Mobility Model

This mobility model determines the distance between *Base Transceiver Station* (BTS) and MS in every time step (a radio block period) of the simulation. During start-up time a randomly located start point on the cell border (determined by a start angle α_{start}) and a given velocity of movement are assigned to every MS.

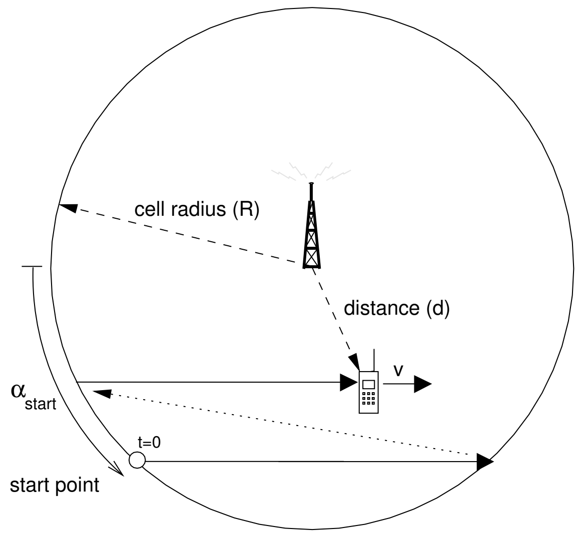


Figure 6.15: Secant Mobility model

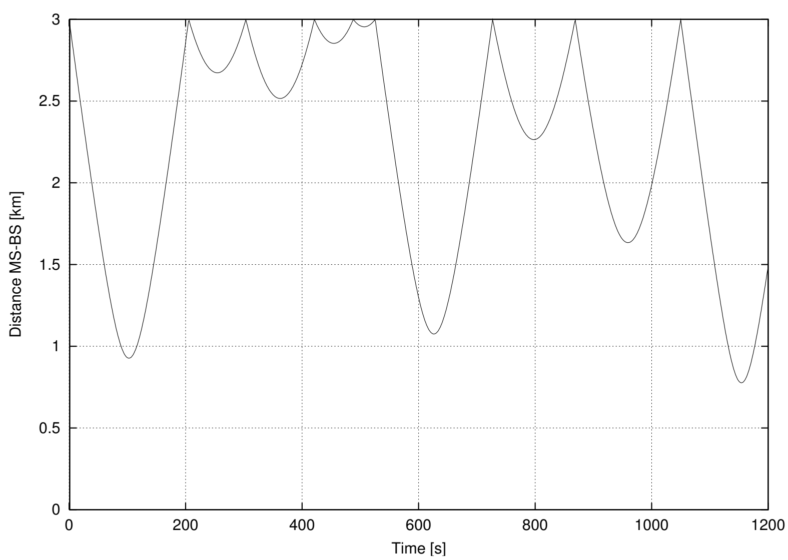


Figure 6.16: Distance between MS and BS using the Secant Mobility Model

When the simulation starts, the MS starts moving with the assigned velocity on a secant of the circular cell with radius R. If a mobile station reaches the cell boundary, it starts on the cell boundary and moves again on another secant.

Figure 6.15 illustrates how a single MS moves through the cell, reaches the cell boundary and starts again moving from a random location on the cell boundary. This model is a rather simple model and does not take into account the fact that mobile stations could change their direction or their velocity. In Figure 6.16 the actual distance between one MS and the BTS is shown over the simulation time where the MS has a velocity of 100 km/h.

6.9.2.2 Brownian Motion Mobility Model

This is an enhanced model to simulate movement of mobile stations more realistically. It is based on the *Brownian motion* of molecules [Karatzas and Shreve (1991)], whose

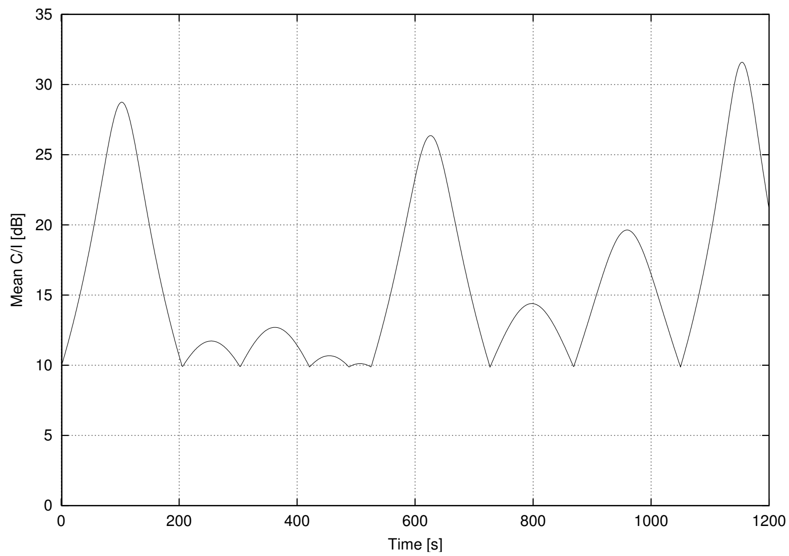


Figure 6.17: Mean C/I ratio over time, using the Secant Mobility Model

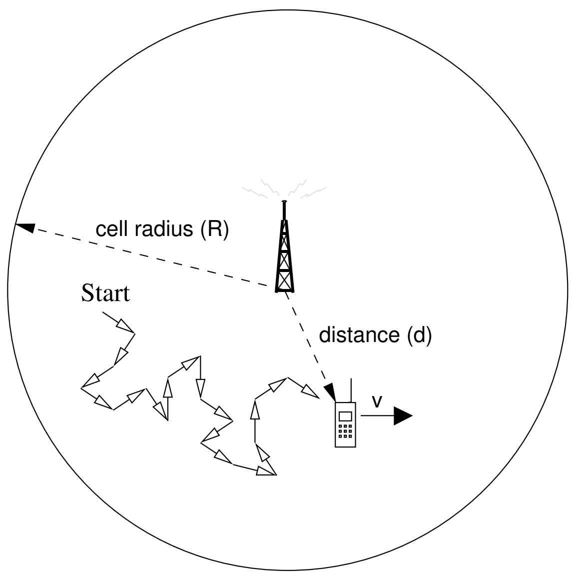


Figure 6.18: Brownian motion

speed and direction depend on the collisions among each other. This mobility model determines the distance between BTS and MS in every time step (a radio block period) of the simulation. During start-up time a randomly located starting point in the cell and a given velocity of movement are assigned to every MS. When the simulation starts, the MS starts moving in a randomly defined direction. In every time step, the direction is changed with a certain probability. If a mobile station reaches the cell boundary, it restarts at a randomly defined position in the cell. Figure 6.18 illustrates the movement of a single MS in the cell.

6.9.3 Radio Propagation Model

The influence of radio propagation is estimated in two steps. First, the C/I ratio can be calculated using:

$$\frac{C}{I}[\text{dB}] = 10 \cdot log\left(\frac{d^{-\gamma}}{6 \cdot D^{-\gamma}}\right) \text{ with } D = R \cdot \sqrt{3 \cdot C} \tag{6.1}$$

where d is the distance between MS and BS, R the cell radius, C the cluster size, D the co-channel distance and γ the path-loss factor.

Since the long-term fading mean value changes relatively slowly with the distance between MS and BTS, the distance-dependent calculation of the pathloss-related C/I value is carried out only once per radio block. The actual distance between MS and BTS required for this calculation is gained from the mobility model.

In a second step, long-term fading through signal shadowing is modeled by adding a normally-distributed random variable with a mean of zero and a given variance to the mean pathloss value (see Table 6.1). The calculation of the impact of shadow fading on C/I is performed on a burst basis (four times per radio block). Figure 6.17 illustrates how the mean C/I ratio alters over time if the distance between MS and BTS changes as shown in Figure 6.16. Short-term fading is already included in the link level simulations and needs therefore no further consideration at system level.

6.10 Statistics

After a simulation run, statistics are produced in a results file. *Inter alia* the following statistics (mean, variance, quantiles and distribution functions) in uplink and downlink and for all different application and subscriber classes are available (see also Section 8.2):

- IP user throughput
- IP datagram delay
- Application response time
- Session blocking rate
- CS call blocking rate
- IP cell throughput
- PDCH utilization
- Assigned PDCHs

6.11 Statistical Reliability of Simulation Results

For evaluation of stochastic simulations the statistical uncertainty concerning the accuracy of results especially for rare events has to be considered.

The control of the relative error is realized with the LRE algorithm [Schreiber (1988); Schreiber and Görg (1996)] which quantifies the needed number of samples based on the correlation between measured values. Because of the correlations in the traffic source and the channel error events a relatively large number of samples is necessary. A relative error of 5% was predefined in the LRE algorithm for the events with the lowest probability. For the calculation of the complementary distribution function of the IP datagram delay up to $P(T > t) = 10^{-3}$ a simulated time of $10^8\,T_{Slot} \approx 128\ h$ is necessary. This corresponds to several thousands of executed sessions per mobile station in the scenario.

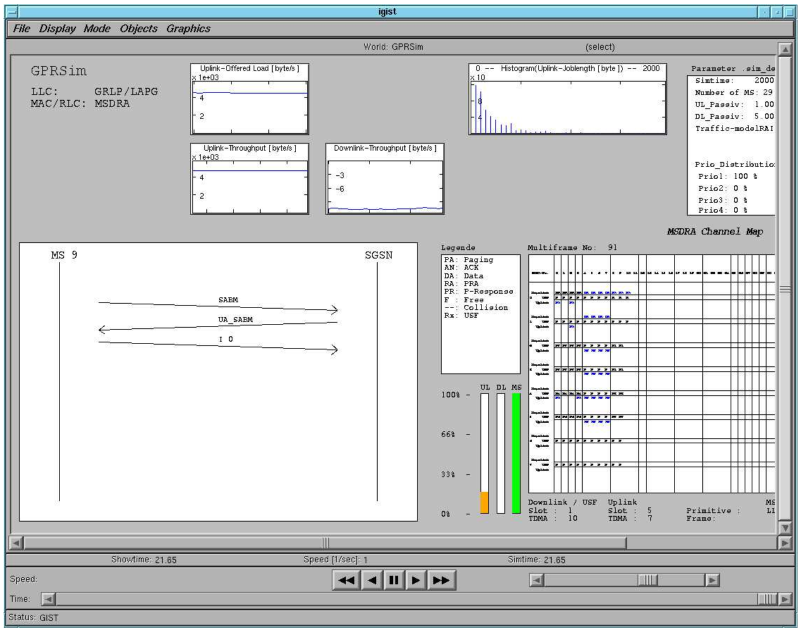


Figure 6.19: Graphical GIST presentation of the MAC and LLC transfer

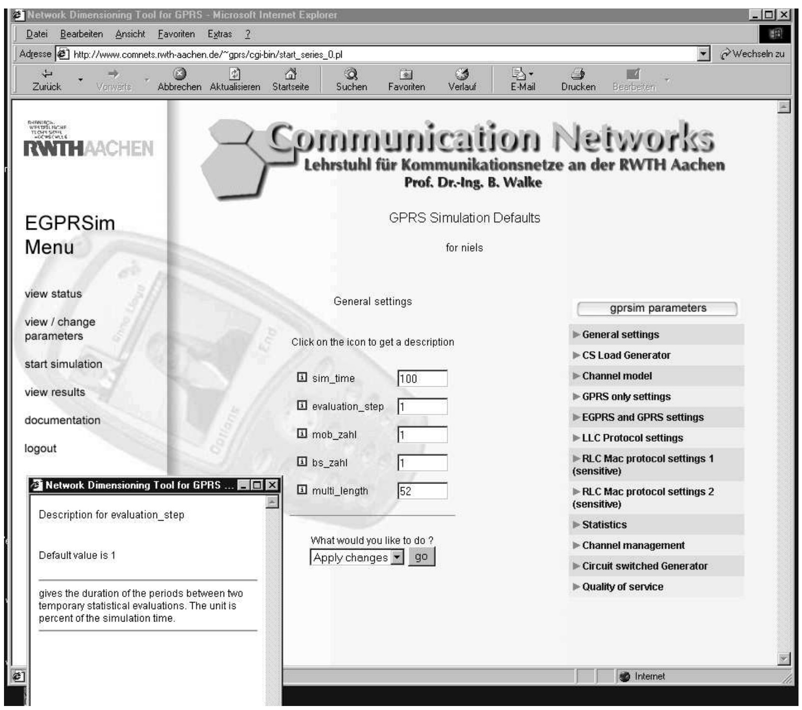


Figure 6.20: The GPRSim web interface

6.12 Graphical Presentation

To make the protocol functionality transparent the simulation flow can be presented with a graphical user interface. The tool *Graphical Interactive Simulation result Tool* (GIST) is utilized for this. GIST is an independent process called by the main program. An interface to exchange data with the simulation program is provided by this tool. Navigation over the simulation time can be done by using the buttons used in a video recorder. They provide the control of simulation sequences with a configurable speed.

The GIST graphic (see Figure 6.19) shows a 52-multiframe. For every time slot the bursts on uplink and downlink and the USF are pictured. Also the LLC protocol sequence is represented by LLC message sequence charts.

6.13 Web Interface

The web interface was created to simplify the simulator handling and to enable start and evaluation of simulation runs over the Internet.

After authentication, the user arrives at the start page of the simulator, where he can decide between the menu points simulation parameters and simulation progress. At the menu simulation parameters the user can change a certain number of parameters and start a new simulation after confirmation (see Figure 6.20). It is also possible to use a help function that explains the existing parameters. To avoid conflicts between several users and to guarantee confidentiality between users from different companies, for each user a new domain is created. At the menu simulation progress the user can supervise the progress of a current simulation run and read the temporary evaluation files that are created at specified intervals and at the end of the simulation.

6.14 Validation of the Simulation Model

Although it is presently not possible to estimate the performance of complex scenarios for Internet applications over GPRS by measurement or analysis, it is necessary to validate the correct functionality of the emulator at least in simplified scenarios. Especially the effects of *Temporary Block Flow* (TBF) management in the RLC/MAC layer, the protocol overhead caused by the air interface and the Internet protocol stacks and the effects of statistical multiplexing of several logical connections on the shared radio channels need to be validated.

The plausibility of the GPRSim results is shown by an analytical estimation of the TCP performance for typical WWW and e-mail object sizes in a simplified GPRS scenario. Additionally GPRSim results are validated by measuring the FTP performance in a real operational GPRS network.

6.14.1 Analytical Model for TCP over GPRS

In this section simulation results for a scenario with one mobile station per cell is compared with an analytical estimation of the TCP performance over GPRS.

A TCP transmission can be structured into three phases: TCP connection setup phase, TCP slow start phase and TCP steady state phase. The steady state performance of TCP has been studied extensively in literature [Chockalingham et al. (1999); Laksman and Madhow (1999)]. The results of these studies show that the TCP performance can be estimated, if the slow start phase of a TCP connection is small compared to the total transmission time. However, because of small file sizes and high round-trip times up to one second [Stuckmann et al. (2002)] in mobile packet networks the slow start phase of TCP often is in the same order of the steady state phase or even longer. Therefore, the slow start phase has been included into this analytical TCP performance estimation as described in [Peisa and Meyer (2001)]. The TCP setup time T_{init} can be added to the duration of the data transmission phase composed of slow start phase and steady state phase.

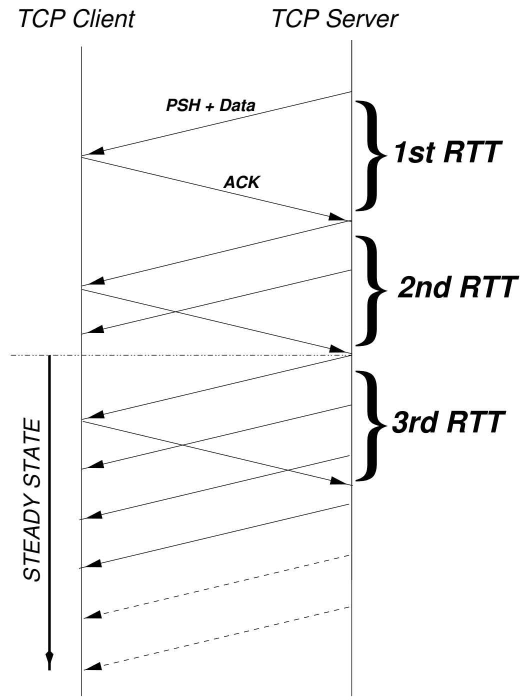


Figure 6.21: Message sequence chart for TCP slow start and steady state

6.14.1.1 TCP Slow Start

In the slow start phase the window is increased exponentially over time to approach the available bottleneck capacity. A round-trip period is defined as the period between the transmission of one or more segments and the reception of the acknowledge for this batch of TCP segments (see Figure 6.21). After each round-trip period the congestion window size is doubled. The steady state is reached, when the time to transmit the batch of TCP segments exceeds the round-trip period.

To be able to cover TCP implementations, which include delayed acknowledgements, the factor k_{ss} is introduced instead of the factor 2 [Cardwell et al. (2000)]. The factor k_{ss} is defined by 1 plus the reciprocal of the number of TCP segments, which are acknowledged by one acknowledgement message.

For the calculation of the transmission duration of a TCP slow start phase, files are regarded that can be entirely transmitted in the slow start phase ($F \leq B_{ss}$). With the file size F, the maximum segment size MSS and the number of started round-trip periods X, the following relation holds for a TCP transmission in slow start phase:

$$\left\lceil \frac{F}{MSS} \right\rceil \leq W_{init} \left(k_{ss}^0 + k_{ss}^1 + k_{ss}^2 + \ldots k_{ss}^X \right) = W_{init} \left(\frac{1 - k_{ss}^X}{1 - k_{ss}} \right) \tag{6.2}$$

Solving this inequation for X delivers the number of complete round-trip periods N_{RTT} needed for the transmission of the file.

Within these N_{RTT} round-trip periods B_{RTT} bytes can be transmitted. For X, N_{RTT} and B_{RTT} the following equations apply:

$$X = \left\lceil \frac{\log\left(1 - \left\lceil \frac{F}{MSS} \right\rceil \frac{1 - k_{ss}}{W_{init}}\right)}{\log(k_{ss})} \right\rceil \tag{6.3}$$

$$N_{\mathrm{RTT}} = X - 1 \qquad B_{\mathrm{RTT}} = W_{init} MSS \left(\frac{1 - k_{ss}^{N_{\mathrm{RTT}}}}{1 - k_{ss}} \right) \tag{6.4}$$

The TCP connection operates in the slow start phase until the slow start transmission rate reaches the steady state transmission rate R_{TCP}, which is the minimum of the capacity of the radio access bearer R_{radio} and the rate of the bottleneck link in the Internet R_{internet}.

After [Mathis et al. (1997)] the TCP transmission rate of an Internet link can be calculated to:

$$R_{\mathrm{internet}} = \frac{MSS \cdot C}{RTT \cdot \sqrt{p}} \tag{6.5}$$

RTT is the TCP round trip time and p is the Internet packet loss ratio. Assuming the radio link as the bottleneck link, the time needed to transmit an object of the size F in the status *slow start* using a radio access link with the capacity of R_{TCP} can be calculated as the number of complete RTTs N_{RTT} added to the time needed for the transmission of the remaining bytes of the file and added to the fixed end-to-end delay D_{LCH} for one segment. D_{LCH} is the end-to-end delay composed of the propagation delay of the air interface and of the delay of the core network and the Internet.

$$t(F \leq B_{ss}) = N_{\mathrm{RTT}} RTT + \frac{(F - B_{\mathrm{RTT}})}{R_{\mathrm{TCP}}} + D_{\mathrm{LCH}} \tag{6.6}$$

6.14.1.2 TCP Steady State

To calculate the transmission duration of files that are greater than the number of bytes that can be transmitted in slow start ($F > B_{ss}$), both the slow start phase and the steady state phase have to be considered.

Equation (6.6) can be applied for a TCP transmission until the sender transmission rate reaches the steady state transmission rate R_{TCP} (see Figure 6.21). This is the case, when the time to transmit the batch of TCP segments after X_{SS} round-trip periods exceeds the next round-trip period RTT.

$$RTT \leq \frac{W_{init} MSS}{R_{\mathrm{TCP}}} k_{ss}^{X_{ss}-1} \quad \Leftrightarrow \quad X_{ss} = \left\lceil \frac{\log\left(\frac{R_{\mathrm{TCP}} RTT}{W_{init} MSS}\right)}{\log(k_{ss})} + 1 \right\rceil \tag{6.7}$$

This condition for the transition to the steady state can be solved for X_{ss}. Now the number of round-trip periods the connection stays in slow start N_{ss} and the number of bytes the connection is able to transmit during the slow start phase can be calculated to

$$N_{ss} = X_{ss} - 1 \quad with\ B_{ss} = W_{init} MSS \left(\frac{1 - k_{ss}^{N_{ss}}}{1 - k_{ss}} \right) \tag{6.8}$$

Now the total transmission time can be calculated. It is the time of the slow start phase added to the transmission duration of the remaining bytes during steady state and added to the fixed end-to-end delay D_{LCH}.

$$t(F \geq B_{ss}) = N_{ss} RTT + \frac{(F - B_{ss})}{R_{\mathrm{TCP}}} + D_{\mathrm{LCH}} \tag{6.9}$$

6.14.1.3 GPRS Overhead and Signaling Delay

The analytical estimation for slow start and steady state can now be extended by additional delays and overhead introduced by the GPRS radio interface protocol stack.

After TCP segments are encapsulated with a 40 byte TCP/IP header, which is assumed to be reduced by the SNDCP header compression function to 5 byte on average, the SNDCP header of 3 byte is added. SNDCP Protocol Data Units are the payload of LLC frames. The LLC frame format comprises an address field consisting of 1 byte, a control field typically consisting of 3 bytes for information frames and a FCS field of 3 bytes. Assuming a typical TCP MSS of 512 byte, the LLC frame size results in 527 byte.

The LLC frames are segmented into a number of RLC blocks depending on the GPRS Coding Scheme (CS). Using CS-2 with an RLC block payload size of 30 byte and an average radio block transmission duration of 20 ms the RLC/MAC data rate $R_{RLC/MAC}$ is resulting in 12 kbit/s per PDCH. Assuming a multislot capability of 4 in downlink direction, up to 4 timeslots per TDMA frame can be used.

Block errors on the radio channel additionally reduce the throughput. Since a *Selective-Repeat* (SR) ARQ protocol is applied in the GPRS RLC layer the normalized throughput v_{SR} is reduced by the block error rate [Lin and Costello (1983)].

$$v_{\mathrm{SR}} = 1 - BLER \tag{6.10}$$

Considering the overhead of TCP/IP, SNDCP and LLC headers and an RLC/MAC block error rate of 13.5 %, the effective bearer data rate R_{TCP} can be calculated to:

$$R_{\mathrm{TCP}} = 4 \cdot R_{\mathrm{RLC/MAC}} \cdot (1 - 0.135) \cdot \frac{512}{527} = 40.34\,\mathrm{kbit/s} \tag{6.11}$$

In GPRS the round-trip time is mainly influenced by the establishment of logical RLC/MAC connections in uplink and downlink, whenever a new block flow has to be transmitted. These logical connections are called *Temporary Block Flows* (TBFs) (see Section 3.4.5.1). It is assumed that for each batch arrival of TCP segments and for the corresponding TCP acknowledgement message a new TBF has to be established.

Adding the uplink and downlink TBF setup times TBF_{setup} to the round-trip time RTT in Equation (6.7), X_{ss} can be calculated to:

$$X_{ss} = \left\lceil \frac{\log\left(\frac{R_{\mathrm{TCP}}(RTT+TBF_{setup})}{W_{init}MSS}\right)}{\log(k_{ss})} + 1 \right\rceil \tag{6.12}$$

For the uplink TBF setup (two-phase access) one half of a block period as the waiting time before the next random access channel is detected, 4 block periods for the exchange of the messages *packet channel request*, *packet uplink assignment*, *packet resource request* and *packet resource assignment*, and 4 block periods as processing delays after the reception of these messages are assumed. For the downlink TBF setup one half of a block period waiting time before the detection of the packet access grant channel, 1 block period for the packet downlink assignment and 1 block period for processing are assumed. Hence, the round-trip time is increased by 220 ms through the uplink and downlink TBF setup.

Like in the GPRSim the transmission delay in the core network and external networks is neglected compared to the TBF setup time corresponding to a scenario where the server is located in the operator's domain. The round-trip time is only composed of the transmission delay of one TCP segment and one TCP acknowledgement message over the air interface added to the TBF setup times.

The end-to-end delay D_{LCH} equals the propagation delay of the air interface and the delay of the core network and the Internet and is assumed to 100 ms.

With these assumptions the duration of the slow start and steady state phases for a TCP transmission over GPRS can be calculated. Finally the TCP connection setup and the client request have to be modeled.

6.14.1.4 TCP Connection Setup

For the three-way handshake two uplink and one downlink TBF setup durations (19.5 periods), three radio block periods for three RLC data blocks, one half of a radio block period for the initial waiting time and two radio block periods for the processing delay after the radio block reception are assumed resulting in 500 ms. For the client request a 200 byte TCP segment is assumed that is segmented into 8 RLC data blocks and transmitted in two radio block periods. Together with the initial waiting time for the random access channel and the uplink TBF setup, the duration of the client request transmission is estimated to 220 ms. Hence, the total setup time T_{init} including the three-handshake and the client request results to 780 ms.

6.14.1.5 Result Comparison

The simulation results are compared with this analytical estimation for a scenario with one mobile station per cell with 4 downlink slots, and a BLER of 13.5 %. Additionally RTT is assumed to 360 ms (transmission of one TCP segment and one ACK) and D_{LCH} to 110 ms [Sachs and Meyer (2001)].

With the average application packet sizes according to the traffic models presented in Section 5.1.1 and Section 5.2.2 of 3700 byte (WWW) and 10000 byte (e-mail), Equation (6.12) and Equation (6.9) can be applied with the parameters RTT = 360 ms, TBF_{setup} = 220 ms, D_{LCH} = 110 ms, R_{TCP} = 40.34 kbit/s and T_{init} = 720 ms. The results of the mean TCP throughput for the simulation with the GPRSim and for the analytical estimation are presented in Table 6.2.

This comparison shows that the results are in the same order of magnitude, which proves the plausibility of the simulation results produced by the GPRSim. The differences can be explained with the stochastic model of the traffic sources and the radio channel in the simulation model. For the analytical model only the mean value of the file size and the mean block error rate are regarded.

Table 6.2: Comparison of simulation and analysis

	WWW (3700 byte)	e-mail (10000 byte)
analytical model	15.0 kbit/s	24.8 kbit/s
simulation (4 fixed PDCHs)	17.2 kbit/s	22.9 kbit/s

6.14.2 Validation by Measurement

While in Section 6.14.1 the behavior of TBF management and protocol overhead in the GPRSim is validated for a scenario with one mobile station per radio cell, the effects of statistical multiplexing of several stations per cell are validated by measurement in an operational GPRS network.

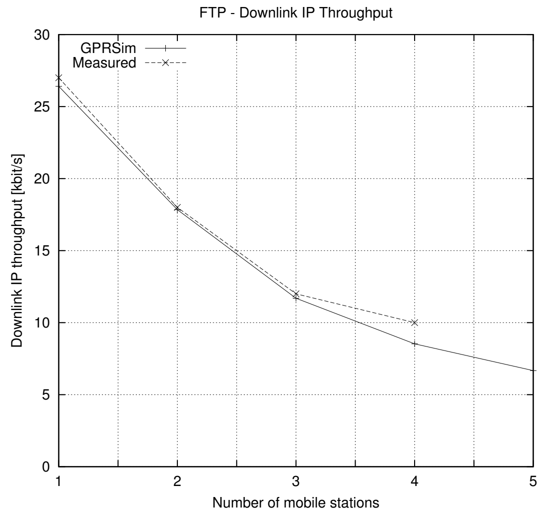


Figure 6.22: FTP throughput performance compared to GPRSim results

The measurements were performed in December 2001 in the Vodafone NL building in Maastricht [Stuckmann et al. (2002)]. The GPRS terminals were connected to laptops. The load generator is represented by files located on a high-speed Web server in the Internet (round-trip time via local area network smaller than 40 ms). The average throughput was measured in intervals of five seconds so that a time-weighted mean value could be calculated out of these samples.

The settings in the GPRS simulator GPRSim were adjusted to meet the measurement conditions. The multislot capability of three downlink slots and the FTP traffic model were chosen. The comparison shows very similar results in simulation and measurements (see Figure 6.22). For one, two and threee active stations the throughput is the same, while for four active stations the simulation result is slightly lower, which can be explained by cell reselections with four active stations in the measurement scenario. This comparison can be seen as a validation of the GPRSim regarding the statistical multiplexing function of the RLC/MAC layer and the overhead of the whole protocol stack.

CHAPTER 7

Applicability of Analytical Traffic Engineering Approaches

After a comprehensive emulation tool for simulative performance analysis of (E)GPRS networks has been presented in the previous chapter, this one here gives an overview of the state of the art in analytical modeling of mobile packet networks carrying Internet traffic. To show the limitations of analytical approaches the results, gained by analysis, are compared with simulation results.

Analytical modeling often leads to the problem of finding an acceptable trade-off between accuracy and analytical tractability. The process of finding the optimal compromise between both aspects has to be guided by an understanding of the various influence factors of the regarded quantities and the relation between them. Based on such considerations, the dominant aspects of the real quantities can be identified, depicted in the selected model and unimportant aspects can be neglected.

In the first analytical approach presented in Section 7.1, the applicability of the *Fluid-flow Model* (FFM) [Kosten (1974); Mitra (1988)] is examined. The FFM has been successfully applied for analytical dimensioning of ATM multiplexers [Fiedler and Voos (1997)].

The main result of this modeling approach is the prediction of the *Cumulative Distribution Function* (CDF) of the equilibrium buffer size under a given traffic load which allows the calculation of the mean equilibrium buffer size by integration. In using Little's Law, the mean waiting time can be calculated from the mean buffer size.

In the second approach the applicability of two different models based on queueing systems is examined (see Section 7.2). The functional principle of most communication protocols is well suited for modeling as a queueing system (see Figure 7.1). Regarding, for example, an ARQ protocol (i.e., the GPRS RLC protocol), the service offered by protocol instances to their users can be interpreted as a queueing system. The arrival of data units at the corresponding *Service Access Point* (SAP) can be described by an arrival process, which is defined by the inter-arrival times. Additionally, the distribution of the size of arriving packets can be used to determine the service time distribution of the system.

In Section 7.2.1 and Section 7.2.2 the arrival process and the service process are determined by simulation using the GPRSim (see Chapter 6). Based on the evaluated mean values of the interarrival and service times, an M/M/n model is analyzed and compared to simulation results in Section 7.2.3 to show the high influence of the source traffic model. The second queueing analysis approach in Section 7.2.4 is based on a much more precise characterization of the arrival process based on the *Marked Markovian Arrival Processes* (MMAP) [He (1996)], that was introduced for GPRS in [Vornefeld (2002a,b)]. Taking into account that the segmentation of IP datagrams into RLC blocks creates batch arrivals on RLC level and approximating the behavior of an empirically derived traffic source model for WWW with 10 negative exponentially distributed phases, a closer approximation of the arrival process is gained. Accordingly, GPRS performance measures are derived by analysis of an MMAP/G/1 queueing system.

The results of both approaches, FFM and MMAP, are compared with results of the GPRS emulation tool GPRSim that is representative for the performance of real GPRS networks (see Section 6.14). The GPRSim employs a detailed WWW model for traffic generation and is modeling the influence of TCP, contrary to the analytical models. The

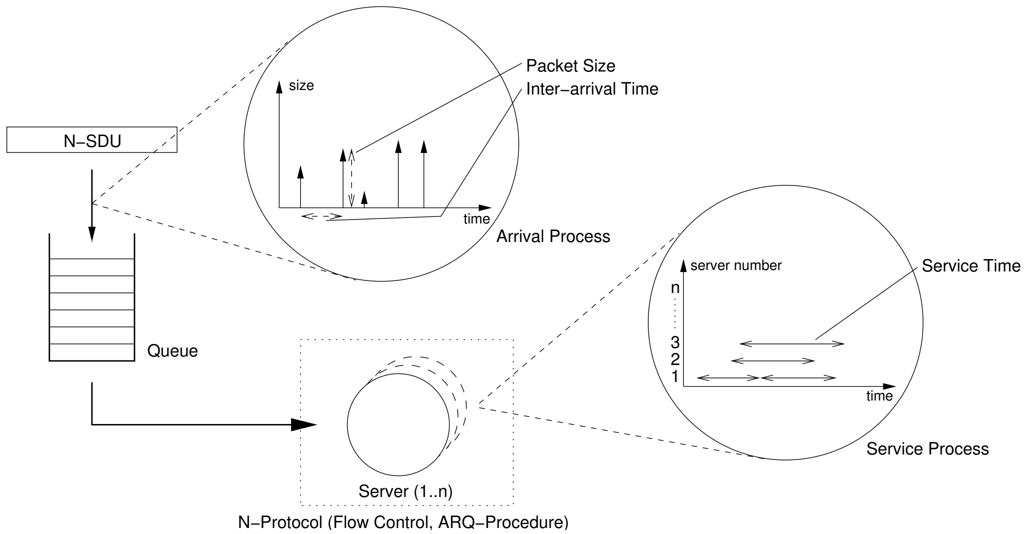


Figure 7.1: Modeling a communication protocol as a queueing system

result comparison illustrates the influence of accurate source modeling and the limitations of the analytical approaches to model the GPRS system carrying elastic traffic.

7.1 Applicability of the Fluid-flow Model

The *Fluid-flow Model* (FFM) is based on a concept developed by [Kosten (1974)], and was extended by [Anick et al. (1982)]. The analysis in this thesis is proceeding from the description and the notation of [Fiedler and Voos (1997)].

The main application of the FFM is predicting the CDF of the equilibrium buffer size under a given traffic load. After calculating the mean amount of backlogged traffic in using Little's Law, the mean waiting time of an IP datagram can be calculated.

In the scope of the FFM, the arrival of data generated by a number of traffic sources at the point of interest is compared to water falling into a reservoir (the network element's buffer memory), which depletes at a constant rate C. The unit of data is assumed to be infinitely small. Comparing traffic sources to water taps, which are asynchronously turned on and off, the sources alternate between an *ON* state and an *OFF* state. The sojourn times in the ON and OFF states are exponentially distributed. While being in the ON state, a source transmits data at a constant rate h. Superimposing arbitrary numbers of not necessarily equal ON/OFF sources allows the coverage of a large range of load scenarios.

The scheduling strategy applied in the simulation of the GPRS system is *Round Robin* (RR). Thus the FFM waiting time is not directly compared to the waiting time of data packets in a real GPRS system, but the *FFM Sojourn Time* is introduced as an analytical quantity which is equivalent to the IP datagram delay, evaluated in the simulation system. It is derived by adding an offset for the mean transmission duration to the FFM mean waiting time, based on evaluation of the mean IP datagram size, the mean number of RLC blocks per IP datagram and the minimum time that is required for transmission of these blocks.

To employ the FFM in a parameter range representing typical GPRS load scenarios, the according parameters needed for definition of the ON/OFF sources are determined. To achieve this, ON/OFF source parameters representing the mean offered traffic in typical GPRS load scenarios are derived by stochastic simulation. By using a synthetic traffic source model for simulation an equivalent load scenario for both simulation and analysis can be established eliminating inaccuracies evolving from approximating WWW traffic source behavior.

7.1.1 Source Modeling

A single traffic source is modeled by a *Markov Modulated Rate Process* (MMRP), called ON/OFF source or *Interrupted Rate Process* (IRP). Multiple equal subscribers are modeled by superposition of several IRPs, called NIRP [Fiedler and Voos (1997)].

7.1.1.1 Interrupted Rate Process (IRP)

Each IRP is controlled by a two-state *Markov Chain* (MC) with states Λ_0 and Λ_1. In state Λ_0 the source transmits packets at the rate $r_0 = 0$ and in state Λ_1 the source transmits packets at the rate $r_1 = h$. The transition rates between the ON state Λ_1 and the OFF state Λ_0 are λ and μ (see Figure 7.2). The sojourn times in each state are negative exponentially distributed.

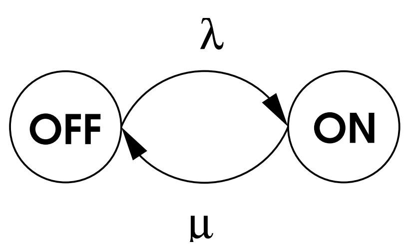


Figure 7.2: State transition diagram of a single IRP

The steady state probability of the ON state is called the *activity factor* α, which represents the fraction of time the source is active.

Each IRP needs the following parameters to be described completely:

- Activity factor α
- Mean burst length EN_B, $[EN_B] = byte$
- Peak bit rate h, $[h] = \frac{byte}{s}$

From these parameters, the transition rates λ and μ can be calculated to

$$\mu = \frac{h}{EN_B} \tag{7.1}$$

$$\lambda = \mu \cdot \frac{\alpha}{1-\alpha} \tag{7.2}$$

and the mean transmission rate of the source is

$$M = \alpha \cdot h \tag{7.3}$$

7.1.1.2 Superposition of ON/OFF Sources (NIRP)

Superimposing N equal IRPs, a so-called N *Interrupted Rate Process* (NIRP) can be defined [Anick et al. (1982); Fiedler and Voos (1997)], adding the number of superimposed IRPs N to the otherwise unaltered set of parameters. A NIRP can be described by a one-dimensional MC (see Figure 7.3). The state variable of this MC is the number of active IRPs at an instant, so that the total number of states becomes $N + 1$.

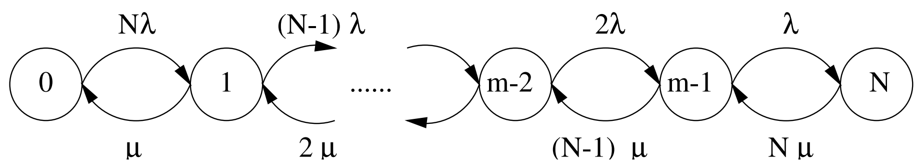


Figure 7.3: NIRP state transition diagram

With the number of sources in the ON state q the corresponding transmission rate r_q can be calculated for each activity state Λ_q to:

$$r_q = q \cdot h \tag{7.4}$$

and the NIRP's mean transmission rate is

$$M = N \cdot \alpha\, h. \tag{7.5}$$

7.1.1.3 State Space Subdivision

Knowing the total rate r_q and the capacity C, the state space of a NIRP can be subdivided into:

- $\Lambda^u = \{\Lambda_q \in \Lambda \text{ with } r_q < C\}$: set of underload states
- $\Lambda^e = \{\Lambda_q \in \Lambda \text{ with } r_q = C\}$: set of uniform load states
- $\Lambda^o = \{\Lambda_q \in \Lambda \text{ with } r_q > C\}$: set of overload states

In an underload state, the buffer content depletes at the rate $C - r_q$, in an overload state the buffer content rises with rate $r_q - C$ and in a uniform load state the buffer content remains constant.

7.1.2 Fluid-flow Analysis

Following the presentation in [Fiedler and Voos (1997)], the general case of a Fluid-Flow multiplexer with buffer size K and N equal ON/OFF sources is regarded, which means that the complete arrival process is represented by a NIRP.

7.1.2.1 The Fundamental Differential Equation System

The equilibrium probability of state q, with the buffer of maximum capacity K being filled with x bytes of data waiting for transfer is:

$$F_q(x,K) = Pr\{X \leq x \text{ and } \Lambda_q\} \quad ; x \leq K$$

Combining $F_q(x,K), q = 0, \cdots, N$ to the vector

$$\vec{F}(x,K) = \{F_0(x,K), \cdots, F_N(x,K)\}\,, \tag{7.6}$$

a system of differential equations describing the system behavior can be calculated [Anick et al. (1982); Fiedler and Voos (1997)]. Matrix notation leads to

$$\mathbf{D}\frac{d}{dx}\vec{F}(x,K) = \mathbf{M}\vec{F}(x,K). \tag{7.7}$$

$\mathbf{D}$ is the so-called *Drift Matrix*

$$\mathbf{D} = \begin{pmatrix} r_0 - C & & 0 \\ & \ddots & \\ 0 & & r_n - C \end{pmatrix}$$

With *Rate Matrix* $\mathbf{R} = diag\{r_0, \cdots, r_n\}$ and identity matrix $\mathbf{I}$, the Drift Matrix can be written as

$$\mathbf{D} = \mathbf{R} - C\,\mathbf{I}.$$

7.1.2.2 General Solution Approach

The general approach to solve the differential equation system (7.7) is:

$$\vec{F}(x,K) = a(K)\,\vec{\varphi}\,e^{(zx)} \tag{7.8}$$

Applying Equation (7.7) to this approach leads to the following eigenvalue problem:

$$z\vec{\varphi} = \mathbf{D}^{-1}\mathbf{M}\vec{\varphi} \tag{7.9}$$

where z is an eigenvalue and $\vec{\varphi}$ is the corresponding right eigenvector of matrix $\mathbf{D}^{-1}\mathbf{M}$. The coefficients $a(K)$ in Equation (7.8) are needed to fit the solution to the boundary conditions and thus depend on the available buffer memory K.

7.1.2.3 Eigenvalues

In [Stern and Elwalid (1991)] it is shown that in case of an irreducible MC (which is given in case of a NIRP), the eigenvalue of activity state Λ_0 is

$$z_0 = 0$$

and the corresponding eigenvector is

$$\vec{\varphi}_0 = \vec{\Pi}.$$

The eigenvalues z_q can be calculated to:

$$\begin{aligned} z_q \;=\; & \frac{1}{2\,(C - qh)\,(C - (N-q)\,h)} \\ & \cdot\Big(NC\,(\lambda+\mu) - N^2h\lambda + 2\,(N-q)\,qh\,(\lambda-\mu) + \\ & (2q - N)\cdot\sqrt{C^2\,(\lambda+\mu)^2 + (Nh\lambda)^2 - 2NhC\lambda\,(\lambda+\mu) + 4\,(N-q)\,qh^2\lambda\mu}\Big) \end{aligned} \tag{7.10}$$

Special cases that have to be treated separately are:

1. $C = qh$:
   The eigenvalue is undetermined.
2. $C = (N-q)h$:
   The eigenvalue is given by:

$$z_q = \frac{2(N-q)q(\lambda+\mu)^2}{N(2C - Nh)(\lambda+\mu) - (N-2q)^2(\lambda-\mu)}$$

7.1.2.4 Eigenvectors

In [Anick et al. (1982)] the eigenvector components are calculated in closed form except two special cases ($q = 0$ and $q = N$), which are treated separately.

In [Fiedler and Voos (1997)] the authors present an alternative way of calculating the eigenvector components.

The generating function for the eigenvector components is

$$\Phi_q(x,z) = \sum_{i=0}^{N} \vartheta_{q,i}(z) \cdot x^i, \tag{7.11}$$

where $\vec{\vartheta}_q(z)$ is the eigenvector of the inverse eigenvalue problem of Equation (7.9). It is related to the eigenvector $\vec{\varphi}_q$ by:

$$\vec{\vartheta}_q(z_q) = \vec{\varphi}_q \quad \forall q. \tag{7.12}$$

$\Phi_q(x,z)$ can be determined by solving the following differential equation:

$$\frac{\partial}{\partial x}\Phi_q(x,z) = \frac{z\gamma_q(z) - N\lambda + N\lambda x}{\lambda x^2 + (\mu - \lambda + zh)x - \mu} \cdot \Phi_q(x,z) \tag{7.13}$$

Defining $res_{q,1/2}(z)$ to be the distinct real roots of the denominator of the right-hand side

$$res_{q,1/2}(z) = \frac{1}{2\lambda}\left((\lambda - \mu - zh) \pm \sqrt{(\mu - \lambda + zh)^2 + 4\mu\lambda}\right), \tag{7.14}$$

Equation (7.11) can be written as

$$\Phi_{\vec{q}}^{(j)}(x, z_{\vec{q}}) = \sum_{i_1=\Gamma_1}^{0}\left(\binom{\Gamma_1}{i_1}\left(-res_{\vec{q},1}^{(j)}(z_{\vec{q}})\right)^{\Gamma_1 - i_1} \cdot \sum_{i_1=\Gamma_1}^{0}\binom{\Gamma_2}{i_2}\left(-res_{\vec{q},2}^{(j)}(z_{\vec{q}})\right)^{\Gamma_2 - i_2} x^{i_1+i_2}\right) \tag{7.15}$$

with

$$\Gamma_1(z) = N_j - \Gamma_2(z) = \begin{cases} q & z > 0 \\ (N-q) & z < 0 \end{cases} \tag{7.16}$$

Every addend contributes to the coefficient of the (i_1+i_2)th power of x and thus to the (i_1+i_2)th component of eigenvector $\vec{\varphi}_q$. Thus, the components are not calculated separately. Except for the factor $x^{i_1+i_2}$ the sums are decoupled, which causes the factor

$$f_1(i_1) = \binom{\Gamma_1}{i_1}\left(-res_{q,1}(z_q)\right)^{\Gamma_1 - i_1}$$

to be constant while the second sum is running. For each decrement of the first sum's index, the corresponding factor can be derived from its predecessor by

$$f_1(i_1 - 1) = f_1(i_1) \cdot \left(-res_{q,1}(z_q)\right) \cdot \frac{i_1}{\Gamma_1 - i_1 + 1}$$

starting with $f_1(\Gamma_1) = 1$.

After calculating all components of an eigenvector, the sum of the eigenvector's components can be normalized to one by dividing each component by the component sum of the unnormalized eigenvector.

In the context of this work, the sum of all components of eigenvector $\vec{\varphi}_q$ is of special interest, which can be obtained by evaluating Equation (7.15) for $x = 1$:

$$\Phi_q(1, z_q) = \sum_{i_1=0}^{\Gamma_1}\left(\binom{\Gamma_1}{z_q}\left(-res_{q,1}(z_q)\right)^{\Gamma_1}\sum_{i_2=0}^{\Gamma_2}\binom{\Gamma_2}{i_2}\left(-res_{q,2}(z_q)\right)^{\Gamma_2 - i_2}\right) \tag{7.17}$$

7.1.2.5 Coefficients

After calculating the eigenvalues and eigenvectors, the coefficients $a_q(K)$ remain to be determined (see Equation (7.8)).

Since in the simulation model as well as in real GPRS systems high memory capacity is available compared to the offered downlink traffic at the base station, the solution for the pure storing system is regarded in the following. According to [Fiedler and Voos (1997)], the following assumptions can be established for this case:

1. In an overload state, the buffer cannot be empty:
$$\lim_{x \to 0} F_q(x,\infty) = 0 \quad \forall q \quad \text{with} \quad \Lambda_q \in \Lambda^o$$
2. The eigenvalues of the underload states are positive, except for $z_0 = 0$. The solution has to be bounded for all values of $x, 0 < x < \infty$. This leads to
$$a_q(\infty) = 0 \quad \forall q \quad \text{with} \quad \Lambda_q \in \Lambda^u$$
An exception to this is given by the coefficient $a_0(\infty)$, which can be determined to
$$a_0(\infty) = 1$$
3. The coefficients of the uniform load state have to be zero, because the eigenvalue for this state is undetermined:
$$a_q(\infty) = 0 \quad \forall q \quad \text{with} \quad \Lambda_q \in \Lambda^e$$

According to [Anick et al. (1982)], the coefficients can be calculated in closed form as

$$a'_q(\infty) = -\alpha^N \prod_{\substack{s \in \Lambda^0 \\ s \neq q}} \frac{z_s}{z_s - z_q}$$

In [Fiedler and Voos (1997)], the eigenvectors are normalized to the sum of all components being equal to one (see Section 7.1.2.4). Since the eigenvectors in [Anick et al. (1982)] are normalized to the n-th component being $\varphi_{q,N} \equiv 1$, the coefficients have different values than the coefficients according to [Fiedler and Voos (1997)].

In [Fiedler and Voos (1997)] a relation between the $a'_q(\infty)$ and the $a_q(\infty)$ is presented, which is used in the scope of this work:

$$a_q(\infty) = -\frac{\alpha^N}{\varphi_{q,N}} \prod_{\substack{s \in \Lambda^0 \\ s \neq q}} \frac{z_s}{z_s - z_q} = -\alpha^N \, \Phi_q(1,z_q) \prod_{\substack{s \in \Lambda^0 \\ s \neq q}} \frac{z_s}{z_s - z_q} \tag{7.18}$$

where $\varphi_{q,N}$ represents the N-th component of the corresponding eigenvector, the component sum of which is normalized to one, and where $\Phi_q(1,z_q)$ denotes the sum of eigenvector components without normalization.

7.1.3 Performance Measures

7.1.3.1 Equilibrium Buffer Size

$F_q(x,K)$ represents the equilibrium probability of state q regarding a buffer of size K filled with x packets (see Section 7.1.2.1):

$$F_q(x,K) = Pr\{X \leq x \text{ and } \Lambda_q\}$$

Assuming an unlimited buffer size the parameter K can be omitted. In order to be able to distinguish the index for the solution vector component from the sum index, the index representing the component of the solution vector will be substituted with the letter i.

$F_i(x)$ can be used to calculate the distribution function of the equilibrium buffer size regarding an unlimited buffer, which is defined to

$$F(x) = Pr\{X \leq x\}.$$

Then the *Complementary Cumulative Distribution Function* (CCDF) or, as it is called in [Anick et al. (1982)], the probability of overflow beyond x is defined as

$$G(x) = Pr\{X > x\} = 1 - F(x).$$

and can be calculated from the vector $\vec{F}(x) = (F_0(x), \cdots, F_N(x))$ as

$$G(x) = 1 - \vec{1} \cdot \vec{F}(x) = 1 - \sum_{\forall i} F_i(x)$$

Then, the distribution function of the equilibrium buffer size can be derived to:

$$F(x) = \sum_{\forall q} F_i(x) \tag{7.19}$$

The i-th component of vector $\vec{F}(x)$ can be derived from Equation (7.8) to

$$F_i(x) = \sum_{\forall q} a_q(\infty)\, \varphi_{q,i}\, e^{z_q\, x},$$

where $\varphi_{q,i}$ represents the i-th component of eigenvector $\vec{\varphi}_q$. Replacing $F_i(x)$ in Equation (7.19) leads to

$$F(x) = \sum_{i=0}^{N} \sum_{\forall q} a_q(\infty)\, \varphi_{q,i}\, e^{z_q\, x}.$$

Rearranging the sums delivers

$$F(x) = \sum_{\forall q} a_q(\infty)\; \Phi_q(1,z_q)\; e^{z_q x}.$$

Taking into account that the eigenvalue z_0 equals zero, the eigenvector $\vec{\varphi}_0$ equals the vector of the NIRP steady state probabilities, the component sum of which is one, and regarding Equations (7.18) and (7.18), we finally receive for the buffer content distribution function:

$$F(x) = 1 + \sum_{q \in \Lambda^o} a_q(\infty) e^{z_q x} \tag{7.20}$$

and thus for the CCDF (compare [Fiedler and Voos (1997)], pp 92-93):

$$G(x) = -\sum_{q \in \Lambda^o} a_q(\infty) e^{z_q x}. \tag{7.21}$$

The mean equilibrium buffer size is equal to the area below the CCDF and can be calculated to

$$\begin{aligned} E[x] &= \int_0^\infty G(x)dx = \int_0^\infty (1 - F(x))\, dx \\ &= \sum_{q \in \Lambda^o} \frac{a_q(\infty)}{z_q} \end{aligned} \tag{7.22}$$

7.1.3.2 Waiting Time and Sojourn Time

From the equilibrium buffer size $E[x]$ and the arrival rate M, the mean waiting time can be obtained by Little's Law [Kleinrock (1975)]:

$$t_w = \frac{E[x]}{M} = \frac{1}{M} \sum_{q \in \Lambda^o} \frac{a_q(\infty)}{z_q} \tag{7.23}$$

Due to the different scheduling paradigms in the scope of FFM and GPRS, respectively, the waiting time cannot be compared directly. The GPRS MAC protocol does not apply strict FIFO queueing of the aggregated traffic (see Section 3.4.5.5 and [Stuckmann (2002a); Stuckmann and Möller (2003)]).

Starting with the mean IP datagram waiting time delivered by the FFM analysis, a measure is defined that can be compared with the IP datagram delay at the GPRS air interface. To achieve this, an offset is added for the minimum transmission time of an IP datagram over the air interface. The resulting quantity is called the *FFM Sojourn Time* [Irnich and Stuckmann (2003)].

First the duration of a *Radio Block Period* (RBP) and of a TDMA frame is introduced

$$t_{\mathrm{RBP}} = 18.46ms \tag{7.24}$$

$$t_{\mathrm{TDMA}} = 4.615ms. \tag{7.25}$$

The determination of the mean transmission duration is based on the evaluation of the mean IP datagram size $\overline{N_{\mathrm{IP}}}$ in our simulation system. The CDF of the downlink IP datagram size shown in Figure 7.4 has a mean of 340 bytes. Since this measure is only depending on the traffic source and not on the underlying network, this graph can be regarded as representative for all simulation scenarios.

Division of $\overline{N_{\mathrm{IP}}}$ by $N_{\mathrm{RLC,CS}}$, the number of data bytes in one RLC block (depending on the regarded coding scheme, see Section 3.4.5.6), leads to the mean number of RLC blocks needed for transmission of an IP datagram. The multislot capability (MSC) of a GPRS MS determines how many time slots the MS is able to use in a TDMA frame (see Section 3.4.5). Thus, the number of RLC blocks per IP datagram additionally has to be divided by the multislot capability, denoted by N_{msc}, to determine how many radio block periods the transmission of an IP datagram takes. The number of radio block periods is denoted by the letter R:

$$R = \left\lceil \left\lceil \frac{\overline{N_{\mathrm{IP}}}}{N_{\mathrm{RLC,CS}}} \right\rceil \cdot \frac{1}{N_{\mathrm{msc}}} \right\rceil \tag{7.26}$$

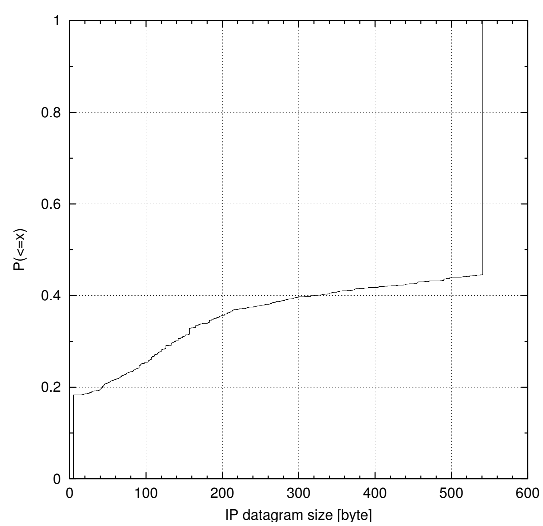


Figure 7.4: Downlink IP datagram size distribution

Multiplying with t_{RBP} then leads to the mean transmission delay of an IP datagram. Considering a processing offset between uplink and downlink by two TDMA frame durations and one idle frame after every 3 radio block periods, $\lfloor R/3 \rfloor + 2$ times the duration of one TDMA frame has to be added:

$$t_s = t_w + \left(R \cdot t_{\mathrm{RBP}} + \left(\left\lfloor \frac{R}{3} \right\rfloor + 2 \right) \cdot t_{\mathrm{TDMA}} \right) \tag{7.27}$$

During the simulations that were performed to determine the FFM source parameters (see Section 7.1.4.2) the parameters needed for deriving the FFM sojourn time were evaluated additionally. In case of an error-free radio channel the mean RLC block service time is 28 ms. Assuming CS-2 (29 byte payload per RLC block, see Section 3.4.5.6) this leads to a mean number of 12 RLC blocks per IP datagram, which results in a mean service time of 336 ms per IP datagram.

7.1.4 GPRS Modeling with ON/OFF Sources

In order to determine FFM parameter values representing typical GPRS applications, the FFM parameters h, α and EN_B of the GPRSIM WWW traffic generator module (see Section 6.2 and Section 5.1.1) have been evaluated. Using these parameters to define an ON/OFF source, an equal mean load can be generated.

7.1.4.1 IRP Parameter Evaluation Procedure

The evaluation of the ON period is triggered by monitoring the TCP flags in the uplink direction. Once an uplink packet carrying the first segment of a TCP connection (SYN flag), the next downlink packet is counted as the first packet of a burst. The last packet of this burst is defined to be the last TCP segment of a TCP connection (FIN flag). Thus, one burst is equal to the amount of data transmitted during the lifetime of one TCP connection.

This evaluation procedure is affected by the HTTP *keep-alive* feature. If keep-alive is used, several Web objects belonging to the same Web page are transferred using the same TCP connection, which is equivalent to defining the download of a whole Web page as the ON phase of the WWW traffic model. Disabling keep-alive causes one TCP connection per Web object to be established. In this case an ON phase of the WWW model is defined as the download of a single Web object.

7.1.4.2 ON/OFF Source Parameters for WWW-equivalent Load

To generate a mean offered traffic equivalent to real WWW traffic, the IRP source model parameters have been evaluated in the following simulation scenario:

- 1 MS
- Traffic Mix 100 % WWW
- 4 fixed PDCHs
- CS-2
- Error-free radio bearer
- Multislot capability (dl/ul) 4/1

The resulting values, which allow definition of ON/OFF sources generating an equal mean load are presented in Table 7.1. It is shown that in fact the HTTP keep-alive feature has a significant influence on the parameters that define the corresponding equivalent ON/OFF source.

The enabled keep-alive feature results in a higher activity factor, caused by the longer duration of the TCP connections in this case. The mean transmission rate is higher, because the longer lifetime of a TCP connection leads to a higher maximum TCP send buffer size and thus reduces the impact of TCP flow control. Since the mean burst size is equal to the mean amount of data transferred in a TCP connection, a longer lifetime of a TCP connection and consecutive downloading of several Web objects leads to a higher mean burst size.

Table 7.1: FFM parameters for WWW over GPRS (1 MS)

keep-alive enabled	keep-alive disabled
$\alpha = 0.187346$	$\alpha = 0.0859222$
$h = 3272\,{}^{\mathrm{byte}}/_{\mathrm{s}}$	$h = 2735\,{}^{\mathrm{byte}}/_{\mathrm{s}}$
$EN_{\mathrm{B}} = 9149\,\mathrm{byte}$	$EN_{\mathrm{B}} = 3677\,\mathrm{byte}$

7.1.4.3 GPRS System Capacity

In case of an error free radio channel the GPRS system capacity can be calculated to

$$C = N_{\mathrm{PDCH}} \cdot R_{\mathrm{PDCH}},$$

where N_{PDCH} denotes the number of PDCH available for GPRS and R_{PDCH} denotes the maximum IP throughput of a single PDCH, depending on the applied coding scheme.

7.1.5 FFM Results and Comparison with Simulation Results

To eliminate inaccuracies evolving from deviations in source behavior between simulation and analysis, an implementation of the ON/OFF source model, parameterized to generate the same mean offered traffic as a single WWW traffic source was used on top of the GPRS protocol stack. Thereby an identical source behavior in both analysis and simulation was achieved, allowing an evaluation of the potential of the Fluid-flow approach to model the GPRS protocol stack and its performance.

7.1.5.1 IRP Implementation

Each `IRP` instance is operating on its own LLC connection, which is kept established throughout the whole simulation. This means that data from all IRPs is served by a round robin algorithm [Stuckmann (2002a)], which is similar to the FFM assumption that all data in the buffer is treated equally.

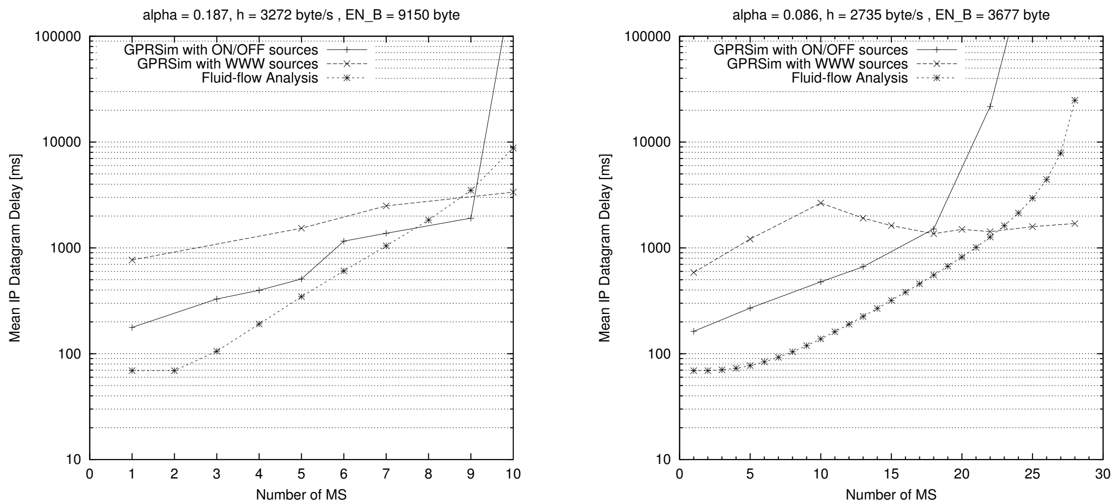
In order to generate a mean data arrival rate h during the ON state, the given IP MTU is divided by h to determine the required inter-arrival time. Once a source switches to the ON state, it draws a sample from the burst size random generator (the burst size is characterized by a negative exponential distribution with mean `enb`) and starts to generate IP datagrams with an inter-arrival time as calculated before. Once the given burst size is reached, the source draws the length of the following OFF interval and switches to the OFF state.

7.1.5.2 Performance Evaluation

The ON/OFF source parameters were chosen according to Section 7.1.4.2. These parameter sets have been used in the simulation with ON/OFF sources and the Fluid-flow analysis, using both parameter sets (with and without HTTP keep-alive). The results for WWW with keep-alive are shown in Figure 7.5(a) and for WWW without keep-alive in Figure 7.5(b).

In both cases the Fluid-flow analysis and the simulation with ON/OFF sources show a similar tendency and the same order of magnitude. This shows that if the elastic property of WWW traffic is eliminated, the performance degradation of GPRS under an increasing traffic load can be depicted using FFM analysis. On the other hand it is visible in both cases that the FFM underestimates the IP datagram delay (more than 100%), which is caused by the conservative assumptions for deriving the transmission delay offset (see Section 7.1.3.2). The selected definition of the FFM sojourn time leads to a lower bound of the IP datagram delay and thus allows prediction of the theoretical maximum performance of GPRS under the given traffic load.

The third graph in both diagrams additionally represents a simulation series with the traditional WWW application model generating GPRS traffic with the same mean offered traffic. In the range of low system load the mean IP datagram delay is higher than predicted by FFM analysis. Advancing to higher system load a lower performance degradation than in both FFM analysis and



(a) keep-alive enabled

(b) keep-alive disabled

Figure 7.5: Mean IP datagram delay, FFM parameters representing WWW over GPRS

simulation with inelastic traffic sources can be recognized. The TCP flow control procedures (i.e., TCP slow start and congestion avoidance) applied by the WWW traffic source are responsible for this behavior, because they influence the burstiness of the IP arrival process. These procedures represent a means of adaptation of the offered traffic to the available bandwidth. With increasing offered traffic in the system less capacity is available for each TCP connection, which causes TCP to reduce the offered traffic by decreasing its sliding window [Stevens (1996)].

One single WWW load generator instance causes a higher load than the same instance would, when operating on a bearer sharing the same capacity with several sources. To express this in terms of the Fluid-flow source parameters, the load of the GPRSim simulation scenario with 1 MS has a lower activity factor and a higher ON state transmission rate than the load of the simulation with 20 MSs. In the range of small load this causes higher delays than predicted by the FFM, because the source transmission rate in the active state is higher. Advancing into the range of high load, the delay does not rise as fast as the FFM predicts, because the traffic sources reduce the transmission rate during the active phases.

7.2 Queueing Systems

Since queueing systems in general are defined by an arrival process, a service process and a queueing discipline, modeling a communication protocol by means of queueing systems requires definition of these components.

In this section, two different queueing systems are applied for modeling the GPRS system. The arrival process is defined on IP level (see Section 7.2.1). Since for the service process a load-independent behavior is needed, it is defined on RLC level (see Section 7.2.2).

In Section 7.2.3 a simple M/M/n model parameterized by characteristic values of the arrival and service processes evaluated in Section 7.2.1 and Section 7.2.2 is considered. The second examination in Section 7.2.4 is based on a precise characterization of the arrival process based on the MMAP [Vornefeld (2002a,b)].

In the next two sections the IP arrival process and the RLC/MAC service process are evaluated by simulation with the GPRSim.

7.2.1 Characterization of the IP Arrival Process

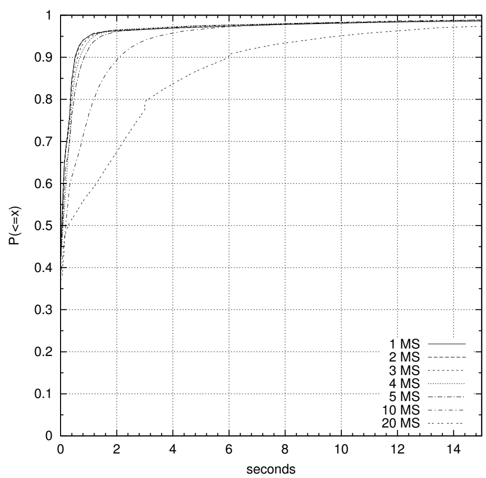
The GPRS system load is defined by random processes of IP datagram size and inter-arrival time, which defines the IP arrival process. In Section 7.2.1.1 the IP datagram inter-arrival time is characterized by mean value and CDF evaluation. It is shown that WWW users tend to adapt the offered traffic to the capacity share available for the single user.

7.2.1.1 IP Datagram Inter-arrival Time

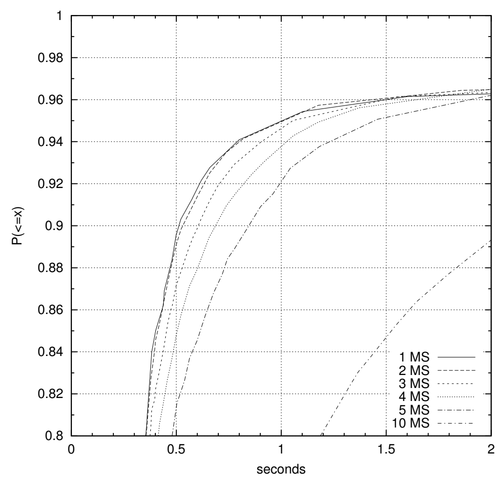
The IP datagram inter-arrival time can either be defined on a per-user basis or on datagrams arriving from all users in a sum process. It is evident that a higher system load causes the IP inter-arrival times to become longer, because the TCP flow control algorithm reduces the load generated by a single source in an environment of degrading performance.

Regarding the interval from 0 s to 2 s and the probability range from 80 % to 100 % (see Figure 7.6(b)) shows that this effect is already taking place in the range of a relatively low system load. Only the CDFs for 1 MS and 2 MS do not show the influence of TCP flow control.

In Figure 7.6(a) the IP inter-arrival time CDF that was evaluated based on the offered traffic of a single user is depicted.



(a) Total range



(b) Lower range

Figure 7.6: IP datagram inter-arrival time per single source

7.2.1.2 IP Datagram Size

The CDF of the IP datagram size as already presented in Section 7.1.3.2 is again shown in Figure 7.7. It is firstly characterized by two discrete values, the minimum and maximum datagram size. The minimum datagram size is defined by the TCP/IP header length. If TCP header compression is applied, this length has the value of 8 byte. The maximum datagram size is defined by the IP MTU, resulting in a maximum value of 540 byte. Between these two borders the values are nearly uniformly distributed.

7.2.2 Characterization of the RLC/MAC Service Process

The RLC service process is defined by the duration needed for the successful transfer of one RLC block to its peer entity. Since for most applications the downlink direction represents the bottleneck, focus is set on the characterization of the downlink RLC service process.

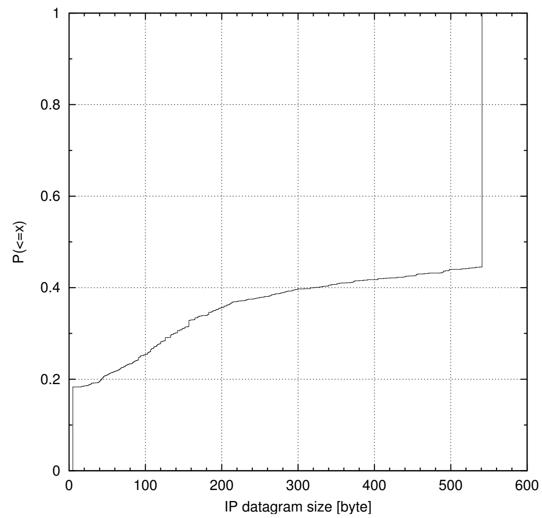


Figure 7.7: Downlink IP datagram size distribution

When a radio block is transmitted over the air interface, a block error occurs with a certain probability, depending on the condition of the radio channel and the applied CS.

A radio block that is erroneous and can not be corrected by the FEC function is discarded and thus has to be transmitted again. This is controlled by the ARQ protocol that assigns a sequence number to each transmitted data block, which enables the peer entity to detect the loss of data. After detection of gaps in the order of sequence numbers or after time-outs, the peer entity informs the sender about the missing datagrams by sending a NACK message (see Section 3.4.5.3).

Since an erroneous transmission is an event with a certain probability, the user of a reliable transfer service encounters a non-deterministic service time, i.e. the service time of a data block is a random variable. The lower bound of the service time distribution is the time needed for a single (error-free) transfer of a data block.

In case of GPRS, the RLC protocol is responsible for the reliable transfer of data blocks. The efficiency of the ARQ procedure executed by the RLC protocol has a significant influence on the transfer delay over the air interface.

7.2.2.1 RLC Service Time Evaluation Procedure

In order to evaluate the probability density function of the RLC/MAC service time process, the GPRSIM was instrumented to calculate the service duration for each radio block transmitted in downlink direction. The evaluation samples have been rounded to an integer number of milliseconds.

To achieve this, a timestamp was added to the RLC block data structure, which represents the instant of the block being handed from the RLC to the MAC layer for the first time as response to a polling signal. The polling signal is transferred at the beginning of a radio block period when radio resources are allocated to the regarded logical connection. Thus, this instant denotes the start of service. The end of service is defined by the instant when the RLC block is correctly received by its peer entity. The difference between these two points in time is defined as the *RLC block service time*.

A typical simulation scenario was chosen, in the first step comprising one mobile station with a multislot capability of 4 downlink slots, 4 fixed PDCHs and CS-2. The radio channel BLEP was set to 13.5 %, which corresponds to a mean C/I of 12 dB. To discover possible influences of the total system load on the results, simulations with 2 MSs and 5 MSs have been performed in the second step.

The ARQ protocol executed by the GPRS RLC protocol performs *PENDING ACK* retransmissions (see Section 3.4.5.5.3). Once an RLC block is transmitted for the first time it is marked as *PENDING ACK* by the sending entity. Assume there is no new (i.e., not yet sent) data, this block is retransmitted the next time the sending RLC instance is polled.

This behavior is intended to reduce the transmission delay, because it increases chances that the radio block is received correctly without the necessity of waiting for the NACK message by the peer instance. On the other hand logical connections have more radio resources assigned than it really needs to handle its given load. Connections under high traffic load will almost anytime have new RLC blocks to send, which practically disables retransmissions of *PENDING ACK* blocks.

This causes a load-dependent relation between the RLC block arrival process and the service process.

7.2.2.2 Expectations to the Service Process

Regarding an error-free radio channel and in a first step neglecting some details related to the GPRS multiframe structure, a deterministic *Probability Density Function* (PDF) would be expected, since the service time is identical to the size of an RLC block divided by the transmission rate at the air interface.

If the radio block can be disturbed during transmission over a radio channel, the service time distribution is expected to become an n-point distribution, each point representing the event that the radio block is disturbed a certain number of times (including zero) before it is received correctly.

Additionally, the height of the n-th peak (corresponding to the probability of the event that the radio block is disturbed n-1 times and not disturbed the n-th time it is transmitted) is expected to be lower than the probability of its predecessor (radio block is disturbed n-2 times and correctly received the n-1st time), since the error probability is assumed to be constant every time a transmission is performed. Figure 7.8 summarizes these considerations.

7.2.2.2.1 Expected Minimum RLC Service Time

In addition to the duration of one radio block period (18.46 ms) we have to consider the processing offset assumed to 9.23 ms or two TDMA frames in the GPRSim (see Figure 7.9) that occurs similarly in real GPRS implementations. Thus, a minimum service time of 27.67 ms is expected.

These considerations are based partly on implementation details specific to the GPRSim, which means that in a real GPRS network small differences e.g. in the processing delay are possible.

7.2.2.2.2 Expected Service Time for Retransmitted RLC Blocks

To estimate the time needed for the service of an RLC block that is disturbed during its first transmission attempt, we assume that the MS is capable of transmitting NACK messages in every uplink radio block period following a downlink radio block period that is used for downlink transmission to the regarded MS.

In a real GPRS system, the MS is polled for acknowledgement messages by the BS. If the BS wants the MS to transmit an RLC acknowledgement message, it sets a poll bit in the RLC block header of a downlink radio block. This implicitly assigns the uplink PDCH with the same

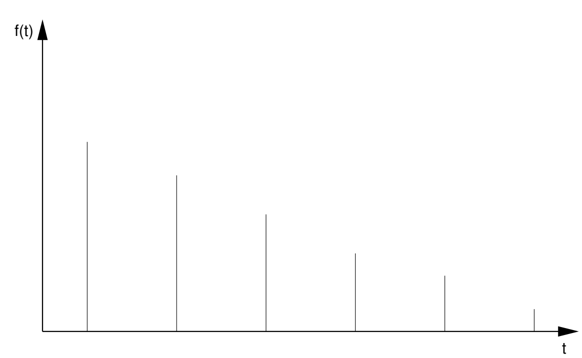


Figure 7.8: The distribution of the RLC service time expected to be an n-point distribution

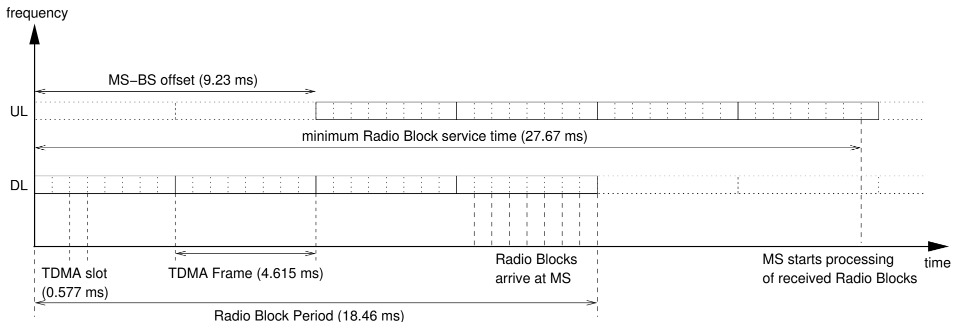


Figure 7.9: Successful RLC block transmission at the first attempt

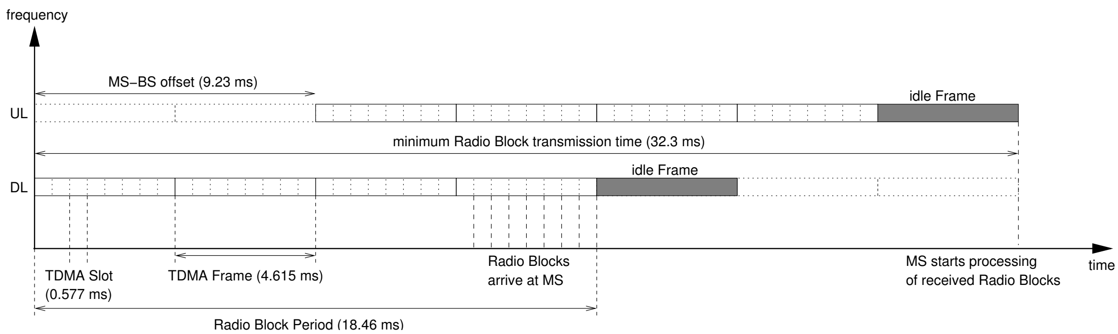


Figure 7.10: RLC block transmission involving one idle frame

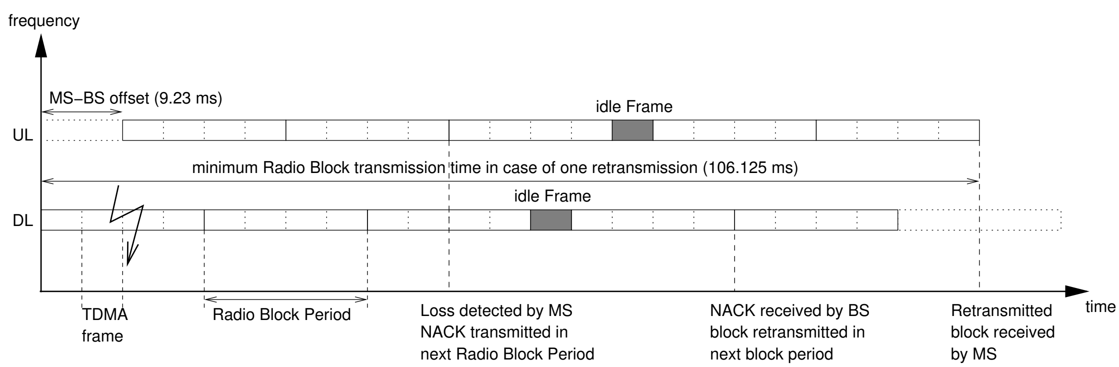


Figure 7.11: Retransmission of the first radio block in a radio block sequence

slot number during the next uplink radio block period to the receiving MS for transmission of the acknowledgement message.

Additionally, we assume that the BSS transmits several RLC blocks in a sequence and only the first one is disturbed, which is indicated by the flash in Figure 7.11. The MS detects the missing RLC block after reception of the second RLC block. This causes a NACK message to be transmitted in the next uplink radio block period. The BS receives the NACK message at the beginning of the next downlink radio block period after reception of the NACK message. Four downlink radio block periods have passed since the beginning of the first RLC block service time. The RLC block service time for the regarded block up to this instant accumulates to 78.455 ms (4 radio block periods plus one TDMA frame), because in a sequence of 4 radio block periods at least one idle frame duration has to be added (see Section 3.3.3).

The requested RLC block is retransmitted immediately and received correctly at this time. The time needed for the second transmission equals the duration for an undisturbed transmission, which above was calculated to 27.67 ms. Thus, we expect a total RLC service for this block time of 106.125 ms.

With an increasing number of active MSs in the regarded system, only little change of the evaluated service process is expected, since the queueing delay of RLC blocks that increases significantly with increasing load is not part of the RLC block service time.

7.2.2.3 Results and Discussion

As shown in Figure 7.12(a) the expected result qualitatively matches the simulation results. The service duration in case of successful transmission at the first attempt is either 27 ms or 31 ms (see Section 7.2.2.2.1). The next peak after the gap is a double-peak at 105 ms and 110 ms, representing the event of one erroneous transmission during an RLC block service time (see Section 7.2.2.2.2).

Further characteristics can be notified and explained:

- Two types of peaks are occurring, which are call double-peak and single-peak, respectively.

  The existence of double-peaks can be explained with GPRS multiframe structure and considering some implementation details.

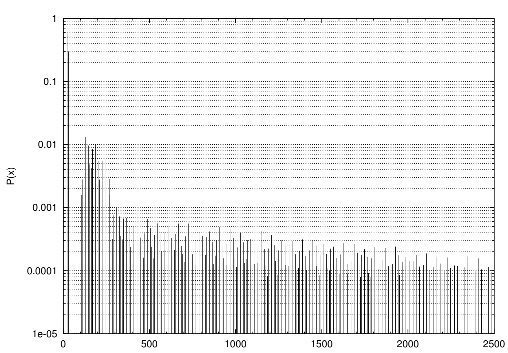
  A multiframe consists of 52 TDMA frames, i.e. 4 *idle* frames and 48 frames that can carry GPRS radio blocks (see Section 3.3.1). The latter are split into 12 groups of four, each group is called a *radio block period*. After every third radio block period an idle frame is transmitted.

  At the beginning of each radio block period, the radio blocks received during the previous radio block period are handed to the RLC layer (see Figure 7.10). This causes the service time to be increased by the duration of one TDMA frame in case the radio block period utilized for transmission is followed by an idle frame. Thus, every event represented by a double–peak can take place involving a different number of idle frames, and every event that is only possible involving a fixed number of idle frames is represented by a single peak.

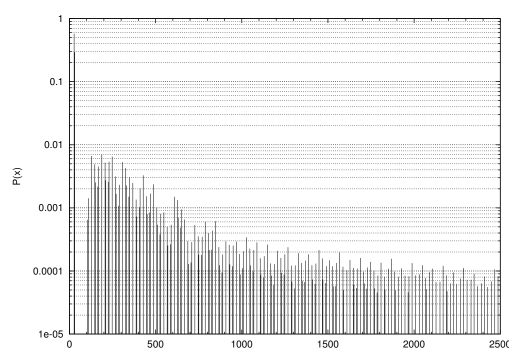
  The gap between the two peaks of a double peak is representing the duration of the additionally involved idle frame, which is 4.615 ms.

- Two double–peaks are followed by a single peak. Since after three radio block periods an idle frame is transmitted, a constant number of idle frames are involved every third event, when the number of radio block periods needed for successful transmission is a multiple of three.

In Figure 7.12(b), the simulation was performed with 5 MSs. Besides the fact that the upper bound for the RLC service time increases, the service process does not change significantly with an increasing number of MS.



(a) 4 fixed PDCHs, 1 MS



(b) 4 fixed PDCHs, 5 MSs

Figure 7.12: RLC service time

7.2.3 Modeling as M/M/n System

After evaluation of the characteristic arrival and service time distributions in GPRS systems by simulation, the mean value of these distribution is now taken for the analysis of an M/M/n queueing system.

Both arrival process and service process of an M/M/n queueing system are modeled by Poisson processes, which only depict first order characteristics. The queue is assumed to be infinite. The number of servers is set to the number of PDCHs in the simulation scenario regarded for comparison. Comparison with the simulated system is done for a scenario with 4 fixed PDCHs ($n = 4$).

The quantities of interest are comprehensively derived in [Kleinrock (1975)]. Let $E[\tau_b]$ be the mean service time and $E[\tau_a]$ be the mean inter-arrival time. Then the *offered traffic* is

$$a = \frac{E[\tau_b]}{E[\tau_a]}$$

and the system load is

$$\rho = \frac{a}{n} = \frac{E[\tau_b]}{E[\tau_a] \cdot n}\,; \quad \rho \leq 1.$$

The probability of an arriving job finding all servers busy (i.e., the waiting probability) is

$$p_w = \frac{\frac{a^n}{n!}\frac{1}{1-\rho}}{\sum\limits_{i=0}^{n-1}\frac{a^i}{i!} + \frac{a^n}{n!}\frac{1}{1-\rho}}.$$

The mean queue length is

$$\Omega = p_w \cdot \frac{1}{1-\rho}.$$

With knowledge of Ω and with the mean arrival rate λ the mean waiting time can be calculated using Little's Law:

$$E[\tau_w] = \frac{\Omega}{\lambda}.$$

Finally, the mean sojourn time is calculated by adding the mean service time to the mean waiting time:

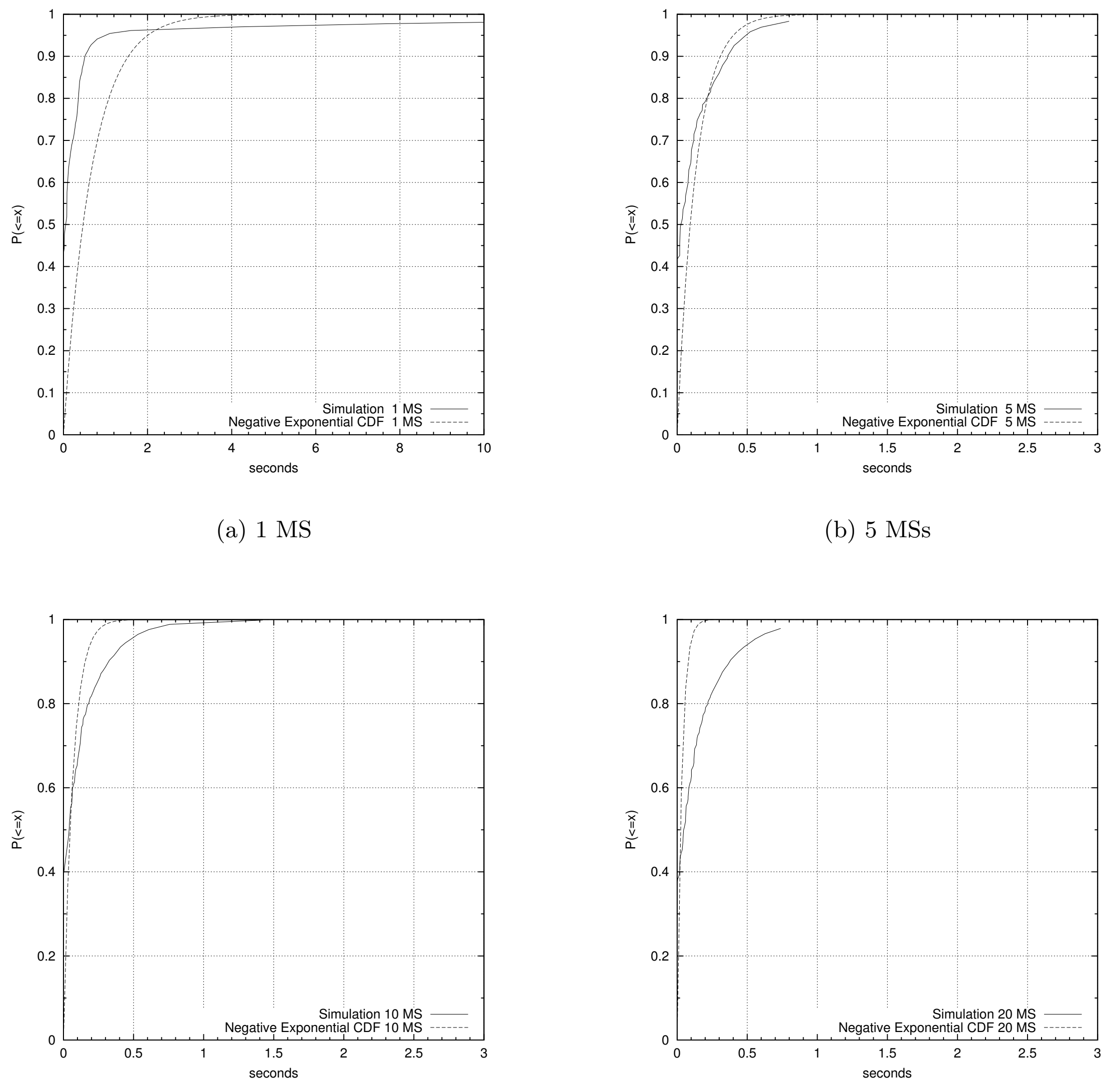
$$E[\tau_d] = E[\tau_w] + E[B] = E[\tau_w] + \frac{1}{\mu}. \tag{7.28}$$

Based on the evaluation of the IP datagram arrival process (see Section 7.2.1), the arrival rate is chosen to the reciprocal value of the mean IP datagram inter-arrival time, evaluated in a scenario with only one active MS. To illustrate the accuracy of this approximation in Figure 7.13(a) the CDF of the IP datagram inter-arrival time evaluated in this scenario is compared with a negative exponential CDF with same intensity. The Poisson approximation neither depicts that the probability for zero inter-arrival times is about 45%, nor the heavy-tail characteristic, which is visible in the CDF evaluated by simulation.

To consider increasing offered traffic with an increasing number of subscribers modeled in the system, the related number of superimposed Poisson processes is applied and compared with the IP inter-arrival time CDF in a corresponding simulation scenario (i.e., a scenario which comprises the same number of MS, see Figure 7.13(b), Figure 7.13(c) and Figure 7.13(d)).

Table 7.2: M/M/4 service process parameters

Mean IP datagram size	340 byte
RLC Block payload (CS-2)	29 byte
Mean number of RLC Blocks per IP datagram	12
Mean RLC Block service time	105 ms
Mean IP datagram service time $\overline{\tau_s}$	1.26 s



(a) 1 MS (b) 5 MSs

(c) 10 MSs (d) 20 MSs

Figure 7.13: IP datagram inter-arrival (simulation vs. M/M/4 model)

Although the heavy-tailedness of the sum arrival process becomes less significant with an increasing number of MSs, the Poisson approximation still fails to depict important characteristics of the real system beyond the mean value.

The mean service time is based on the evaluation of the mean IP datagram size, $\overline{S_{\mathrm{IP}}}$. The calculations of this section are based on the mean IP datagram size of 340 byte evaluated by simulation (see Table 7.2). Dividing by the number of bytes fitting into one RLC block (under the influence of the applied coding scheme), the mean number of RLC blocks needed to transfer an IP datagram, $\overline{N_{\mathrm{RLC}}}$, is derived. CS-2 was assumed, which means that 29 bytes of payload are transported within each RLC block. Assuming a multislot capability of 4 downlink blocks per radio block period, the number of servers, n, was set to 4. The mean number of RLC blocks per IP datagram was multiplied with the mean RLC block transmission delay, which was evaluated to 105 ms (see Section 7.2.2). This results in the mean IP service time:

$$\overline{\tau_s} = \frac{1}{\mu} = \frac{\overline{N_{ip}}}{\overline{N_{rlc}}} \cdot \overline{t_{rlc}}$$

In Table 7.3 the results for the M/M/n modeling approach are summarized. The arrival rate λ characterizing the arrival Poisson process is calculated by taking the reciprocal value of the mean IP inter-arrival time from the 1 MS scenario and multiplying with the number of MSs to be modeled. In the second column, $\overline{\tau_a}$, the mean inter-arrival time of the superimposed Poisson process for the corresponding number of MS is shown, while the third column represents $\overline{t_a}$, the real IP inter-arrival time for the same scenario evaluated by simulation with the GPRSim. The fifth column shows the system load.

The last two columns compare the mean sojourn time calculated with Equation (7.28) with the mean IP datagram delay evaluated by simulation. While the mean IP datagram delay evaluated by simulation does not exceed 3 s with 5 active MS, the M/M/4 model predicts a delay of more than 6 s already with 2 active stations. Regarding the system load ρ, shown in the 5th column of Table 7.3, it is shown that the model is in an overload situation for already 3 MS. Thus, the M/M/4 model does only deliver results for scenarios with very low traffic load.

As conclusion, the M/M/4 model overestimates the mean IP datagram delay by far (compare to simulated value $\overline{t_{\mathrm{IP}}}$), which indicates that an M/M/n model is not suitable for GPRS performance analysis regarding Internet traffic. As stated in general in [Paxson and Floyd (1995)] simple Poisson process modeling fails for Internet traffic in GPRS networks, since Poisson processes cannot represent inherent correlation structures or effects like long range dependence.

Table 7.3: M/M/4 result summary

MS	$\overline{\tau_a} = \frac{1}{\lambda}$ [s]	$\overline{t_a}$ [s]	$\rho = \frac{a}{n}$	$E[\tau_d]$ [s]	$\overline{t_{\mathrm{IP}}}$ [s]
1	0.669	0.669	0.470	**1.477**	0.913
2	0.335	0.314	0.942	**6.341**	1.157
3	0.223	0.216	1.413	—	1.264
5	0.134	0.136	2.350	—	2.268
10	0.067	0.107	4.701	—	10.975
20	0.033	0.127	9.545	—	41.521

7.2.4 Markovian Arrival Process with Marked Arrivals (MMAP)

In [Vornefeld (2002a,b)] an analytical modeling framework for Internet traffic in cellular packet radio networks is proposed. Numerical algorithms are used to derive an analytically tractable representation of Choi's Web browsing traffic model [Choi and Limb (1999)] (see Section 5.1.2) in terms of MMAP. The behavior of the Web model is divided in active (ON) and passive (OFF) phases and the sojourn times in these states are characterized by PH-type distributions. During the ON phase, Poisson packet arrivals are assumed.

The MMAP introduced by [He (1996)] is a stochastic process that combines inherent correlation structures with analytical tractability. It extends the concept of the *Markovian Arrival Process* [Neuts (1979)] by allowing to assign an individual service time distribution to each arrival event, called *marking* an arrival. Thus, it is possible to account for individual conditions of radio link and resource allocation strategies regarding multiple users.

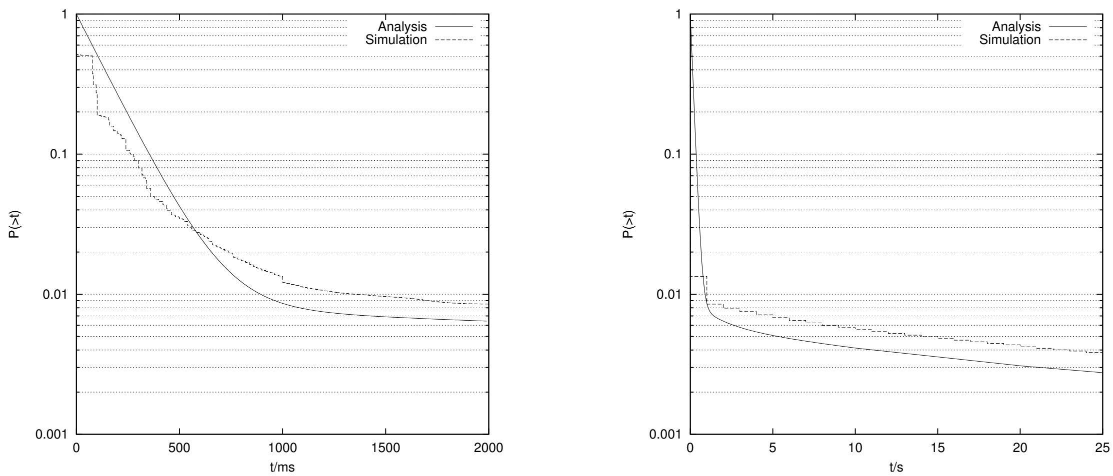
7.2.4.1 MMAP Representation of WWW Traffic

To be analytically tractable, a source description has to consist of exponential phases only. In order to derive such a description Vornefeld fits a PH-distribution to a simulated ON phase distribution, using the EM algorithm of [Asmussen et al. (1996)]. For generating the input distributions of this algorithm, Vornefeld has implemented Choi's application model (see Section 5.1.2) in a simulation environment. Since the duration of the OFF phase of Choi's model is described by a heavy-tailed Weibull distribution, which is available in closed form, the OFF phase duration's Weibull distribution can be approximated by a Hyper-exponential distribution. To achieve this, Vornefeld uses an extended version of the algorithm proposed by Feldmann and Whitt [Feldmann and Whitt

(1998)]. For a comprehensive description of the procedure see [Vornefeld (2002a),Vornefeld (2002b)].

The arrival events of the resulting stochastic process are represented by the arrival of IP datagrams with maximum size, which is assumed to be 536 bytes of IP payload and 40 bytes of TCP and IP header, 576 bytes in total.

Comparing the IP datagram inter-arrival time resulting from Vornefeld's analytical representation of Choi's WWW model with the inter-arrival time evaluated by simulation with the GPRSim shown in Figure 7.14, a good similarity can be observed between source characteristics in simulation and analysis, respectively.



(a) Lower range (b) Tail

Figure 7.14: CCDF of IP datagram interarrival time after Choi's model (simulation vs. analysis in [Vornefeld (2002a)])

7.2.4.2 Service Process

Vornefeld uses link level simulations to determine the *Laplace-Stieltjes-Transform* (LST) of the service time distribution [Vornefeld (2002a)]. His simulation of the GPRS radio channel comprises a non-frequency selective radio channel with slow and fast fading and Doppler shifts.

After the arrival of an IP datagram of 576 bytes in size at the top of the GPRS protocol stack, protocol overhead and segmentation into RLC blocks is modeled. Thus, an IP datagram arrival leads to a batch arrival of RLC blocks on RLC level. The coding scheme determines the size of these batches by the number of payload bytes an RLC block can carry (see Section 3.4.5.6).

Vornefeld determines the n-point distribution describing the number of transmissions needed for successful transmission of all RLC blocks carrying one IP datagram. In a next step he approximates the obtained distribution by a continuous PH-type distribution, using the EM algorithm [Asmussen et al. (1996)].

7.2.4.3 MMAP/G/1 Results and Comparison with Simulation Results

The results in [Vornefeld (2002a)] are compared with simulation results gained with the GPRSim. The IP datagram delay is chosen as the performance measure of interest. The analytical results regarded here can be found in [Vornefeld (2002a)].

In Figure 7.15 the mean IP datagram delay is displayed over the mean C/I ratio for a scenario comprising one WWW user employing CS-2 and using 2 PDCHs and 4 PDCHs, respectively. In

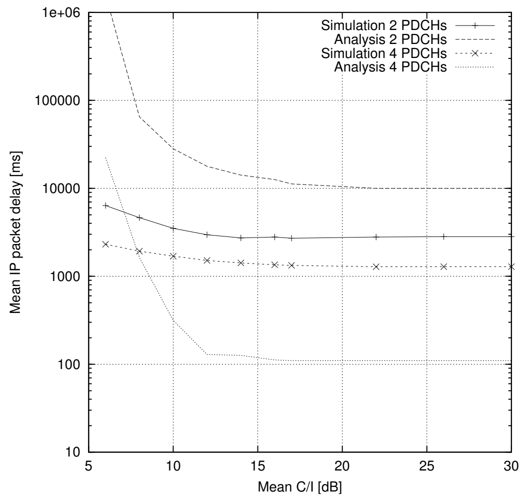


Figure 7.15: Mean IP datagram delay vs. mean C/I (1 MS, analysis see [Vornefeld (2002a)])

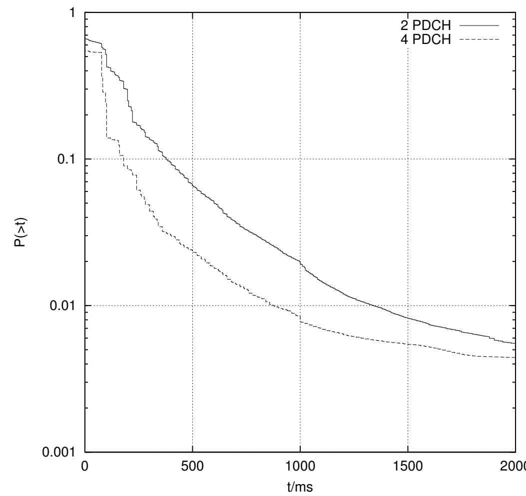


Figure 7.16: CCDF of IP datagram inter-arrival time (simulation with 1 MS, 30 dB C/I)

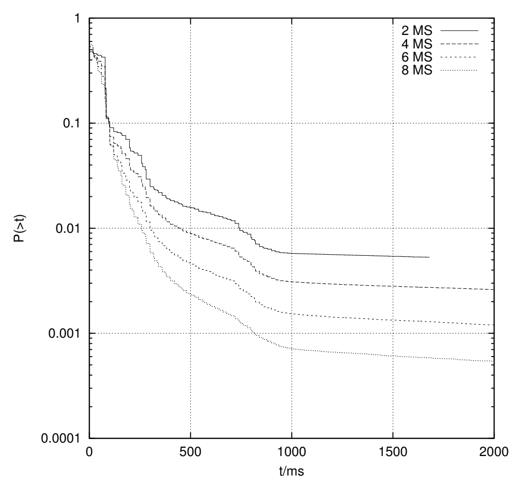


Figure 7.17: CCDF of IP datagram inter-arrival time (simulation with 8 PDCHs, 30 dB C/I)

both scenarios the downlink multislot capability is set equal to the number of available PDCHs (2 and 4 downlink slots).

In the case of negligible BLER (represented by the high C/I values at the right part of Figure 7.15) the impact of the system capacity has different effects in simulation and analysis, respectively. While the analysis predicts the mean IP datagram delay to rise two orders of magnitude when the system capacity is reduced from 4 to 2 PDCHs, simulation predicts only a doubled mean IP datagram delay.

This behavior is caused by the TCP protocol, which is considered in the simulation model (see Section 6.2), but not in the analytical model. TCP performs flow control to avoid network congestion [Stevens (1996)]. This means that on IP level the offered traffic is limited when the system capacity is reduced. As a result, in the simulation model a so-called elastic source behavior [Roberts (2001)] is resulting, while in the analysis we have an inelastic traffic source. This effect is visualized in Figure 7.16. Without considering packet loss at the air interface the reduction from 4 to 2 PDCHs causes higher IP datagram inter-arrival times. Thus, in the same time interval the mean number of datagram arrivals decreases when the system capacity is reduced. A reduced offered traffic leads to a better delay performance and thus represents the reason for the different effects of variations in system capacity [Irnich and Stuckmann (2002)].

In the analytical model the traffic is not elastic, since the traffic load is determined by the poissonian datagram arrival process in the ON phase and not by the underlying system performance or the available capacity. This leads to a delay performance difference in the analytical model regarding the cell scenario with 4 PDCHs, where less traffic than the available channel data rate is offered compared to the cell scenario with 2 PDCHs that represents an overload condition during the ON phase (see Figure 7.15). Since the simulation model represents TCP-based traffic that is bursty in nature due to the flow control mechanisms, the mean IP datagram delay in the 4 PDCH scenario of the simulation model is higher than in the analytical model. While the IP datagrams generated by a Poisson traffic source can be served immediately by the channel with relatively high capacity, batch arrivals generated by the TCP source lead to higher waiting times. Once the available channel capacity becomes smaller than the offered traffic during the ON phase like in the 2 PDCH scenario the delay increases dramatically in the analytical model, while the TCP source is adapting the offered traffic to the available channel capacity leading to a lower offered traffic with shorter queue lengths and lower delays than in the analytical model.

Evaluating the effect of the mean C/I ratio to the mean IP datagram delay, which is visible in the left part of Figure 7.15, we see that the simulation model is less sensitive to a degradation of channel quality. Again, this is caused by TCP flow control. An increasing BLER leads to performance degradation, because additional time needed for retransmission adds to the IP delay. As a reaction TCP reduces the offered traffic, which partly compensates the performance degradation.

Additionally, the *PENDING ACK* feature (see Section 3.4.5.5.3) deployed by the GPRS RLC layer has some effect in situations under bad channel condition. When an RCL block is transmitted for the first time, it is marked as *PENDING ACK* at the transmitting entity. The next time radio resources are allocated to this MS, the radio block is transmitted again, provided that no new data has arrived meanwhile. Thus, the receiving entity does not have to discover the loss of radio blocks and request the radio block for retransmission in case the first transmission attempt was not successful.

In Figure 7.15 the effect of elastic traffic is illustrated. For a scenario with WWW users, 4 PDCHs, multislot capability 4, CS-2 and an error-free radio channel the number of users is increased from one to nine. When system load increases, the traffic load of a single user is gradually decreased (see Figure 7.17), which results in a slower increase of the total system load, compared to linear superposition of multiple users.

Traffic Performance of GPRS and EGPRS

This chapter presents a comprehensive traffic performance analysis based on the GPRSIM introduced in Chapter 6 for typical GPRS and EGPRS traffic sources like WWW and e-mail regarding different cell configurations, system parameters and radio channel conditions. Both fixed and on-demand PDCHs are considered. To compare the packet data services GPRS and EGPRS the effect of the new modulation and coding schemes and the *Link Quality Control* (LQC) functions *Link Adaptation* (LA) and *Incremental Redundancy* (IR) is outlined.

8.1 General Simulation Parameter Settings

The cell configuration is defined by the number of *Packet Data Channels* (PDCHs) available permanently or on demand for GPRS or EGPRS.

An RLC/MAC block error rate of 13.5% has been assumed throughout the most GPRS simulations corresponding to a C/I of 12 dB. This parameter value is in accordance with a typical radio coverage planning goal, widely used by operators for GSM networks. CS-2 is used as the coding scheme for user data, which is a relatively robust coding scheme. The Gilbert-Elliot channel model, as described in Section 6.9.1.3, has been applied.

In EGPRS simulations the channel conditions are determined by the cell and cluster sizes that are the basis for the C/I calculation as described in Section 6.9.3. Cluster sizes 3 and 7, cell sizes with a radius of 300 and 3000 meters and a velocity of 100 km/h and 6 km/h, have been regarded. The Brownian motion mobility model is used in the EGPRS simulations modeling an urban environment (see Section 6.9.2.2), while the Secant mobility model is applied for rural environments (see Section 6.9.2.1). The Brownian motion mobility model is parameterized so that each mobile station changes its direction every 15 seconds on average.

Mobile stations with the multislot capability of 1 in uplink and 2, 3 and 4 in downlink are discussed. LLC and RLC/MAC are operating in acknowledged mode. The MAC protocol instances in the simulations are operating with three random access subchannels per 52-multiframe. For the simulations without QoS management, all conventional MAC requests have the radio priority level 1 and are scheduled using a *Round Robin* (RR) strategy with a quantum of 10 radio blocks per TBF per RR cycle. LLC has a window size of 16 frames. TCP/IP header compression in SNDCP is performed. The TCP *Maximum Segment Size* (MSS) is set to 256 byte for WAP and 512 byte for TCP-based applications. In the Internet stack for WWW and e-mail, TCP is operating with a maximum congestion window size of 8 kbyte. The transmission delay in the core network and external networks, i.e., the public Internet, is neglected, as it is assumed that the servers are located in the operator's domain and the core network is well dimensioned. Since the high round-trip time in GPRS networks is mainly caused by *Temporary Block Flow* (TBF) establishment procedures at the air interface, the delay in well-dimensioned IP subnetworks does not have a significant effect on the end-to-end performance [Stuckmann et al. (2002)].

In order to perform network capacity planning the system behavior during the *busy hour* has to be considered. The traffic load generated is characterized by the number of Internet sessions running or by the offered traffic measured in kbit/s per radio cell. The traffic characteristics are related to the traffic models introduced in Chapter 5. While for WWW sessions the adapted Mosaic model is used (see Section 5.1.1), e-mail, WAP and Streaming sessions are characterized by the models in Section 5.2.2, Section 5.4.4 and Section 5.5.1.2, respectively. The inactive period between two sessions is negative exponentially distributed for all kinds of (E)GPRS sessions with the mean value of 60 s. In the following, mobile stations running (E)GPRS sessions with this small inter-arrival time are called 'active stations'.

The parameter settings are summarized in Table 8.1. Parameters of the following scenarios, that differ from these settings, will be explicitly defined at the beginning of the related sections.

Table 8.1: Simulation Parameter Settings

Number of fixed PDCHs	8 (7, 6, 5, 4, 0)
Number of on-demand PDCHs	0 (8)
Multislot capability (DL/UL)	4/1 (3/1, 2/1)
Coding Scheme	CS-2
Mean C/I [dB]	12 (9, 16, 20)
Cluster Size	3 (7)
Cell radius [m]	300 (3000)
MS velocity [km/h]	6 (100)
TCP/IP header compression	activated
TCP MSS [byte]	256 (WAP), 512 (TCP)
TCP version	Reno
TCP maximum window size [kbyte]	8
HTTP version	1.1
Traffic mix WWW/e-mail	30%/70%
Traffic mix WAP/WWW/e-mail	60%/12%/28%
Traffic mix Streaming/WWW/e-mail	10%/27%/63%

8.2 Performance and System Measures

To characterize the traffic performance of GPRS several performance and system measures are defined in the following. For different applications different critical performance measures will be regarded, since different performance characteristics are required for download-oriented, transaction-oriented and real-time applications, respectively. While for the download-oriented and transaction-oriented applications mean values represent the user-perceived performance [Kalden et al. (2000)], variance and distribution functions should be regarded for real-time applications. Since the performance of these applications is limited by the downlink capacity, only downlink measures are regarded.

IP throughput per user is the IP throughput measured during transmission periods, e.g., the transmission period of a single object of a web page. One transmission period is defined by the period between the user request and the complete reception of the file. The throughput measure is evaluated statistically by counting the amount of IP bytes for each transmission period. Dividing by the duration of the transmission period, the throughput for this period is calculated after each transmission period. In order to calculate a statistical measure that is time-weighted, the throughput value is put in the evaluation sequence n times, while n is the duration of the transmission period counted in TDMA frames. At the end of the simulation, the time-weighted mean value and the distribution function can be determined.

80-percentile of IP throughput per user is the throughput performance that can at least be offered to users in 80% of the cases (i.e. 80% of the measured throughput values are higher than this value). This measure corresponds to the 80-percentile of the throughput CCDF and to the 20-percentile of the throughput CDF.

Application response time is the difference between the time when a user is requesting a web page, a WAP deck or an e-mail and the time when it is completely received.

IP datagram delay is the end-to-end delay evaluated by means of time stamps given to the datagrams, when the IP layer generates a new IP datagram. When the datagram arrives at the receiver, the difference between the actual time and the time stamp value is calculated as a sample of the respective evaluation sequence.

Session blocking rate is the number of aborted sessions due to bad performance normalized to the total number of sessions.

IP system throughput per cell is calculated from the total IP data transmitted on all downlink PDCHs and for all users of the regarded radio cell during the whole simulation duration, divided by the simulation duration. Since a loss of IP datagrams over fixed subnetworks is not modeled, this parameter equals the offered IP traffic in the radio cell.

PDCH utilization is the number of MAC blocks utilized for MAC data and control blocks normalized to the sum of data, control and idle blocks. Existing capacity reserves in the scenario under consideration can be seen from this measure.

8.3 Simulation Results for GPRS

In GPRS networks a radio cell may allocate resources by means of one or several physical radio channels in order to support the GPRS traffic. The channels shared by the GPRS mobile stations for packet transmission are taken from the common pool of GSM physical channels available for speech and packet data communication in the radio cell. The allocation of physical channels to circuit-switched services and to the GPRS packet-switched service is done dynamically according to the "capacity on demand" principle, whereby the operator can decide to permanently or temporarily dedicate physical channels for the GPRS. In this context GSM physical channels that are allocated permanently for GPRS are called *fixed PDCHs*, whereas channels that are allocated temporarily for GPRS are called *on-demand PDCHs* (see Section 3.3.1).

In the following subsections simulation results for configurations with fixed and on-demand PDCHs regarding a traffic mix of WWW/e-mail (see Section 5.1.1, Section 5.2.2 and Section 6.2) are discussed. The probabilities for a WWW and an e-mail session are 30% and 70%, respectively.

8.3.1 Configurations with Fixed PDCHs

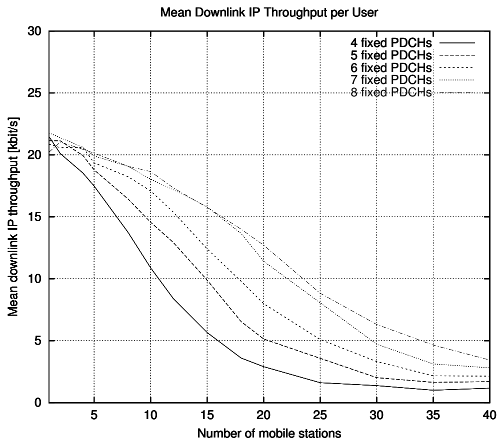
To give an overview on the traffic performance of Internet applications over GPRS, simulation results for configurations with a fixed number of PDCHs are presented in this section.

Figure 8.1, Figure 8.2 and Figure 8.3 show the performance and system measures over the number of mobile stations with the multislot capability of one uplink and four downlink slots for a different number of fixed PDCHs available for GPRS. The maximum mean downlink IP throughput per user during transmission is 22 kbit/s and decreases with the number of mobile stations (see Figure 8.1(a)). This maximum mean throughput is depending on the source traffic characteristics, especially the file size. As explained in Section 6.14.1 the TCP slow start algorithm limits the performance for small files.

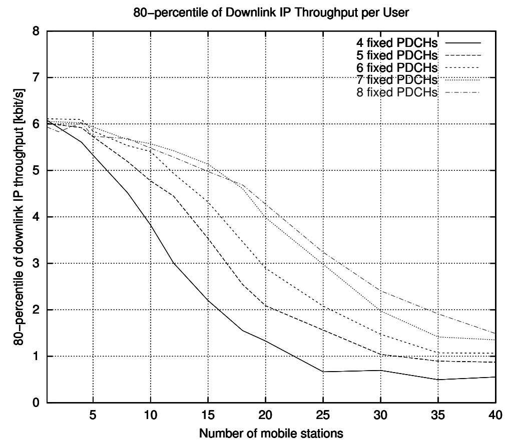
Under low traffic load, e.g., 1–10 MSs and 8 fixed PDCHs the performance decreases only slowly with increasing load, since the PDCHs are not completely utilized (less than 40%, see Figure 8.3(b)). With a lower number of PDCHs the performance degradation happens earlier, because less capacity is available. When the throughput is plotted over the offered traffic normalized to the number of fixed PDCHs (see Figure 8.1(c)), the throughput performance degradation in configurations with more fixed PDCHs is still happening slower with increasing load because of the economy of scale. Since the burstiness of the traffic decreases with growing capacity and constant normalized offered traffic, more traffic can be multiplexed per PDCH with a higher number of PDCHs.

Under medium traffic load, e.g., 10–20 MSs and 6 PDCHs the throughput decreases linearly, since the utilization passes 50% and the probability that the capacity has to be shared between active stations is increasing with an increasing utilization. Once the system reaches saturation, the throughput asymptotically approaches about 2–3 kbit/s with an increasing number of MSs, because the full capacity is utilized and shared between an increasing number of MSs.

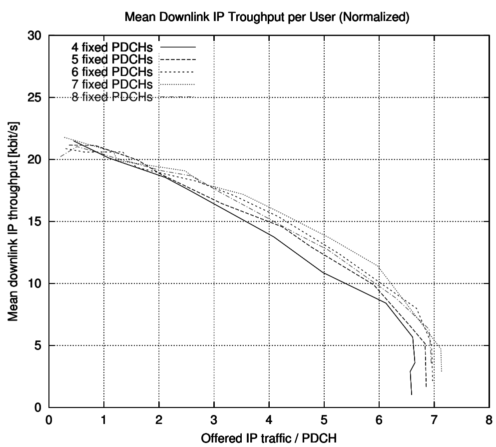
The 80-percentile of the downlink IP throughput per user (see Section 8.2) shows the same behavior with an increasing number of active stations (see Figure 8.1(b)). The 80-percentile under low traffic load is not exceeding 6 kbit/s because of the heavy-tailed file size distribution. While there are some very large objects with good throughput performance, the WWW traffic is dominated by small objects with low throughput performance (see Section 5.1.1). More than 20% of the objects belong to very small objects with poor throughput performance. This can also be seen in Figure 8.1(d), where the throughput distribution functions for 15 MSs with the number of fixed PDCHs as a parameter are shown. Here it can be seen that 50% of the throughput values are smaller than 10 kbit/s and 80% are smaller than 16 kbit/s for 8 fixed PDCHs. For a decreasing



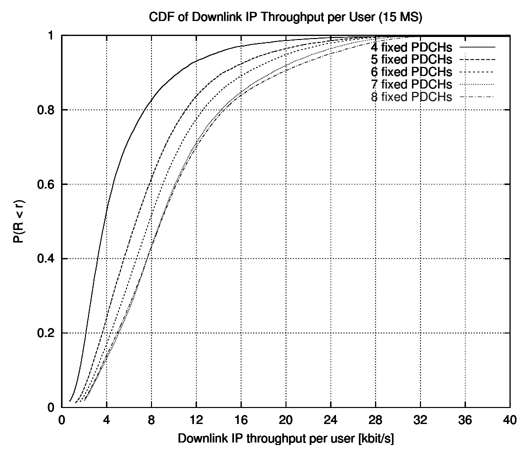
(a) Mean downlink IP throughput per user



(b) 80-percentile of downlink IP throughput per user



(c) Throughput over the normalized offered IP traffic



(d) CDF of downlink IP throughput per user (15 MSs)

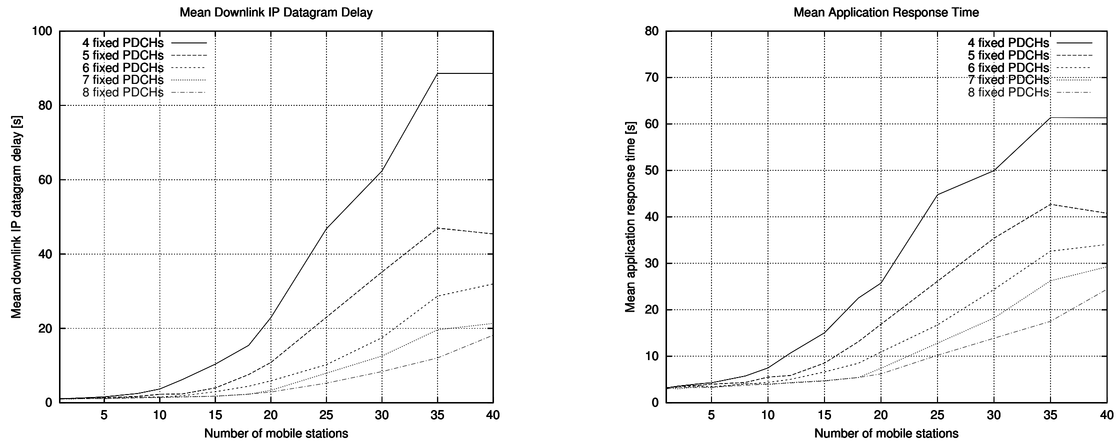
Figure 8.1: Throughput performance for different numbers of fixed PDCHs

number of fixed PDCHs the performance of larger objects decreases because of less available capacity so that there are less values with high throughput and more values in the medium range of 4–12 kbit/s.

The mean downlink IP datagram delay increases asymptotically until the system reaches saturation (see Figure 8.2(a)). Under saturation, TCP limits the offered traffic, which leads to slower increase in traffic load in spite of more MSs in the cell and to slower increase in datagram delay. This can also be seen in Figure 8.3(a) for 4 fixed PDCHs, where the downlink IP system throughput is decreasing, when the number of active stations is further increased after saturation has been reached.

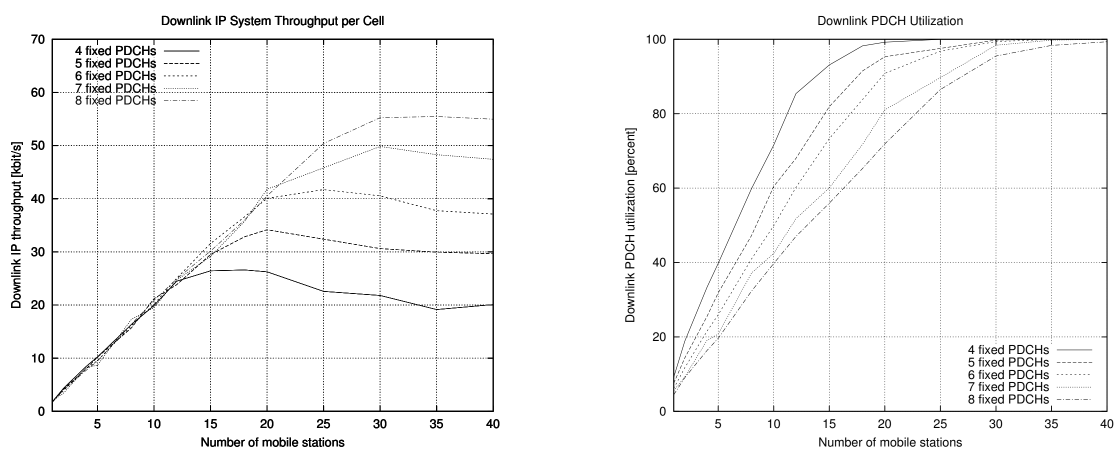
In general the high datagram delays compared to typical delays in fixed networks can be explained by the high TBF set up time (see Section 6.14.1.3) and long IP packet queues in the TCP steady state because of low bandwidth in GPRS networks.

The application response time as shown in Figure 8.2(b) increases similarly from a few seconds in situations with a small number of MSs per cell to unacceptable values of 30–50 seconds for 30 MSs per cell.



(a) Mean downlink IP datagram delay (b) Mean application response time

Figure 8.2: Delay performance for different numbers of fixed PDCHs



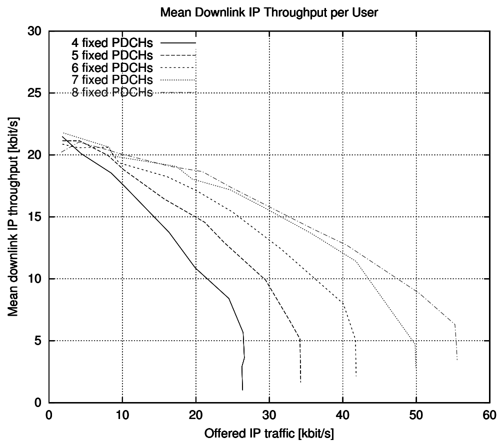
(a) Downlink IP system throughput per cell (b) Downlink PDCH utilization

Figure 8.3: System measures for different numbers of fixed PDCHs

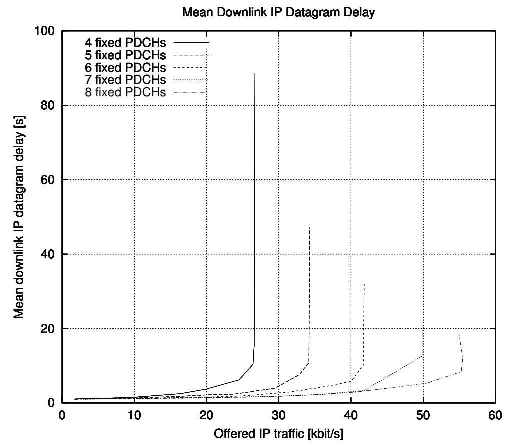
In Figure 8.3 the system measures, namely system throughput and PDCH utilization, are shown. With about 80% PDCH utilization the system throughput goes into saturation. At this point no further offered application traffic can be served, although some remaining capacity is available on average. This effect can be explained with the TCP protocol that is limiting the offered traffic on lower layers performing the congestion avoidance algorithm. The maximum system throughput per cell ranges from 20 to 56 kbit/s depending on the number of fixed PDCHs available.

The graphs in Figure 8.4 are based on the same simulations, but show performance parameters plotted over the offered IP traffic.

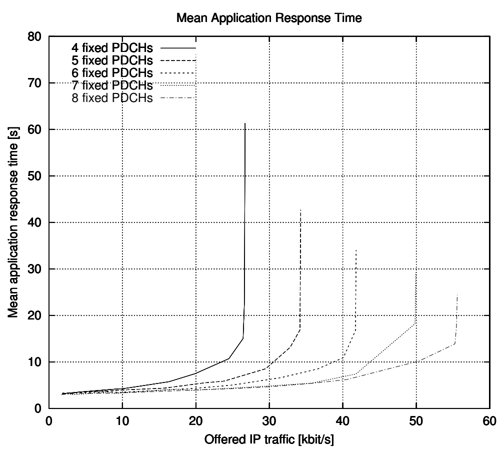
The downlink IP throughput per user during transmission decreases almost linearly with increasing offered IP traffic until saturation is reached (see Figure 8.4(a)). The offered traffic corresponding to saturation is different for different numbers of available fixed PDCHs. If the application traffic is further increased beyond saturation, TCP limits the offered IP traffic and sessions are aborted so that the evaluated IP throughput per user decreases only slightly for the same maximum



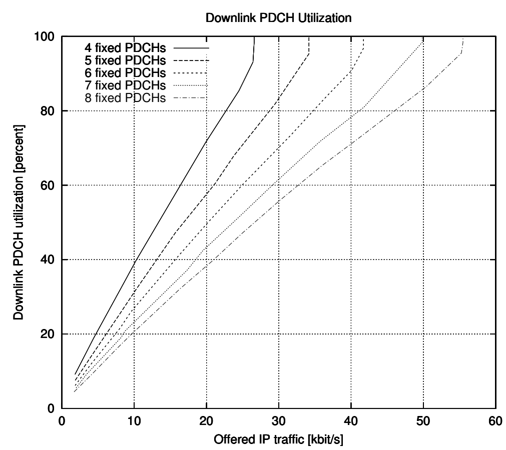
(a) Mean downlink IP throughput per user



(b) Mean downlink IP datagram delay



(c) Mean application response time



(d) Downlink PDCH utilization

Figure 8.4: Performance and system measures over the offered IP traffic

offered IP traffic.

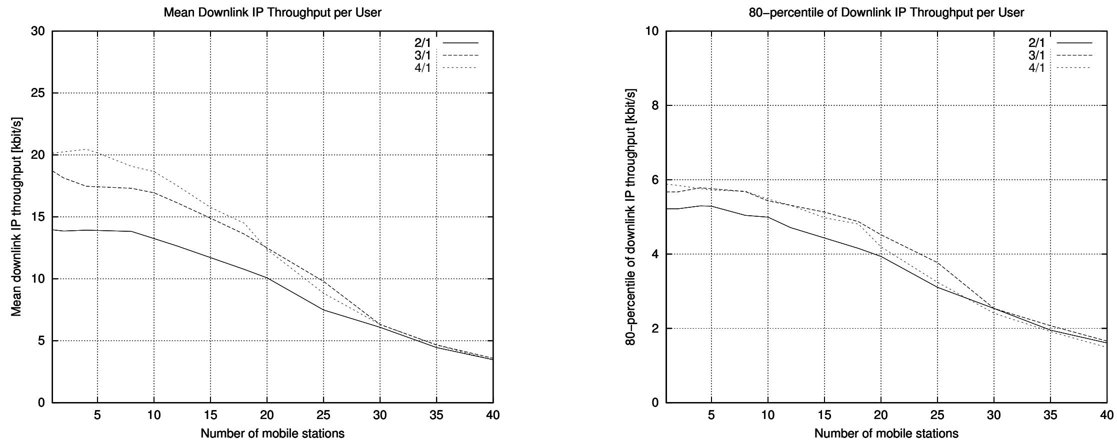
Both IP datagram delay and application response time are approaching a pole for increasing offered IP traffic (Figure 8.4(b) and Figure 8.4(c)). The pole corresponds to the maximum offered IP traffic that can be carried by the regarded number of fixed PDCHs.

The PDCH utilization increases linearly from 0 to 100%. The gradient corresponds to the system capacity represented by the number of fixed PDCHs available in the cell. The maximum offered traffic that is carried by the regarded PDCHs is reached with 90–95% PDCH utilization.

8.3.2 Effect of the Multislot Capability

The performance limitation of GPRS terminals is influenced by the multislot capabilities particularly in the downlink, since Internet applications are characterized by high downlink traffic. In this section, the effect of the downlink multislot capability on performance and system measures is presented.

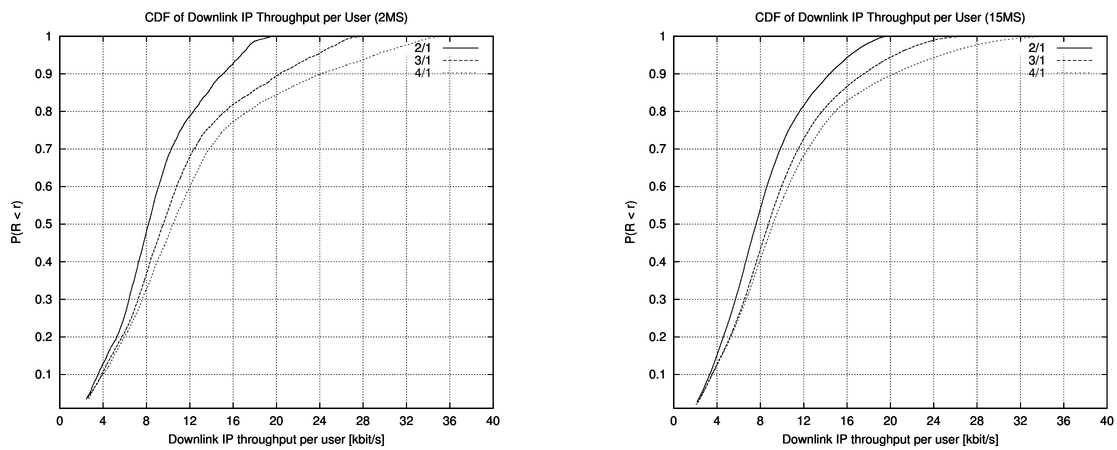
The throughput performance results shown in Figure 8.5 visualize the performance difference between the multislot capabilities of 2, 3 and 4 downlink slots. A cell configuration with 8 fixed PDCHs is assumed.



(a) Mean downlink IP throughput per user

(b) 80-percentile of downlink IP throughput per user

Figure 8.5: Throughput performance for different downlink multislot capabilities



(a) CDF of downlink IP throughput per user (2 MSs)

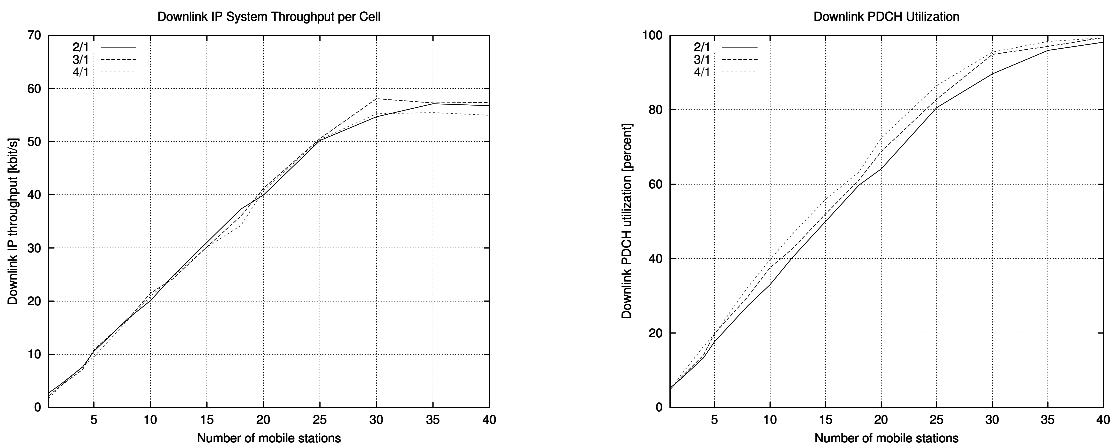
(b) CDF of downlink IP throughput per user (15 MSs)

Figure 8.6: Throughput distribution for different downlink multislot capabilities

The maximum mean downlink IP throughput per user with multislot capability 2 equals about 65% of the value reached with 4 downlink slots, while 85% is reached with 3 downlink slots (see Figure 8.5(a)). At medium traffic load (up to 15 MSs) a similar performance difference can be observed. Under high traffic load, however, equal performance for all multislot capabilities can be observed, because all PDCHs are utilized and the timeslots that can not be used by a mobile station due to multislot limitations can be used by another active station.

The performance difference for the downlink multisolt capablilities 2 to 4 measured as the 80-percentile of the downlink IP throughput as shown in Figure 8.5(b) is less significant, because the performance of small objects that mainly influence the 80-percentile is less sensitive to changing bandwidth. The performance of small objects is mainly influenced by the round-trip delay (see Section 6.14.1).

Figure 8.6 shows the CDFs of the downlink IP throughput per user for 2 and 15 MSs. In



(a) Downlink IP system throughput per cell

(b) Downlink PDCH utilization

Figure 8.7: System measures for different downlink multislot capabilities

the range of low throughput that is corresponding to small objects the throughput distribution functions for different multislot capabilities are similar, since the throughput for small objects is not significantly influenced by a small change in capacity. For larger objects that are corresponding to the range of higher throughput values the sensitivity to the multislot capability becomes visible. While the upper 10% of the throughput samples for 2 downlink slots are reaching 14 kbit/s, 10% of the throughput values for 3 and 4 downlink slots are exceeding 18 kbit/s and 20 kbit/s, respectively (see Figure 8.6(a)). For increasing system load (see Figure 8.6(b)) the throughput distribution functions are increasing faster in the range of medium and high throughput values, because less capacity is available per MS. The system throughput and the PDCH utilization show nearly no difference for various multislot capabilities (see Figure 8.7(a) and Figure 8.7(b)) even for low traffic load, because the offered traffic is not sensitive to small changes in available bandwidth, as long as the system is not congested.

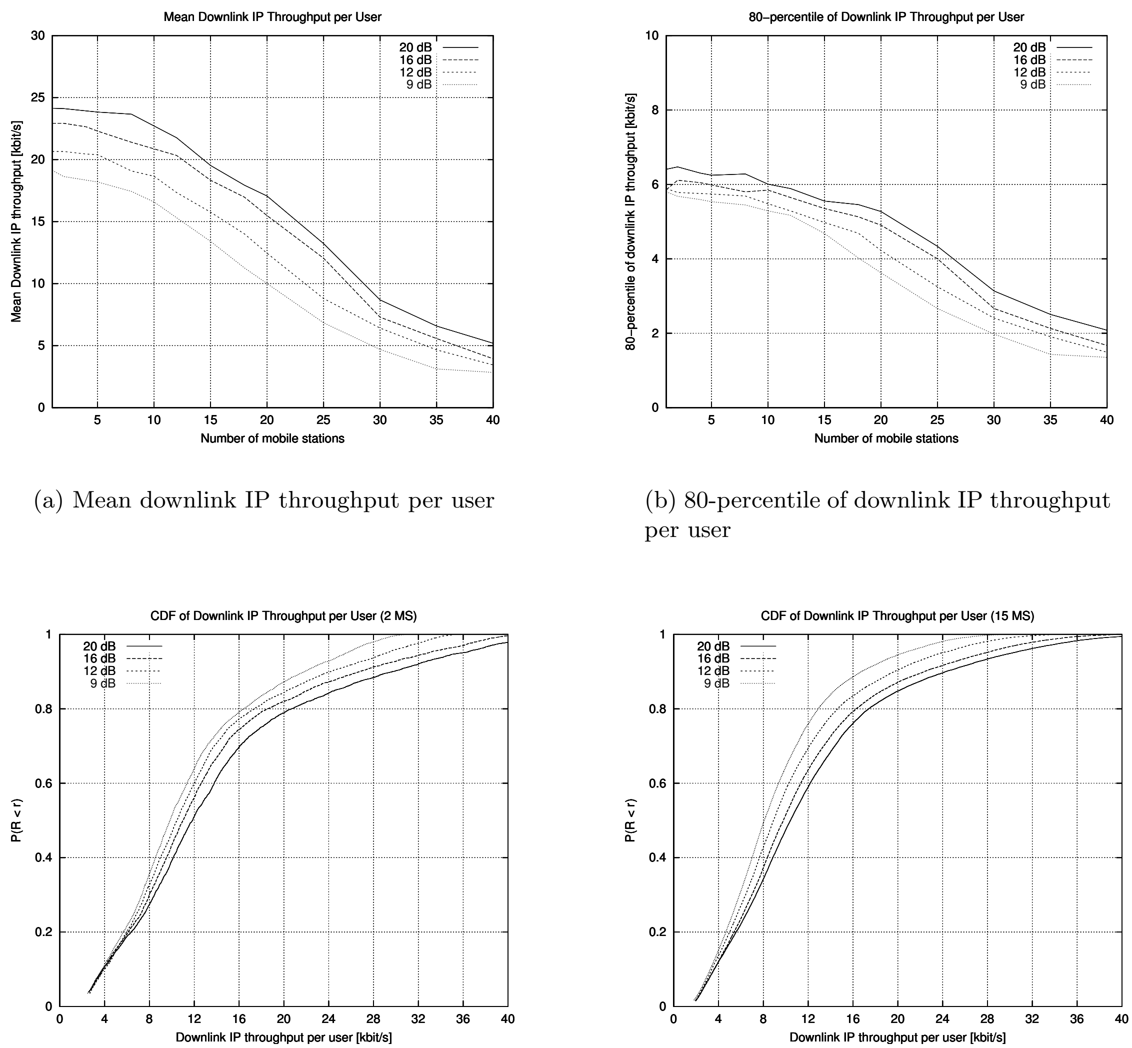
8.3.3 Effect of the Radio Channel Conditions

In modern GSM networks the radio coverage is planned with the C/I goal of 12 dB in 95% of the cell area to maintain high speech quality. Since an ARQ protocol is applied in GPRS to realize a reliable transmission between MS and BSS, the robustness that is necessary for data transmission can be ensured in most coverage areas for all coding schemes that are based on GMSK. Therefore only small GPRS performance differences are expected over the C/I range typical for existing GSM networks.

In this section the performance and system measures for different radio channel conditions are regarded. While a C/I of 12 dB is a typical value for GSM/GPRS environments representing a block error rate of 13.5% for CS-2, the C/I values 16 and 20 dB represent very good channel conditions. For a worst-case estimation the mean C/I value of 9 dB is regarded.

The mean downlink IP throughput per user and the related 80-percentile over the number of mobile stations for different C/I values as a parameter are shown in Figure 8.8(a) and Figure 8.8(b). The mean throughput performance for the scenarios with a mean C/I of 12 dB is only at maximum 15% less compared to the throughput at a C/I of 20 dB that corresponds to a scenario without any block errors. A maximum performance degradation of 10% compared to the 12 dB scenario as experienced with a mean C/I of 9 dB where a high block error rate of more than 25% is observed.

In Figure 8.8(c) and Figure 8.8(d) the CDFs of the downlink IP throughput per user are shown for 2 MSs and 15 MSs. The sensitivity to the mean C/I values becomes visible in the range of medium to high throughput corresponding to the percentage of users that transmit medium and large objects, because the performance of the related objects is sensitive in the change of available



(a) Mean downlink IP throughput per user

(b) 80-percentile of downlink IP throughput per user

(c) CDF of downlink IP throughput per user (2 MSs)

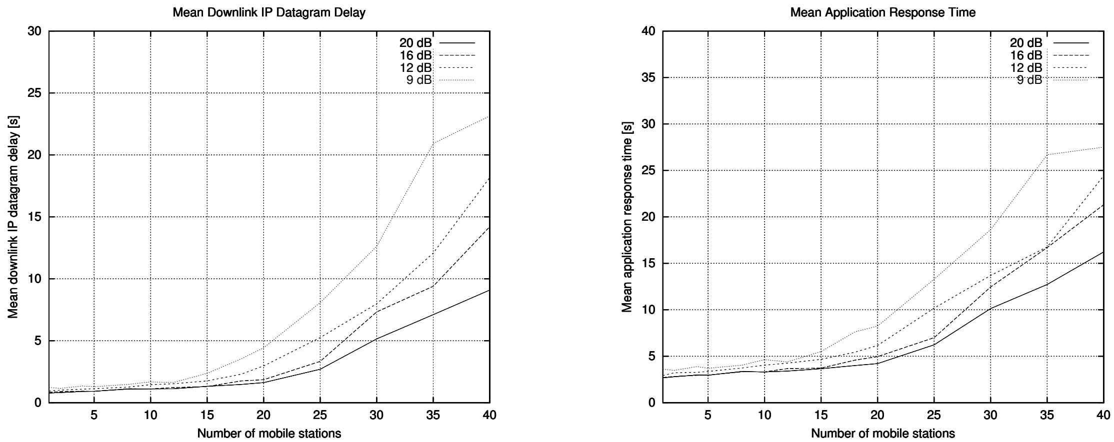
(d) CDF of downlink IP throughput per user (15 MSs)

Figure 8.8: Throughput performance for different channel conditions

capacity and therefore sensitive to block errors (see Figure 8.8(c)). In the load scenario with 15 MSs that represents a high load situation the sensitivity of the throughput distribution to C/I is slightly increasing because less capacity reserves are available and retransmissions owing to block errors are influencing the capacity available per MS (see Figure 8.8(d)).

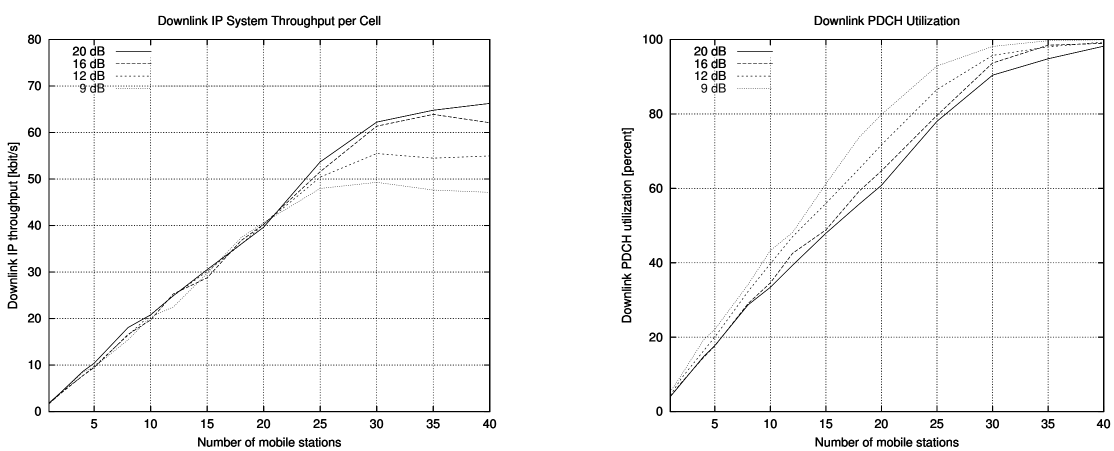
The delay performance shows a sensitivity to the radio channel conditions only under high traffic load. A delay gain of more than 30% is reached under 16 dB compared to 12 dB and the 12 dB delay is 30% higher than at 9 dB, if more than 15 stations are active in the radio cell (see Figure 8.9). The reason for this is the difference in utilization for the various C/I values of 15–20% resulting from retransmitted radio blocks (see Figure 8.10(b)). This effect only becomes visible in situations with higher traffic load, because under low traffic load, retransmissions are automatically performed by the RLC protocol, even if no errors are detected (see Section 3.4.5.5.3). Consequently, a C/I dependency of cell throughput (see Figure 8.10(a)) becomes visible only once the radio cell is highly utilized and approaches saturation. The maximum system throughput ranges from 50 to 65 kbit/s depending on the C/I.

This leads to the conclusion that the sensitivity of the GPRS performance to the C/I is only



(a) Mean downlink IP datagram delay (b) Mean application response time

Figure 8.9: Delay performance for different channel conditions



(a) Downlink IP system throughput per cell (b) Downlink PDCH utilization

Figure 8.10: System measures for different channel conditions

significant for scenarios that are outside the range of mean C/I values for typical radio coverage in GSM networks. The reason for the low sensitivity to a change in the C/I is the use of CS-2 that is a relatively robust coding scheme and the robust RLC scheme in GPRS that is efficient for both uncorrelated errors and error bursts up to a burst length in the order of the RLC window size. This expression only holds for optimized RLC implementations, where RLC acknowledge messages are transmitted regularly to avoid the RLC window stall condition (for an optimized RLC implementation see [Stuckmann (2002b), pp. 166–170]). Both the channel model with uncorrelated errors (see Section 6.9.1.2) and the Gilbert-Elliot channel model (see Section 6.9.1.3) were applied, which have not shown any significant effect on the results.

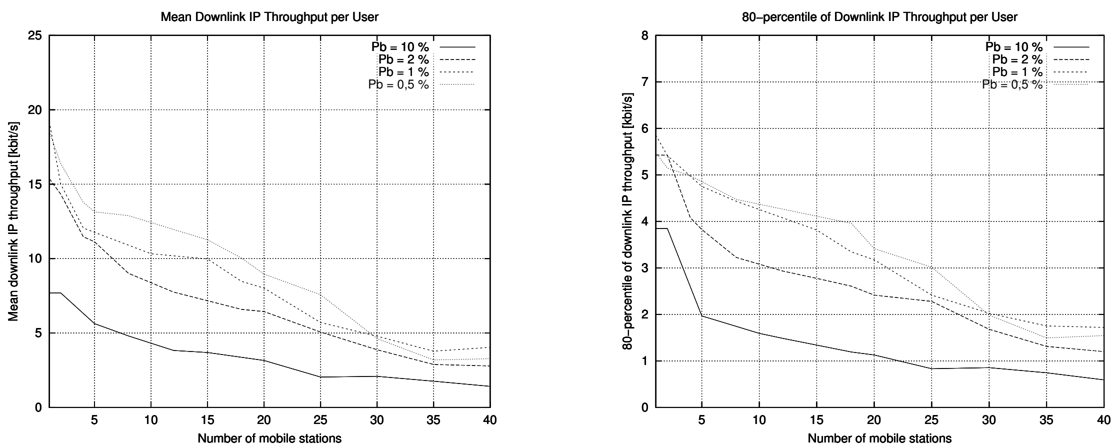
8.3.4 Configurations with On-demand PDCHs

Since GSM networks are dimensioned for a low blocking probability P_b for speech calls, the probability that a number of channels of the trunk of speech channels are unused is high. On the other hand the offered data traffic will be relatively low in GPRS introduction scenarios. As a result the use of on-demand PDCHs that are completely shared between speech and data services leads to an especially efficient radio resource utilization.

In the following, GPRS performance results are presented for cell configurations with on-demand PDCHs. A configuration with three *Transmitter Receiver Units* (TRXs) (21 TCHs) is discussed with speech traffic coexisting to GPRS data traffic that corresponds to a speech call blocking probability P_b of 0.5%, 1%, 2% and 10%, respectively (see Table 8.2). For example, an offered speech traffic of 11.85 Erlang leads to a P_b of 0.5% after the Erlang-B formula. The mean call holding time is assumed to 45 s. Channels are immediately assigned for speech traffic if needed, even if currently used by GPRS (see Section 6.4).

In Figure 8.11 the GPRS throughput performance for different speech call blocking probabilities is shown over the number of mobile stations generating GPRS traffic.

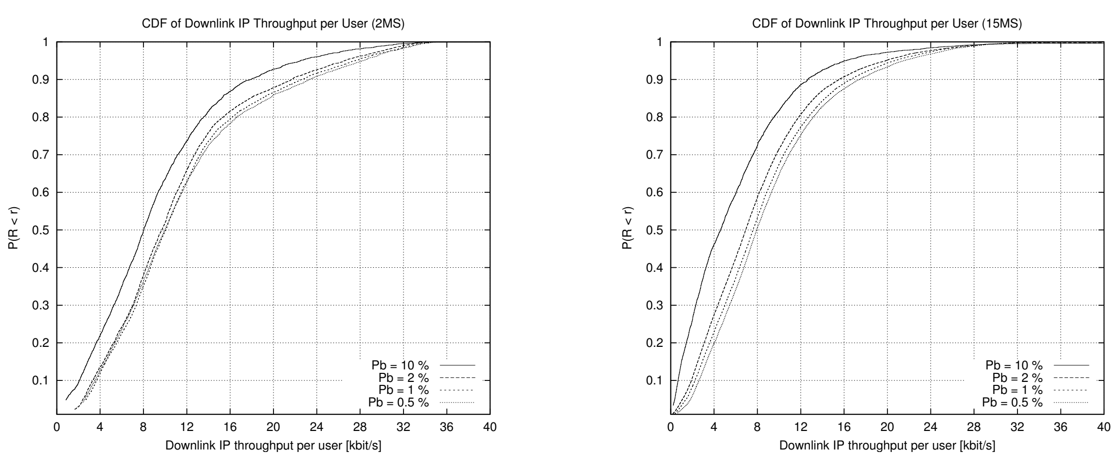
Under low speech traffic load with $P_b = 0.5\%$ the performance is similar to the corresponding



(a) Mean downlink IP throughput per user

(b) 80-percentile of downlink IP throughput per user

Figure 8.11: Throughput performance for different speech call blocking probabilities



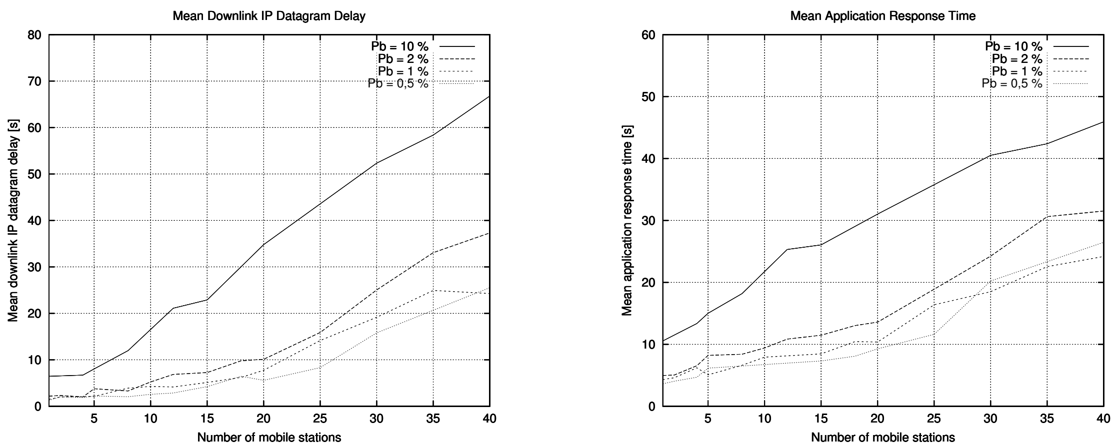
(a) CDF of downlink IP throughput per user (2 MSs)

(b) CDF of downlink IP throughput per user (15 MSs)

Figure 8.12: Throughput distribution for different speech call blocking probabilities

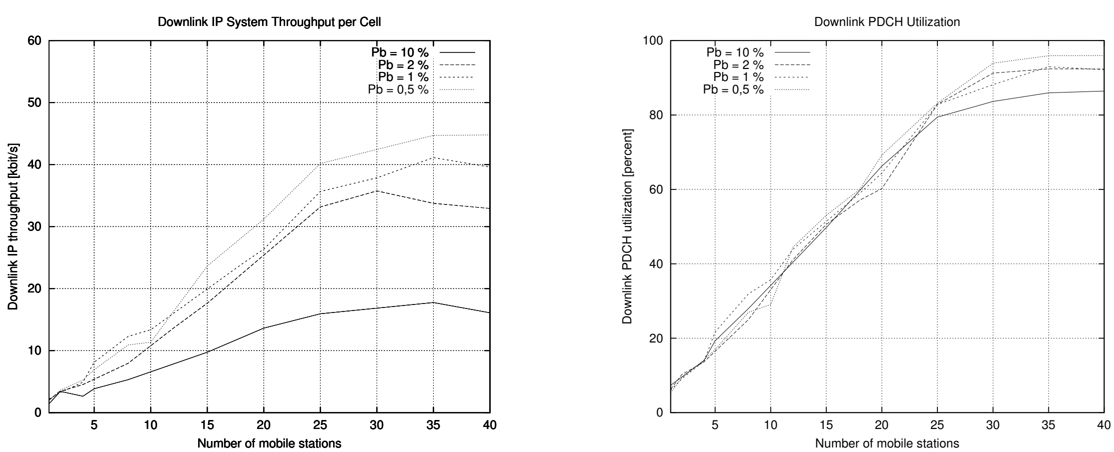
Table 8.2: Offered speech traffic in Erlang for different cell configurations

TCHs	$P_b = 0.5\%$	$P_b = 1\%$	$P_b = 2\%$	$P_b = 10\%$
7 (1 TRX)	2.15	2.50	2.90	4.65
14 (2 TRX)	6.65	7.35	8.20	11.45
21 (3 TRX)	11.85	12.80	14.00	18.65
29 (4 TRX)	18.20	19.45	21.00	27.05
44 (5 TRX)	30.75	32.50	34.65	43.05



(a) Mean downlink IP datagram delay (b) Mean application response time

Figure 8.13: Delay performance for different speech call blocking probabilities



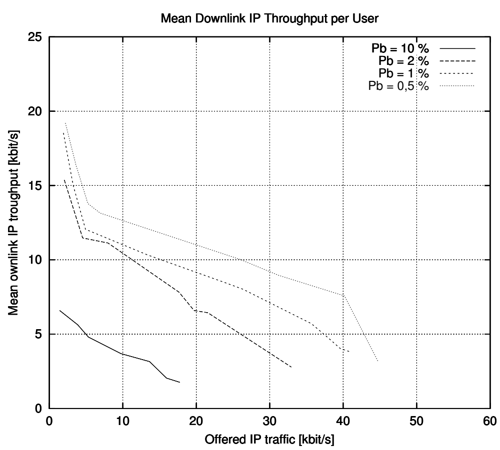
(a) Downlink IP system throughput per cell (b) Downlink PDCH utilization

Figure 8.14: System measures for different speech call blocking probabilities

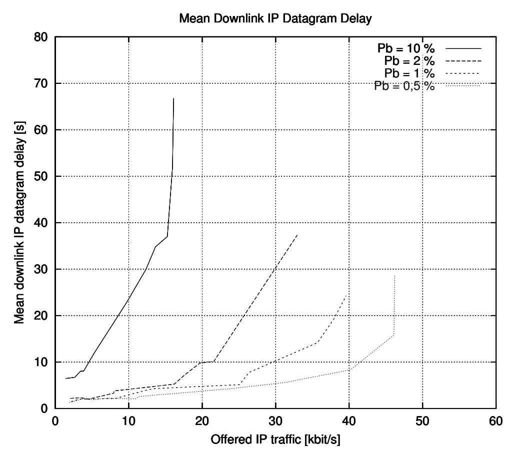
results for 8 fixed PDCHs (see Figure 8.1). For typical speech traffic corresponding to a P_b of 1% and 2% the decrease in mean throughput is in the range of 20–40% compared to 0.5% blocking. This is acceptable, since 2% blocking is already the worst acceptable performance for speech traffic in modern GSM networks world-wide. For a blocking probability of 10% the mean throughput decreases dramatically below 8 kbit/s even with low GPRS traffic load, because nearly every TCP flow is interrupted by an incoming speech call. The 80-percentile of the downlink IP throughput per user is less affected by the coexisting speech traffic, because this measure is determined by the throughput performance of relatively small objects. The probability that a small object is interrupted by a speech call is significantly smaller than for larger objects that are influencing the mean throughput. Therefore an 80-percentile of 2–4 kbit/s can still be reached for a P_b of 10%.

In Figure 8.12 the CDFs of the downlink IP throughput per user is shown for 2 and 15 MSs. For 2 Ms the influence of coexisting speech traffic is only visible for very high speech traffic corresponding to a call blocking probability of 10% (see Figure 8.12(a)). For higher GPRS traffic load (see Figure 8.12(b)) the throughput distribution function is increasing faster for a P_b of 10% already in the range of low throughput values, because even TBFs of small objects are interrupted by speech calls.

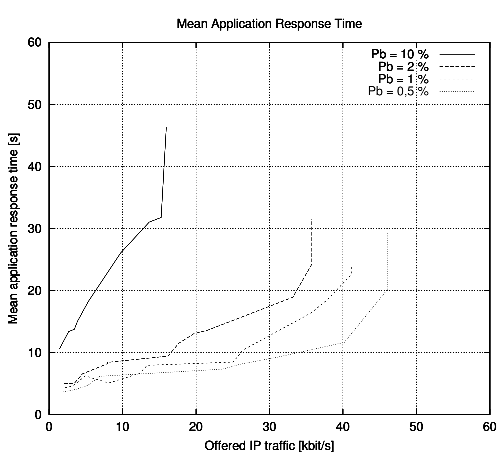
The influence of coexisting speech traffic on the GPRS delay performance under low traffic load becomes only significant, when very high speech traffic, i.e. a P_b of 10%, is regarded (see Figure 8.13). Here, the IP datagram delay exceeds 15 s already with 10 active MSs, whereas the delay performance is similar for different lower speech traffic values (see Figure 8.13(a)). Under



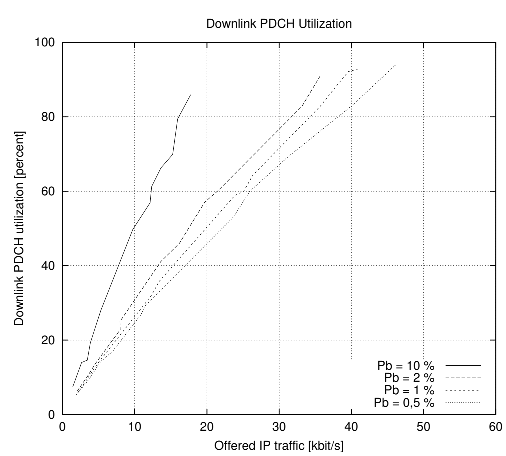
(a) Mean downlink IP throughput per user



(b) Mean downlink IP datagram delay



(c) Mean application response time



(d) Downlink PDCH utilization

Figure 8.15: Performance and system measures over the offered IP traffic

high GPRS traffic load the delay performance is significantly higher for a P_b of 10% exceeding 50 s for 25 MSs. The reason is that nearly every RLC transmission is interrupted by speech calls. The same behavior can be observed regarding the application response time (see Figure 8.13(b)).

The system throughput is also significantly influenced by the coexisting speech traffic especially under high GPRS traffic load (see Figure 8.14(a)), because less capacity is available for GPRS, if more channels are used for speech traffic. Additionally, TCP flows are increasingly interrupted by speech calls with increasing GPRS traffic, so that saturation is reached even earlier. The maximum system throughput per radio cell ranges between 33 and 45 kbit/s for the typical P_b values of 0.5%, 1% and 2%. For a P_b of 10% the system throughput does not exceed 20 kbit/s, even when the utilization of the available PDCHs exceeds 80% (see Figure 8.14(b)).

The graphs in Figure 8.15 are based on the same simulations, but show performance parameters plotted over the offered IP traffic.

Under low traffic load, the mean downlink IP throughput per user decreases similarly for all typical P_b values of 0.5%, 1% and 2%. While the throughput performance for a P_b of 0.5% and 1% decreases slowly in the range of 10–40 kbit/s, a faster degradation can be observed for a P_b of 2% over the whole load range (see Figure 8.15(a)).

The delay performance presented in Figure 8.15(b) and Figure 8.15(c) is characterized by the poles that represent the maximum system throughput for the different blocking probabilities for speech traffic.

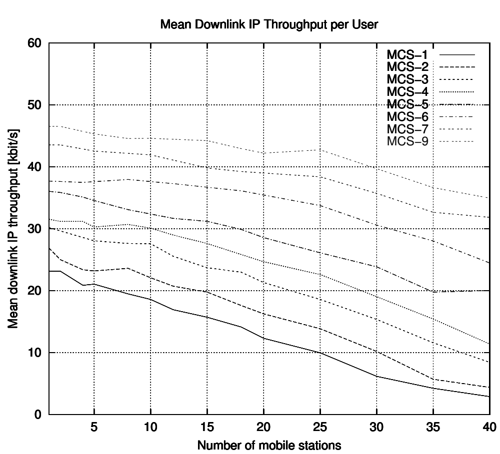
The PDCH utilization shows a linear behavior over the increasing number of active stations (see Figure 8.15(d)). The gradient is depending on the available capacity that differs depending on the average number of available PDCHs for GPRS traffic that are left open by the speech service.

8.4 Simulation Results for EGPRS

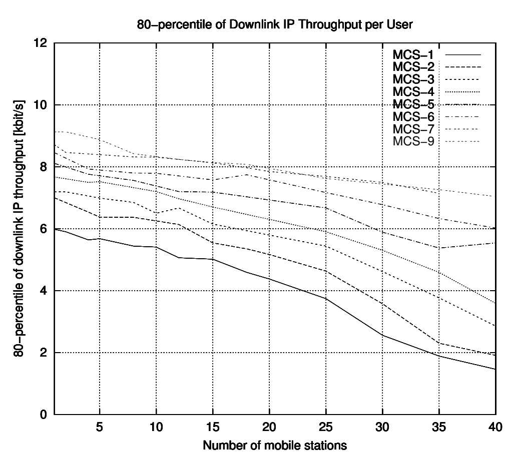
In this section, the impact of the EGPRS *Modulation and Coding Schemes* (MCSs) (see Section 4.2) and of the functions for *Link Quality Control* (LQC) (see Section 4.3) on the performance of typical applications like WWW and e-mail are discussed.

8.4.1 Performance of EGPRS Modulation and Coding Schemes

The following simulation results are based on the general parameter settings that were introduced in Section 8.1. Eight fixed PDCHs and a multislot capability of 4/1 are assumed. The maximum performance achievable with the EGPRS MCSs for this traffic mix is determined. Therefore an



(a) Mean downlink IP throughput per user



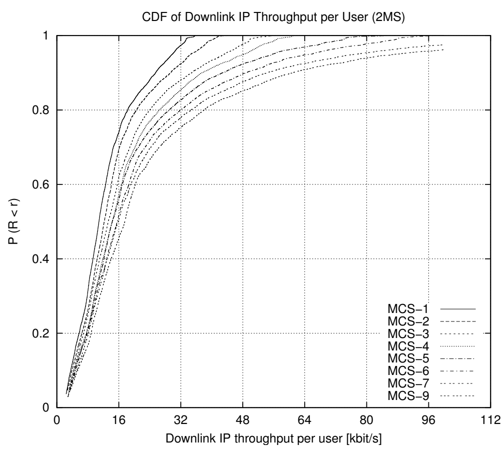
(b) 80-percentile of downlink IP throughput per user

Figure 8.16: Throughput performance for different Modulation and Coding Schemes (MCSs)

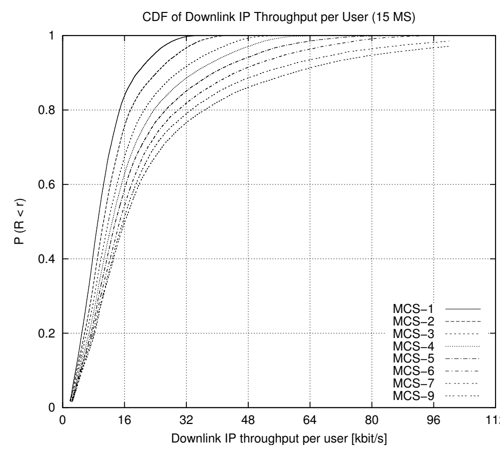
error-free channel is regarded in this section. Channels producing block errors are considered in the next section on the EGPRS LQC options (see Section 8.4.2).

Figure 8.16(a) shows the mean downlink IP throughput per user as a function of the number of active EGPRS users and of the used fixed MCSs. The EGPRS MCSs 1 to 4 are corresponding to the GPRS CSs CS-1 to CS-4. The differences to the results presented in [Stuckmann (2002b); Walke (2001)] can be explained with a more realistic TCP implementation that was applied for the results in this thesis, which is comprising both slow start and congestion avoidance consistent with the specification *TCP Reno* (see Section 6.2).

Under optimal conditions (only one user active, MCS-9, error-free channel) an EGPRS user can reach a maximum mean data rate of 47 kbit/s for the chosen traffic mix. This is an improvement of about 50% compared with the throughput for CS-4 in GPRS. With 30 users the benefits from EGPRS MCSs become even more significant. Because of the higher effective radio capacity in EDGE, the gain exceeds 100% compared to the GPRS CSs. While the mean throughput of MCS-4

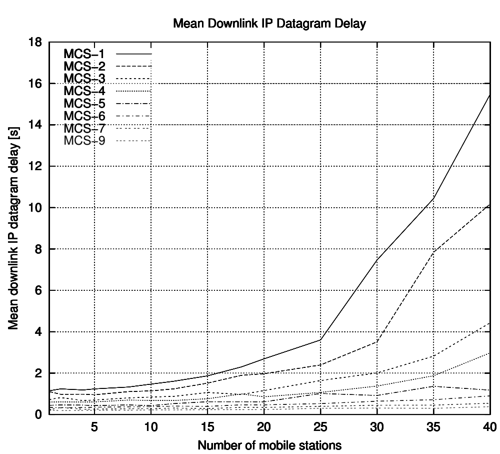


(a) CDF of downlink IP throughput per user (2 MSs)

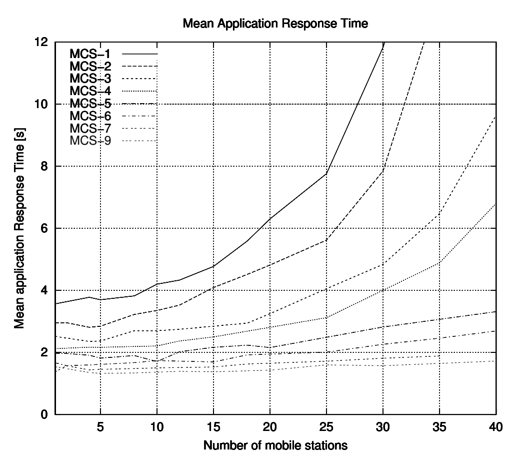


(b) CDF of downlink IP throughput per user (15 MSs)

Figure 8.17: Throughput distribution for different Modulation and Coding Schemes (MCSs)



(a) Mean downlink IP datagram delay



(b) Mean application response time

Figure 8.18: Delay performance for different Modulation and Coding Schemes (MCSs)

passes below 25 kbit/s with already 20 stations, a nearly constant throughput can be achieved up to 30 MSs with MCS-7 to MCS-9. These performance and capacity gains are related to the maximum possible performance for the traffic mix considered assuming an error-free channel. When realistic radio channel conditions are taken into account, the higher MCSs are more sensitive to block errors than lower MCSs and the gain is less significant then (see Section 8.4.2).

The performance gain of higher MCSs compared to lower ones for the 80-percentile of downlink throughput is lower, because this measure is mainly determined the throughput reached for smaller objects that are mostly influenced by the round-trip delay especially over the air interface and less by the available bit rate (see Figure 8.16(b)).

This explanation is confirmed by the cumulative distribution functions of throughput (see Figure 8.17).

In Figure 8.17(a), it becomes visible that the maximum throughput values can be significantly increased for higher MCSs, but the range for smaller throughput values is nearly independent of the MCS. For 15 active MSs as shown in Figure 8.17(b) a dependency in the probability of lower throughput values on MCSs can be observed, because higher MCSs have higher capacity reserves for more mobile stations to be served.

Figure 8.18 shows the delay performance for the MCSs of EGPRS. A comparison between MCS-4 and MCS-9 shows that the application response time can be halved under low traffic load and even reduced to a quarter with high traffic load. With MCSs 5 to 8, more than 40 active MSs can easily be served maintaining good performance, i.e. an application response time of a few seconds (see Figure 8.18(b)).

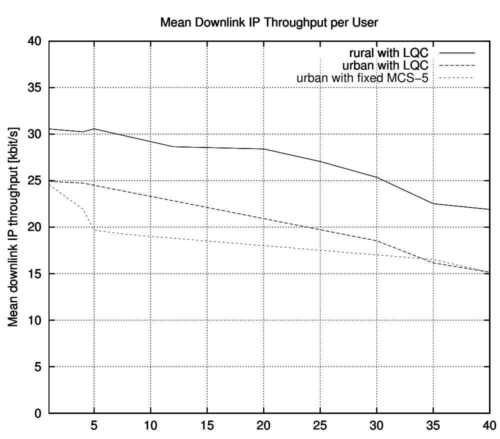
8.4.2 Performance of Different EGPRS LQC Options

While in the previous section an error-free channel is assumed, the channel model considering mobility and C/I dependent block errors (see Section 6.9.1.4) is applied for the examination of the LQC options in this section.

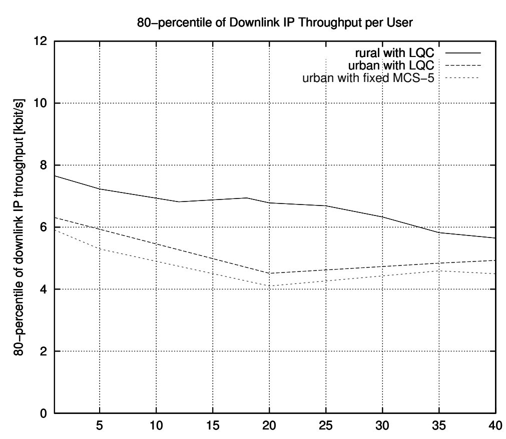
The channel conditions are determined by the cell and cluster size that are the basis for the C/I calculation. A rural scenario with cluster size 7, a cell size with a radius of 3000 m and an MS velocity of 100 km/h as well as an urban scenario with cluster size 3, 300 m cell radius and an MS velocity of 7 km/h are considered.

The same traffic mix and the same protocol settings and capabilities as used in Section 8.4.1 are assumed.

The sensitivity of the EGPRS throughput performance to the channel conditions can be seen

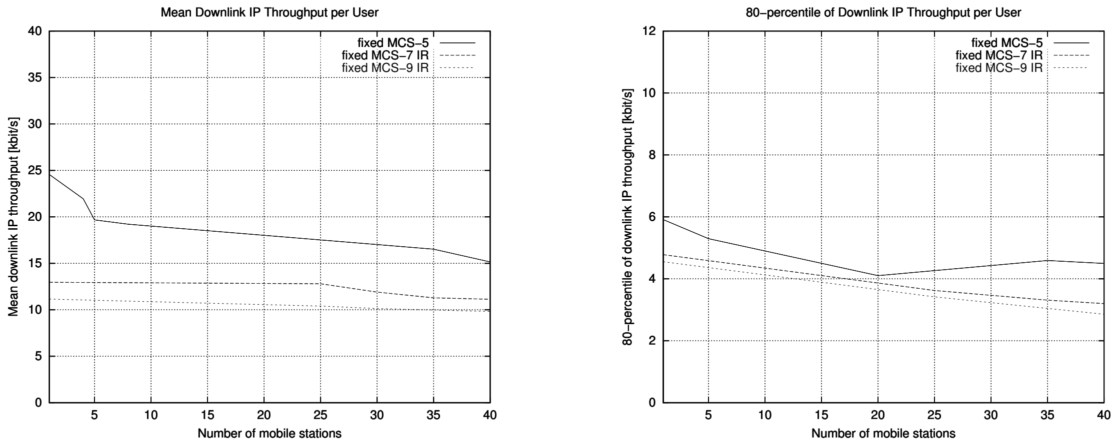


(a) Mean downlink IP throughput per user



(b) 80-percentile of downlink IP throughput per user

Figure 8.19: Throughput performance for LQC vs. fixed MCS



(a) Mean downlink IP throughput per user

(b) 80-percentile of downlink IP throughput per user

Figure 8.20: Throughput performance for different LQC options

from Figure 8.19. The absolute performance difference comparing the mean throughput in the rural and the urban scenario with full LQC capabilities equals about 5 kbit/s for the whole load range. The performance gain achieved by LQC compared to a well-chosen fixed MCS without LQC is 20% under low and 10% under medium traffic load (for this scenario MCS-5 has proven to be the best fixed MCS in sample simulations). Similar conclusions can be drawn from the results for the 80-percentile of the throughput. However, the sensitivity of this measure to the radio channel conditions is lower resulting in a maximum difference of 3 kbit/s for rural and urban with LQC, because the 80-percentile is determined by the throughput performance of small objects that is already limited by TCP slow start (see Section 6.14.1).

Figure 8.20 shows the throughput performance of the IR option compared to the fixed MCS-5 in the urban environment. It becomes clear that the appliance of an aggressive MCS like MCS-7 or MCS-9 together with IR can not maintain a performance comparable to the well-chosen fixed MCS-5. Already MCS-7 with IR looses more than 50% of the performance possible with fixed MCS-5. The reason is the lack of flexibility of IR compared to LA, where the MCS can be chosen dynamically. With MCS-7 and MCS-9 block errors in periods of very bad channel conditions, e.g., at the cell border can not be recovered by IR. Similar conclusions can be drawn for the 80-percentile and the throughput distribution functions, although the performance difference between the aggressive MCSs with IR compared to fixed MCS-5 is less significant.

However, IR enables MCS-9 to realize a performance similar to MCS-7. The 80-percentile of MCS-7 and MCS-9 with IR is almost equal because the dominant small objects are less sensitive to block errors as explained above [Stuckmann and Franke (2001)].

Traffic Performance of Existing and Future Applications

After a comprehensive traffic performance analysis has been carried out in the previous chapter, regarding different GPRS and EGPRS cell configurations, system parameters and radio channel conditions, this chapter presents traffic performance results for different traffic mixes.

These results are useful to estimate the radio capacity counted in the number of PDCHs that is needed during the evolution of GSM networks towards 3G systems under increasing general traffic load [Stuckmann and Hoymann (2002c)]. First a traffic mix of applications based on WAP and conventional Internet applications like WWW browsing and e-mail over GPRS are regarded [Stuckmann and Hoymann (2002a,b)]. In the next step traffic performance results for Streaming applications over GPRS and EGPRS are presented and the feasibility of Streaming with coexisting interactive and background applications like WWW and e-mail is examined [Hoymann and Stuckmann (2002a)]. The same general default parameter settings as described in Section 8.1 are applied. While in Chapter 8 the overall performance measures are presented when averaging the performance of WWW and e-mail traffic, in this chapter the performance measures are presented for each application, separately.

9.1 WAP in Comparison to Internet Applications over GPRS

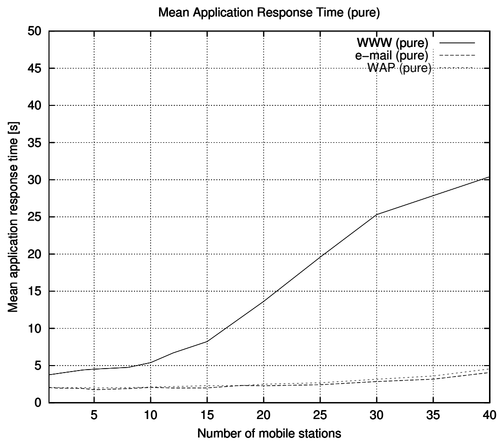
To be able to compare the user-perceived performance of WAP in comparison to conventional Internet applications, the performance and system measures are examined for WWW, e-mail and WAP traffic separately and in a traffic mix scenario. The traffic mix scenario is characterized by the session probabilities for WAP, e-mail and WWW of 60%, 28% and 12%, respectively. Consequently four traffic scenarios are examined: 100% WWW traffic, 100% e-mail traffic, 100% WAP traffic and the scenario with the traffic mix.

9.1.1 Delay Performance

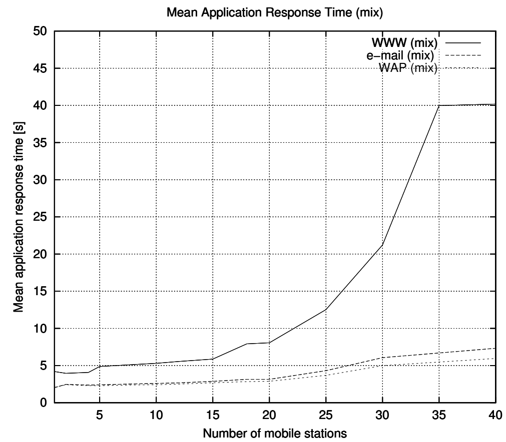
For WAP, the application response time is the critical performance measure, because the file sizes for WAP are comparatively low (1 kbyte for monochrome decks and 3 kbyte for colored decks, see Section 5.4.4). Users of this kind of transaction-oriented applications typically do not notice the effective throughput during transmission, but only the time they have to wait for the response of the short transaction.

Under low traffic load the performance of different applications is characterized by the traffic models themselves, e.g., file sizes and the used transport protocol. The response time for both WAP and e-mail remains below 2 s, whereas the response time for a web page is around 4 s (see Figure 9.1(a)). Although WAP decks are comparatively small, the average response time cannot fall below 2 s because of the long TBF setup time. The reason for the high WWW page response time is that a web page is composed of several objects, which are transmitted in sequence. Even though HTTP version 1.1 is modeled using persistent TCP connections, each object has to be requested separately by a new get request. The e-mail model is characterized by file sizes, comparable to the WWW page size, but e-mails can be transmitted entirely over one TCP connection without interruptions. Therefore the e-mail response time under low traffic load is significantly lower than the WWW page response time.

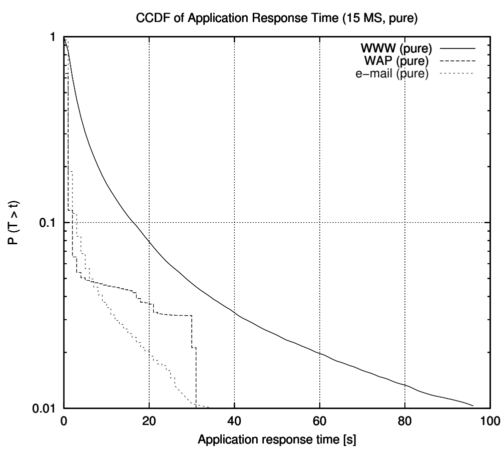
Under higher traffic load, the response time for a WAP deck remains nearly constant for up to 20 MSs and does not exceed 5 s even with 40 MSs offering pure WAP traffic. The response time for e-mail traffic is also remaining nearly constant because of very short sessions in the e-mail traffic model. While WWW sessions consist of several pages per session, an e-mail session only consists of one e-mail download. The inactive period between sessions is equal for all simulations (see Section 8.1), whereas WWW sessions are longer than e-mail sessions. This leads to a faster increase in PDCH utilization for WWW and therefore to a faster decrease in performance for an



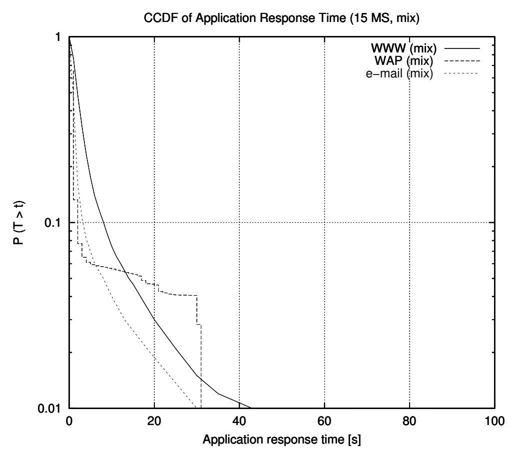
(a) Mean application response time (separate applications)



(b) Mean application response time (traffic mix)



(c) CCDF of application response time (15 MSs, separate applications)



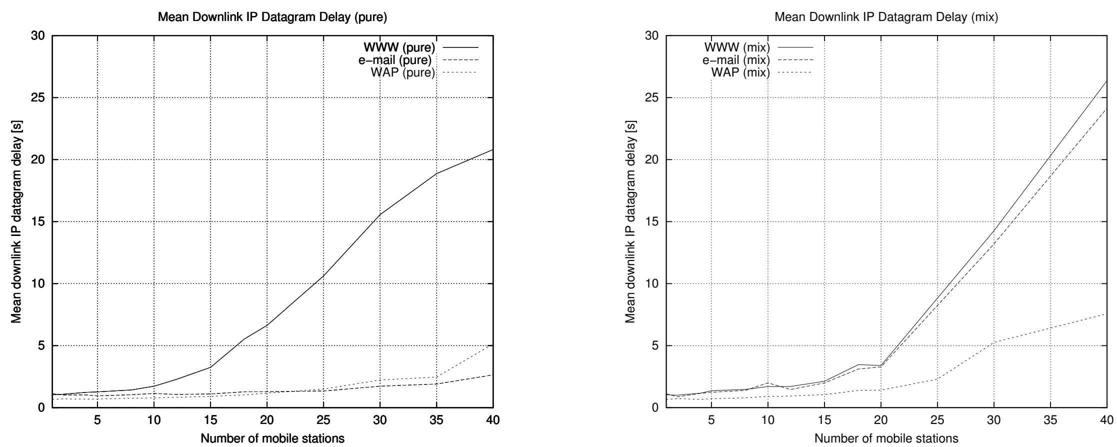
(d) CCDF of application response time (15 MSs, traffic mix)

Figure 9.1: Application response time for WAP, WWW and e-mail

increasing number of active stations (see Figure 9.4). While the scenario with pure WWW traffic already reaches saturation with 18 MSs, the same number of WAP or e-mail stations does not even utilize 50% of the channel capacity.

For more than 30 WWW users the radio cell is congested and TCP reduces the offered traffic, which leads to a slower increase of the WWW page response time compared to the range of 20–30 WWW users.

In the traffic mix scenario both WAP and e-mail response times are increasing faster with an increasing number of users and are exceeding 5 s with 30 active stations offering the traffic mix (see Figure 9.1(b)). However, like in the scenarios with pure WAP and e-mail traffic, the response times for WAP and e-mail remain below an acceptable value of 10 s for the whole regarded load range. The WWW page response time in the mix scenario (see Figure 9.1(b)) is increasing slower than the WWW response time in the scenario with pure WWW traffic (see Figure 9.1(a)), because the load is now characterized by 60% WAP sessions and is increasing slower with an increasing number of users (see Figure 9.4(a)). While a mean WWW response time of 10 s is already exceeded with 17 MSs with pure WWW traffic, a response time of 10 s can be maintained even with 23 MSs in



(a) Mean downlink IP datagram delay (separate applications)

(b) Mean downlink IP datagram delay (traffic mix)

Figure 9.2: Downlink IP datagram delay for WAP, WWW and e-mail

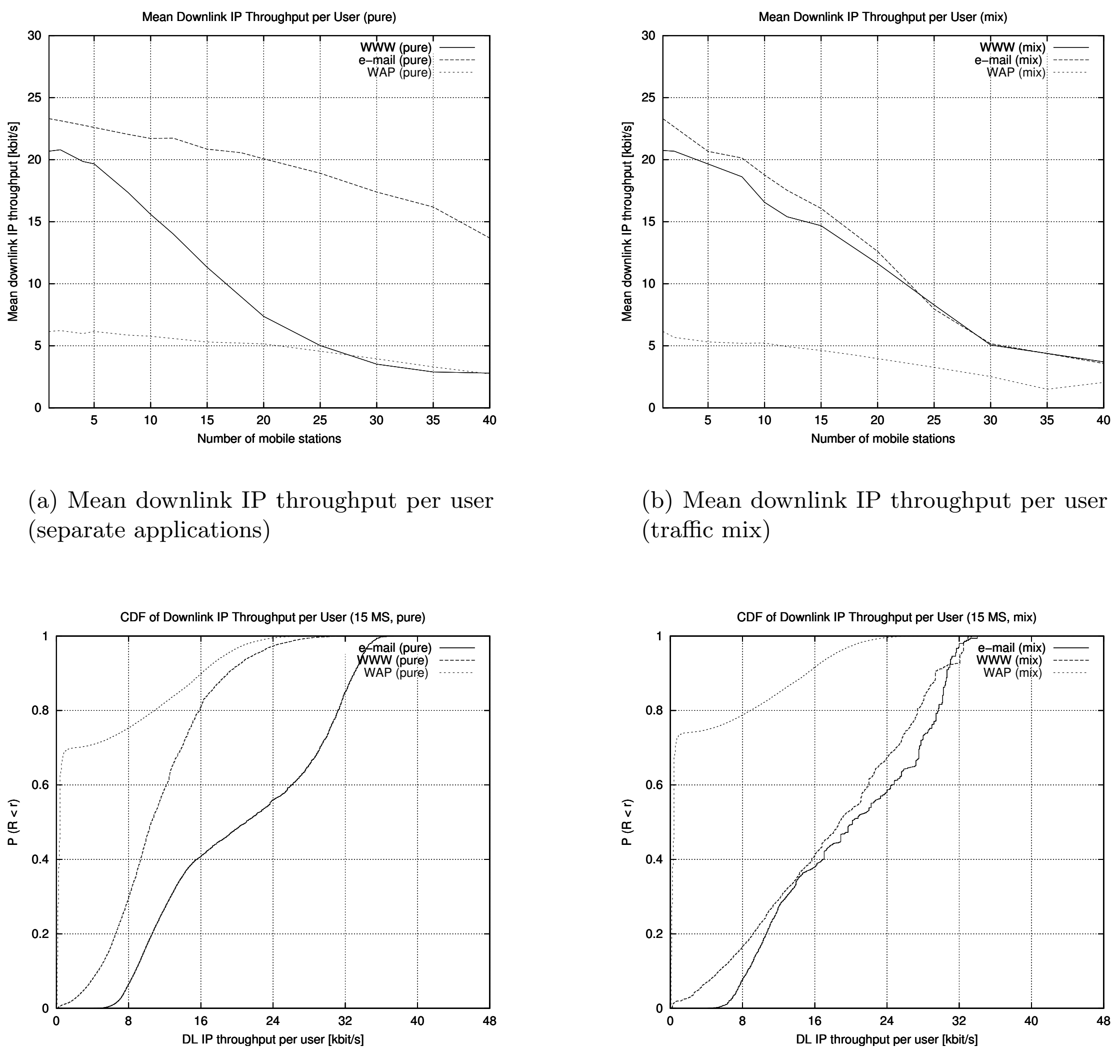
the traffic mix scenario.

Figure 9.1(c) and Figure 9.1(d) show the CCDFs of the WAP, WWW and e-mail application response times for 15 MSs. When the applications are regarded separately (see Figure 9.1(c)) the traffic model of each application determines the distribution function. Since WWW traffic is characterized by a heavy-tailed distribution for the object size, the probabilities for higher response times cannot be neglected. The WAP deck size is limited and therefore the probabilities for response times higher than 30 s is approaching zero. Since the probabilities for very small WAP decks is high, more than 90% of the WAP decks are received within less than a few seconds. In Figure 9.1(d) the influence of WWW traffic on WAP traffic and vice versa becomes visible. While the distribution functions for WAP and e-mail are very similar to the distributions for pure WAP and e-mail traffic, the performance of WWW traffic is better than for pure WWW traffic, since the portion of WAP and e-mail terminals is offering less traffic than WWW terminals do. Therefore 90% of the WWW objects are received within 8 s, whereas in the scenario with pure WWW traffic the 90-percentile is higher than 17 s.

In Figure 9.2 the mean IP datagram delay is shown for the applications regarded separately (see Figure 9.2(a)) and in the traffic mix scenario (see Figure 9.2(b)). For few active stations the datagram delay for pure WWW and pure e-mail traffic is higher than for WAP traffic because of larger files. WAP decks are only segmented into a few UDP/IP datagrams. This leads to short queues in lower layers and short waiting times, while for WWW and e-mail larger queue sizes are existing once the TCP steady state is reached.

With an increasing number of active stations the IP datagram delay for WWW increases to more than 15 s for 30 WWW stations because of the high channel utilization (see Figure 9.4(b)). The datagram delay for WAP exceeds the delay for e-mail with 23 MSs and increases to 5 s with 40 stations offering WAP traffic. The reason for this increase is the higher utilization compared to e-mail traffic (see Figure 9.4(b)) because of relatively long sessions (20 decks per session on average, see Section 5.4.4). E-mail sessions are shorter only composed of one e-mail download with inactive periods following each download. Consequently, for WAP traffic, more stations are competing for the available PDCHs than for e-mail traffic.

In the traffic mix scenario all applications are operating under the same traffic load conditions for a given number of mobile stations. Therefore the IP datagram delay for both WWW and e-mail is increasing rapidly for more than 20 stations exceeding 20 s with more than 35 MSs offering the traffic mix. The datagram delay for WAP traffic also starts to increase significantly for more than 20 MSs, but remains below 10 s for 40 MSs because of the small number of datagrams to be transmitted for each WAP deck.



(a) Mean downlink IP throughput per user (separate applications)

(b) Mean downlink IP throughput per user (traffic mix)

(c) CDF of downlink IP throughput per user (separate applications)

(d) CDF of downlink IP throughput per user (traffic mix)

Figure 9.3: Throughput performance for WAP, WWW and e-mail

9.1.2 Throughput Performance

While the throughput performance is the critical performance measure for download-oriented applications like WWW, it plays only a minor role for WAP applications compared to the application response time (see Section 9.1.1). Response times for these transaction-oriented applications typically lie in the range of a few seconds, even if the throughput performance is low.

Figure 9.3 shows the throughput performance for WAP, WWW and e-mail. The mean downlink IP throughput per user for WAP equals 7 kbit/s for 1 MS in the radio cell. This poor throughput performance for WAP traffic can be explained by the low deck size. WAP applications are more influenced by the high round-trip time, which is mainly caused by the high delay over the air interface, than by the available bit rate. Since the response time for a WAP deck is in the range of a few seconds (see Section 9.1.1), which should be acceptable for a wireless application, the user is not aware of this low throughput performance.

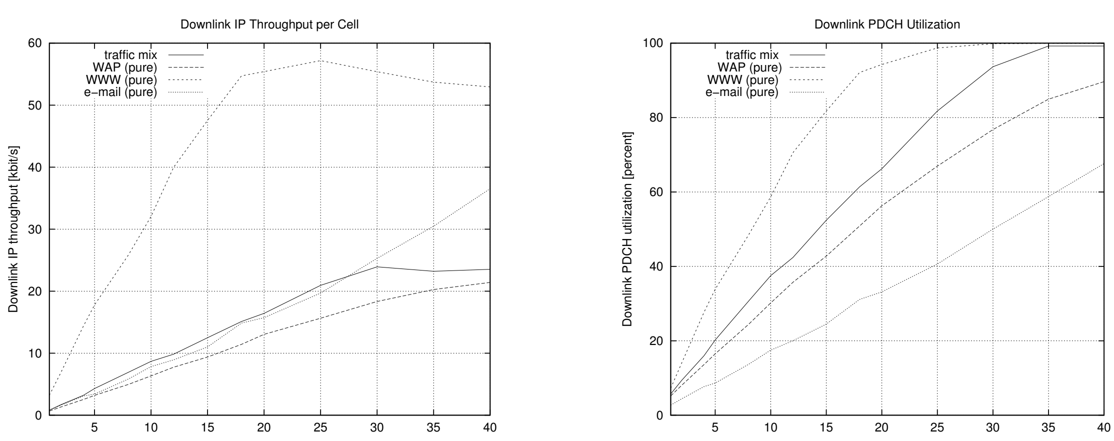
As discussed in Section 6.14.1 the maximum throughput of WWW and e-mail is limited by the TCP slow start function to about 22 kbit/s. While pure e-mail traffic reaches a throughput of

24 kbit/s, WWW is limited to 21 kbit/s because the e-mail traffic model is characterized by slightly larger file sizes. With an increasing number of mobile stations the WWW throughput performance decreases dramatically to 5 kbit/s with 25 active stations. Since for both WAP and e-mail the offered traffic per session is much lower than for WWW, the performance is decreasing slower compared to WWW as explained for the delay performance in Section 9.1.1. In the traffic mix scenario all applications are facing the same PDCH utilization for a given number of users that is dominated by the high traffic load offered by the WWW applications. Therefore the performance of e-mail and WAP traffic is decreasing faster than in the scenario with pure WAP and e-mail traffic and the WWW throughput performance is decreasing slower than for pure WWW traffic. Since the file sizes of WWW and e-mail are comparable, the performance of these two applications is similar for the whole load range in the traffic mix scenario. For more than 25 MSs the difference in the traffic model for WWW and e-mail does no longer affect the traffic performance and the performance is completely determined by the available radio resources.

Figure 9.3(c) and Figure 9.3(d) show the throughput distributions for WWW, e-mail and WAP regarded separately and in the traffic mix scenario. Since WAP traffic is characterized by the small WAP deck size, a high portion of the evaluated throughput values are very low. The throughput performance of WWW is worse than e-mail, because a higher utilization (80 %) is reached for 15 WWW stations compared to 25% utilization for e-mail (see Figure 9.3(c)). For the same reason very low throughput values are occurring scarcely in the distribution function for e-mail. The high gradient in the distribution function in the range of higher throughput values can be explained by the e-mail traffic model that is composed of very small files representing text e-mails and larger files representing e-mails with attachment. In the traffic mix scenario the throughput distribution function for WAP is nearly unaffected, since WAP traffic is not significantly affected by a small portion of WWW and e-mail traffic (see Figure 9.3(d)). The distributions for WWW and e-mail are similar in the range of medium and high throughput values, because both applications are operating under the same traffic load conditions in this traffic mix scenario.

9.1.3 System Measures

In Figure 9.4 the system measures are presented for the four traffic scenarios. WWW sessions are composed of several pages comprising several objects. Therefore they are requesting a higher traffic volume than WAP and e-mail applications. Consequently the system throughput per cell in the scenario with pure WWW traffic is increasing much faster than in the scenarios with WAP and e-mail traffic (see Figure 9.4(a)). The traffic models for WAP and e-mail are differing in the file size, which is significantly higher for e-mail, and in the session length, which is higher for WAP.



(a) Downlink IP system throughput per cell

(b) Downlink PDCH utilization

Figure 9.4: System measures for WAP, WWW and e-mail

On the one hand, the e-mail traffic is offering more traffic during the activity period, on the other hand it is characterized by more and longer inactive periods (the inactivity time after each session is constant for all applications, see Section 8.1). Therefore the system load in the scenario for pure WAP traffic and for pure e-mail traffic is increasing similarly with an increasing number of users.

In situations with higher traffic load the WAP traffic approaches saturation earlier than e-mail traffic (see Figure 9.4(b)), so that e-mail traffic continues to increase even in the range of 40 stations. The reason for the early saturation point for WAP traffic is the high session length and therefore the large number of WAP stations with ongoing sessions that are assigned to radio channels even if they have no new data to transmit (see Section 3.4.5.5.3). This can be confirmed by the PDCH utilization in Figure 9.4(b), where 80% utilization is reached for 32 MSs offering pure WAP traffic, whereas e-mail traffic only reaches around 50%.

The system throughput for the scenario with the traffic mix is not exceeding 25 kbit/s, because it is dominated by WAP traffic. As explained above, a high portion of WAP traffic results in high utilization, but also in low offered IP traffic (see Figure 9.4(a)). Saturation is reached with 30 MSs offering the traffic mix with a PDCH utilization of 92% (see Figure 9.4(b)).

9.2 Streaming over GPRS and EGPRS

As a typical EGPRS introduction scenario, Streaming applications over EGPRS are examined in coexistence with WWW and e-mail applications. The Streaming applications are offering Video traffic with a mean offered data rate of 14.39 kbit/s (see Section 5.5.1.2) and Audio traffic with a constant offered data rate of 6.3 kbit/s representing a low-quality Audio Codec (see Section 5.6.1.1).

Two traffic scenarios are regarded: 100% Streaming traffic and a traffic mix scenario considering Streaming in coexistence with WWW and e-mail. Following the conservative predictions concerning the future usage of Streaming applications the traffic mix contains only 10% Streaming sessions. The remaining part is assumed to 63% e-mail and 27% WWW sessions.

The same GPRS and EGPRS simulation parameter settings as described in Section 8.1 are applied.

9.2.1 Throughput and Delay Performance

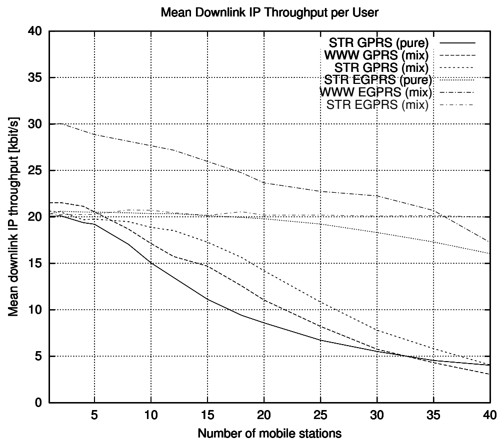
While for download- and transaction-oriented applications the throughput and response time are representing the user-perceived performance, both throughput and delay performance are critical for Streaming applications.

Figure 9.5 shows the throughput and delay performance over GPRS and EGPRS for pure Streaming traffic as well as for Streaming and WWW traffic in the traffic mix scenario. For EGPRS and pure Streaming traffic the throughput performance remains acceptable for up to 20 stations. For more than 20 MSs the offered data rate can no longer be served by EGPRS. Consequently the throughput per user starts to decrease below the mean offered data rate of 20.69 kbit/s (see Figure 9.5(a)) and the mean IP datagram delay starts to increase significantly (see Figure 9.5(b)). Since only 10% of the sessions in the traffic mix scenario are Streaming sessions, EGPRS can easily serve 40 stations that are offering the traffic mix. Therefore the throughput and delay performance for Streaming is constant for the whole load range and the performance of the WWW sessions is decreasing from 30 kbit/s for 1 MS to 17 kbit/s for 40 MSs.

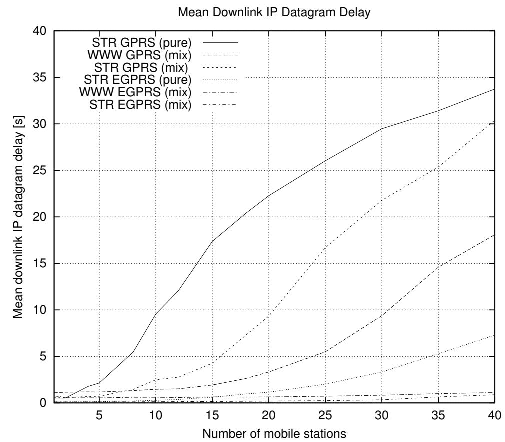
In scenarios with GPRS, Streaming applications can only be realized in situations with low traffic load. While for the GPRS scenario with 100% Streaming traffic, less than 4 MSs can be served, 10 stations offering the traffic mix already result in unacceptable throughput and delay performance for Streaming applications over GPRS.

In Figure 9.5(c) and Figure 9.5(d) the IP datagram delay distributions for 2 and 15 active stations are shown. Here it can be estimated, which scenarios are acceptable for Video Streaming decoders that are sensitive to buffer underflows caused by high delay variations. With 2 active stations as shown in Figure 9.5(c) 90 percent of the delay samples are below 1 s for both EGPRS and GPRS, which is acceptable.

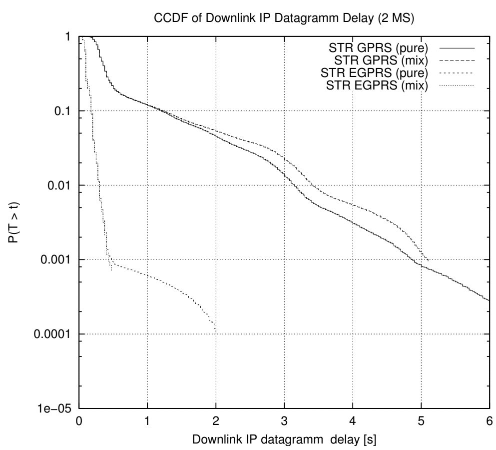
While no high delay variations can be observed in the scenarios for EGPRS, the distributions for Streaming over GPRS are characterized by a larger value range (see Figure 9.5(d)). In the scenario with 100% Streaming traffic 50% of the datagrams are transmitted within 3 s, while 20%



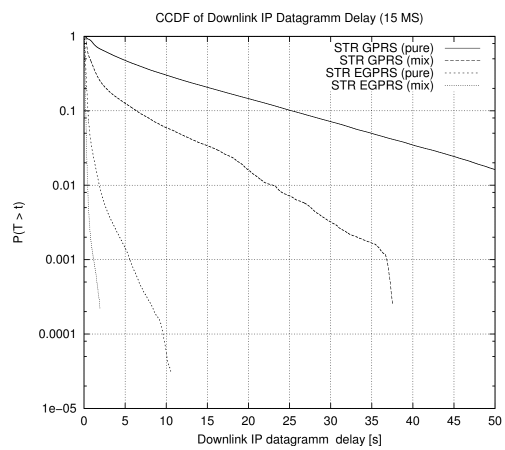
(a) Mean downlink IP throughput per user



(b) Mean downlink IP datagram delay



(c) CCDF of downlink IP datagram delay (2 MSs)



(d) CCDF of downlink IP datagram delay (15 MSs)

Figure 9.5: Throughput and delay performance for Streaming and WWW traffic

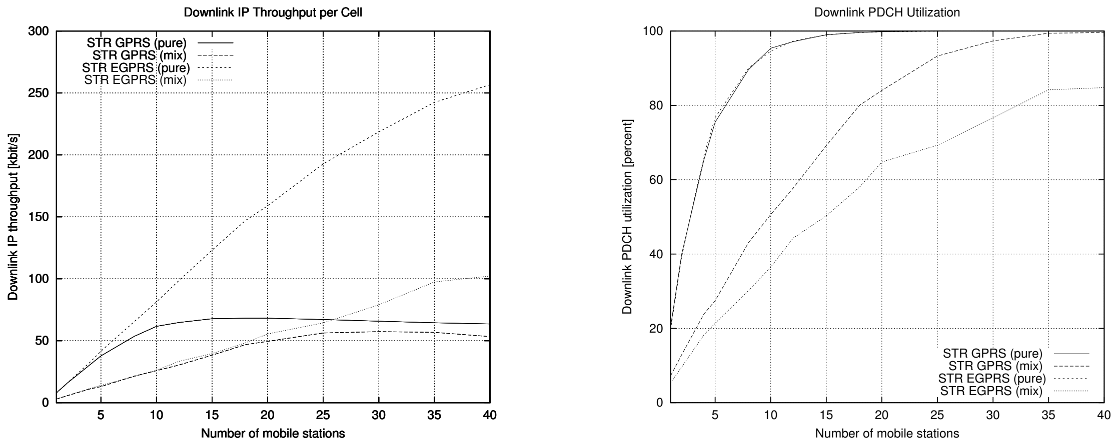
of the datagrams are delayed for more than 20 s. Even with a jitter buffer at the receiver of 10 s such a delay variation is unacceptable. For 15 MSs offering the traffic mix the delay performance is also not acceptable in all cases, since 10% of the datagrams are delayed for more than 6 s. This scenario with 15 MS is just representing the limit between acceptable and unacceptable performance.

9.2.2 System Measures

In Figure 9.6 the system measures are shown for the two traffic scenarios over GPRS and EGPRS.

In contrast to TCP-based traffic, Streaming traffic is inelastic, i.e. the offered traffic does not depend on the available bandwidth. Therefore the IP system throughput per cell increases linearly as long as the offered traffic is completely served by the air interface. Saturation is nearly reached with 260 kbit/s for EGPRS using different MCSs controlled by LA (see Section 8.4.2). The maximum throughput for the traffic mix scenario in EGPRS only slightly exceeds 100 kbit/s, because the offered TCP-based traffic (WWW and e-mail) is adapting to the available bit rate for TCP connections. Therefore only a fraction of the available radio capacity can be used effectively.

Since in GPRS CS-2 is applied, the system throughput limit is around 60 kbit/s. This limit is



(a) Downlink IP system throughput per cell

(b) Downlink PDCH utilization

Figure 9.6: System measures for Streaming traffic

reached with 10 MSs for the scenario with pure Streaming traffic and with 25 MSs for the traffic mix scenario.

Again it becomes visible that the performance for situations with low traffic load is mainly characterized by the offered traffic and the performance for situations with high traffic load is affected by the available radio resources. Therefore the performance for the traffic scenario with pure Streaming traffic and with traffic mix are equal for a low number of mobile stations, whereas the performance of the two GPRS scenarios and the two EGPRS scenarios show similar behavior with a higher number of mobile stations.

In Figure 9.6(b), the PDCH utilization is shown. While 80% utilization is already reached for 5 GPRS MSs in both traffic scenarios, for EGPRS the system approaches saturation with more than 20 MSs. Here the large capacity gain achieved by EGPRS becomes visible again (see Section 8.4.1).

CHAPTER 10

Traffic Engineering Rules and QoS Support

In this chapter traffic engineering rules, which are derived from the results in Chapter 8 and Chapter 9, are presented for GPRS introduction and evolution scenarios with fixed and on-demand PDCHs. While these rules are valid for best-effort services, traffic management techniques with service class differentiation are affecting traffic engineering rules (see Section 2.2.3).

Traffic management techniques for cellular packet radio networks, developed in the framework of this thesis, are introduced in this chapter comprising scheduling and admission control algorithms for best-effort services and traffic class support. The performance of these algorithms is evaluated. Finally the influence of traffic management on traffic engineering rules is discussed for an EDGE introduction scenario.

10.1 GPRS/EDGE Traffic Engineering

Since GPRS will be integrated into existing GSM systems, which are mainly carrying speech traffic, and the GPRS traffic just after service introduction will be rather low, operators will initially deploy GPRS using on-demand PDCHs (see Section 8.3.4). On the other hand the mean number of free channels under a given blocking probability might be too low to carry the packet data traffic and operators might want to allocate a number of fixed PDCHs to be able to guarantee the availability of the GPRS. Therefore, traffic engineering rules for both fixed and on-demand channel configurations as well as for mixed configurations with a combination of fixed and on-demand PDCHs have to be developed [Stuckmann and Paul (2001)].

As in GPRS networks no delay-critical applications will be introduced initially, traffic engineering rules are proposed for mean values first of all. While for Conversational applications the delay and delay variation is critical, for interactive download-oriented applications like WWW and e-mail the average throughput performance is of interest in this context.

The approach introduced in the following can also be applied for engineering the number of PDCHs to reach some mean delay value or to reach a predefined quantile of throughput or delay. Related results are provided in Section 8.3.1 and Section 8.3.4.

10.1.1 Traffic Engineering Rules for Fixed PDCHs

If GSM radio cells have high capacity reserves, i.e., an acceptable blocking probability for speech calls can be guaranteed even if traffic channels are withdrawn for GPRS, then GSM physical channels can be allocated as fixed PDCHs to ensure a guaranteed QoS of GPRS-based services.

As explained in Section 2.2.1.3 the traffic engineering rules are based on an accurate model with realistic parameter settings, while the rules themselves should be applicable only taking into account the user number and traffic volume during the busy hour. In planning scenarios, typically hundreds of potential users per cell are regarded that can be active or inactive offering traffic volumes in the order of several kbyte/h. Thus the traffic engineering rules are derived from the simulation results in mapping a large number of potential active or inactive users generating each a small amount of traffic on the simulation scenarios. Within these ones only the active users are regarded, offering the same aggregate traffic with the same traffic characteristics as all potential users in the planning scenario.

What is needed for engineering the number of fixed PDCHs per cell under a given traffic load is a reliable relation between the mean downlink IP throughput performance and the offered traffic for different numbers of PDCHs.

This relation can be gained from the simulation results shown in Section 8.3.1. The mean downlink IP throughput per user over the offered IP traffic for different numbers of PDCHs is shown again in Figure 10.1 and is representing the proposed dimensioning graph for fixed PDCHs.

The traffic engineering rules for fixed PDCHs are defined by means of the following five steps:

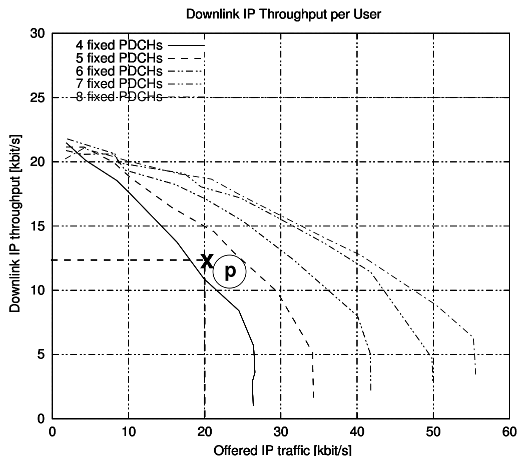


Figure 10.1: Dimensioning graph for fixed PDCH configurations

Step 1: Define the desired QoS (here in terms of average downlink throughput)

Step 2: Estimate the number of users per cell

Step 3: Define the offered IP traffic per user and calculate the offered traffic per cell

Step 4: Establish the operating point p defined by the desired user performance (QoS) on the y-axis and the total offered traffic on the x-axis and choose the adequate PDCH curve in the dimensioning graph as the next that lies above this operating point p.

To visualize this procedure, the following example is regarded. With a desired downlink IP throughput of 12.5 kbit/s, a given offered traffic per user of 90 kbyte/h, and 100 potential users per radio cell, the needed number of PDCHs can be estimated by the following four steps:

Step 1: The mean downlink IP throughput of 12.5 kbit/s is the QoS requirement.

Step 2: Mean expected number of GPRS users = 100

Step 3: Offered traffic/user = 90 kbyte/h = 0.2 kbit/s
total offered traffic = $100 \cdot 0.2$ kbit/s = 20 kbit/s

Step 4: Establish the operating point p (x = 20 kbit/s, y = 12.5 kbit/s). The PDCH curve in the dimensioning graph as the next that lies above is representing 5 fixed PDCHs.

As the result of the dimensioning procedure, a scenario with 5 fixed PDCHs is able to carry the estimated offered traffic, while realizing the QoS target.

10.1.2 Traffic Engineering Rules for On-demand PDCHs

Simulation results have shown that the performance for GPRS-based services does not increase dramatically, when a few fixed instead of purely on-demand PDCHs are used (see Section 8.3.4). If the cell capacity is dimensioned with a low blocking probability, e.g., 1%, for speech services, the use of on-demand PDCHs makes sense. This is plausible, since the probability that with, e.g., a trunk of 21 speech channels more than 2 channels are free, which temporarily can be used for GPRS, is around 90%.

An accurate traffic engineering approach for on-demand configurations can be performed in taking a dimensioning graph in accordance to GPRS performance results for an existing cell scenario with on-demand PDCHs as the basis and by defining the acceptable coexisting CS traffic (corresponding to a given P_b value) so that the GPRS performance for a given offered IP traffic can be guaranteed. This can be done according to the following five steps with a QoS target of 8 kbit/s and the same traffic prognosis as in Section 10.1.1:

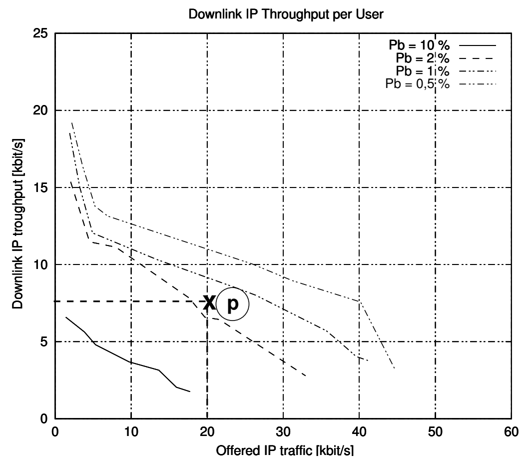


Figure 10.2: Dimensioning graph for on-demand PDCH configurations (3 TRX)

Step 1: Estimate P_b for the related TRX scenario for the CS traffic, which is expected for the scenario considered.
Here $P_b = 2\%$ is assumed.

Step 2: Mean expected number of GPRS users = 100

Step 3: Calculate the total offered traffic per cell:
offered traffic/user = 90 kbyte/h = 0.2 kbit/s
total offered traffic = 100 · 0.2 kbit/s = 20 kbit/s

Step 4: Define the desired average user data rate the operator wants to guarantee (here 8 kbit/s).

Step 5: Establish the operating point p defined by the QoS target on the y-axis and the total offered traffic on the x-axis and choose the adequate P_b curve as the next that lies above this operating point p. Here p equals (x = 20 kbit/s, y = 8 kbit/s).

If the offered CS traffic corresponding to this P_b is predicted to be exceeded, a new TRX should be added. Here p is below the curve with a P_b of 1%. This means that the coexisting CS traffic corresponding up to a P_b of 1% is acceptable for this scenario. Since a P_b of 2% is assumed and the operating point is above the curve for $P_b = 2\%$, an additional TRX would be needed for the radio cell in this example.

10.1.3 Remarks on Mixed Configurations with Fixed and On-demand PDCHs

To be able to guarantee the availability of GPRS-based services an operator might provide one, two or more fixed PDCHs and might provide the rest (up to eight) as on-demand PDCHs. In this case the results from the pure on-demand configurations can be taken as a worst case estimation, since the probability that the first PDCHs are used by CS traffic is quite low (below 10%) [Stuckmann and Müller (2000)].

If four or more fixed PDCHs are provided and the rest (up to eight) are on-demand, there are two possibilities. If an additional TRX is introduced to reduce the blocking probability for CS calls, the on-demand channels will not be highly utilized. Here the capacity can be estimated taking a configuration with eight fixed PDCHs. If no further TRX is provided, the utilization of the on-demand PDCHs will be very high. In this case the capacity can be estimated by a configuration with the allocated number of fixed and no on-demand PDCHs.

10.2 GPRS/EDGE Traffic Management and QoS Support

As explained in Section 2.2.2, today's cellular packet radio networks based on GPRS and EGPRS only include very basic MAC scheduling algorithms based on simple round robin and only realize best-effort services without QoS support for different applications and subscribers with their specific QoS requirements. Admission control to avoid overload situations is also not realized.

While today's GPRS scheduler designs are realizing a logical split between RLC and MAC, the consideration of information from the RLC layer for radio block scheduling on MAC level is proposed. Additionally the usage of information on the actual link quality for adaptive scheduling is considered to increase performance.

In the next step scheduling algorithms for the support of different traffic classes are introduced. Both priority scheduling and fair scheduling approaches are extended by adaptive scheduling concepts. The interworking with admission control in the SGSN is discussed and proposals for parameterization of the weights, defined in the MAC scheduler, and the maximum number of admitted flows for each traffic class in the admission control entity are presented.

10.2.1 Advanced Scheduling Algorithms for Best-effort Services

To support best-effort services all TBFs served by one base station are stored in one TBF queue. They are usually served by a RR strategy (see Section 3.4.5.5.2) to ensure a fair service for all active sessions. In this thesis two extensions of the RR scheduling strategy is proposed:

- the use of acknowledge state information of RLC blocks in the RLC layer (see Section 3.4.5.5.3) for TBF scheduling (*Displaced Pending Acknowledge Round Robin* (DPARR))
- the use of information on the actual link quality for adaptive TBF scheduling (*Link Quality-based Deficit Weighted Round Robin* (LQDWRR))

10.2.1.1 Displaced Pending Acknowledge Round Robin (DPARR)

This scheduling scheme schedules TBFs that have RLC blocks to transmit with the acknowledge state `NACK` always ahead of TBFs that have only RLC blocks with acknowledge state `PENDING_ACK`. With this modification it is avoided that `PENDING_ACK` RLC blocks are transmitted even if other TBFs have `NACK` blocks to transmit that are more urgent. Both the system throughput and the user throughput performance can be significantly increased.

The mechanism is consistent with the GPRS standard, since RLC and MAC need not necessarily be implemented decoupled and the MAC scheduling of TBFs for RLC data block transmission is not explicitly specified in the standard.

10.2.1.2 Link Quality-based Deficit Weighted Round Robin (LQDWRR)

LQDWRR is based on the scheduling algorithm *Deficit Weighted Round Robin* (DWRR) as proposed in [Shreedhar and Varghese (1995)]. The original aim was to assign capacity to class queues accurately and to provide nearly perfect fairness in terms of throughput. Fairness is guaranteed, even if the packet flows contain data packets of different length. This algorithm is extended by the preferred service of TBFs with good channel quality. Fair scheduling of backlogged TBFs is ensured by a deficit counter that adapts the service share dynamically.

The link quality is represented by the optimal *Modulation and Coding Scheme* (MCS), which is reported by the EGPRS *Link Quality Control* (see Section 4.3.1). In case the MCS is below a predefined value MCS_{LIMIT}, the quality of the link is classified as "*bad*", otherwise it is classified as "*good*". A *scheduling state variable* $SV_i(n-1)$ is introduced that represents one of three possible states:

Normal (**N**): The considered TBF_i was able to send data during the last scheduling cycle and the radio link quality was *good*:
$MCS_i[n-1] > MCS_{LIMIT}$

Backlogged (**B**): The regarded TBF_i was not able to send data within the previous scheduling cycle and the reported optimal MCS was below the limit:
$MCS_i[n-1] < MCS_{LIMIT}$

Lagged (**L**): The according TBF_i was backlogged for the maximum duration and had a service lag of Q_{max} and therefore was allowed to send data within the previous scheduling cycle, although the reported optimal MCS was below the limit:
$MCS_i[n-1] < MCS_{LIMIT}$

The deficit counter indicates the number of blocks to be transmitted in the scheduling cycle. It is increased by the RR quantum after each scheduling cycle and is decreased by 1 for each transmitted radio block. In this way backlogged TBFs keep the RR quantum for each scheduling cycle for the following one.

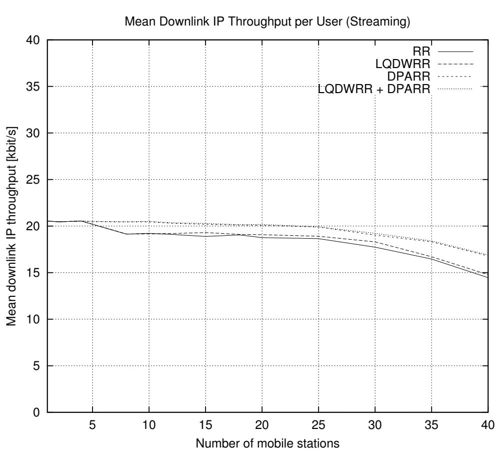
10.2.1.3 Performance Analysis

In this section the performance of the proposed best-effort scheduling algorithms is evaluated and compared to the basic RR algorithm. The scenario is characterized by an EGPRS radio cell with four fixed PDCHs, cluster size 3 and cell radius of 300 m. The EDGE LQC functions LA and IR are used with an MCS_{LIMIT} of MCS-4 for the LQDWRR algorithm. A traffic mix of 10% Streaming, 62% e-mail and 28% WWW sessions (see Section 9.2) is assumed. 4 fixed PDCHs are assumed.

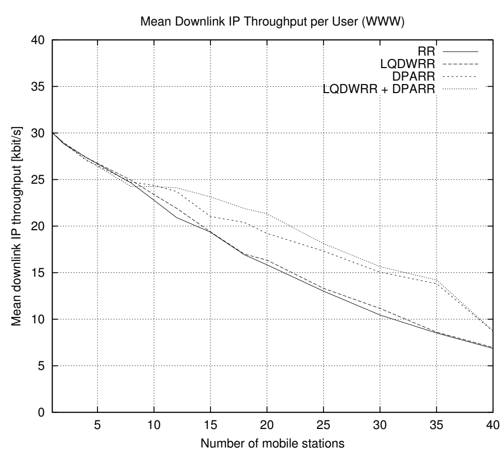
Figure 10.3 shows the throughput performance of Streaming and WWW sessions for the discussed scheduling algorithms over the number of mobile stations per cell. For Streaming traffic the offered data rate can be realized for up to a number of 25 stations with DPARR, whereas without this feature already 10 active stations lead to lower throughput performance. The use of LQDWRR does not show any significant performance gain for Streaming applications.

For WWW applications a performance difference through DPARR also becomes visible for 10 and more MSs. A performance gain of up to 50% is achieved, because in this scenario with low block error rates the transmission of `PENDING_ACK` RLC blocks is not significantly contributing to the user throughput and represents a waste of resources. This problem is overcome with the DPARR algorithm (see Section 10.2.1.1).

With LQDWRR a performance gain can only be achieved together with DPARR, because the RLC protocol without the DPARR feature is less sensitive to differences in link quality. The low performance gain through LQDWRR can be explained by the small sensitivity of the performance of TCP-based applications to an increase in radio capacity. In EGPRS often high MCSs are chosen even in situations with lower C/I values and the throughput is still acceptable because of incremental redundancy (see Section 8.4.2). Additionally TCP timeouts occur for backlogged

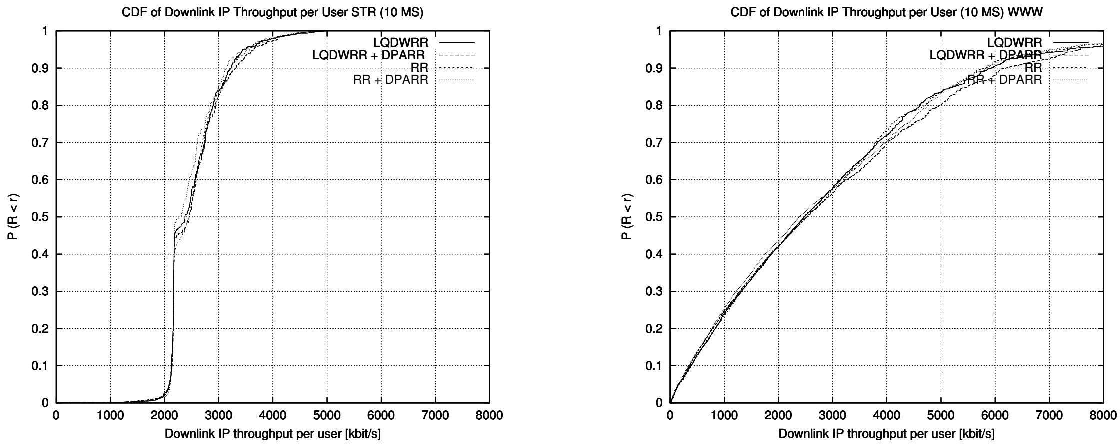


(a) Mean downlink IP throughput per user (Streaming)



(b) Mean downlink IP throughput per user (WWW)

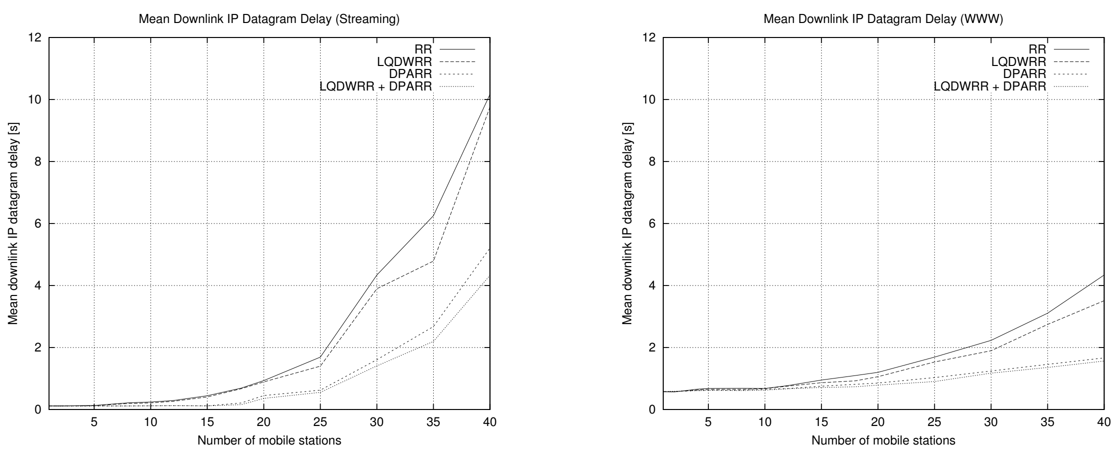
Figure 10.3: Throughput performance for different best-effort scheduling algorithms



(a) CDF of downlink IP throughput per user (15 MS, Streaming)

(b) CDF of downlink IP throughput per user (15 MS, WWW)

Figure 10.4: Throughput distributions for different best-effort scheduling algorithms



(a) Mean downlink IP datagram delay (Streaming)

(b) Mean downlink IP datagram delay (WWW)

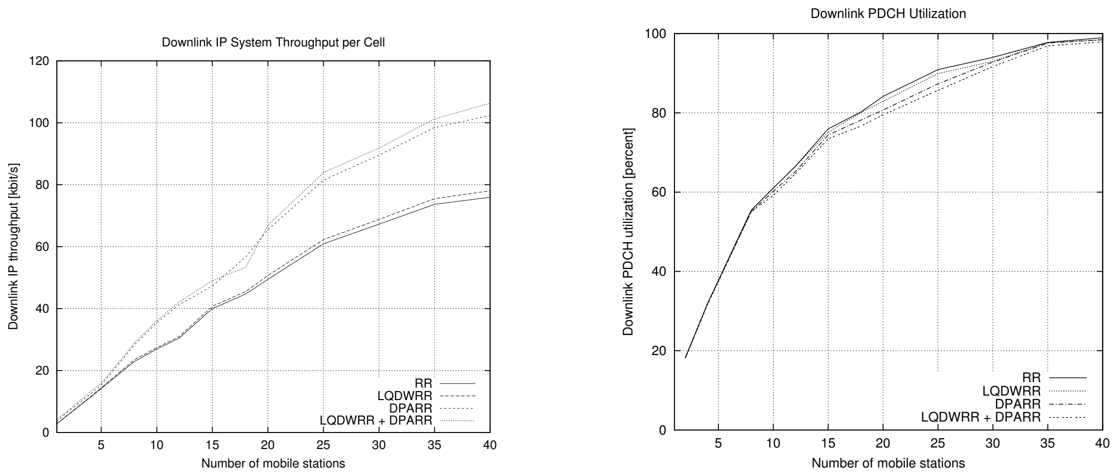
Figure 10.5: Delay performance for different best-effort scheduling algorithms

stations, which limits the performance gain achievable with LQ-based scheduling.

In Figure 10.4 the throughput distribution functions for the discussed best-effort algorithms for 15 MSs is shown. While no significant effect can be observed for Streaming (see Figure 10.4(a)), a slightly better performance is achieved by both DPARR and LQDWRR in the range of medium to high throughput values for WWW (see Figure 10.4(b)).

Similar conclusions can be drawn from the delay performance of the different best-effort scheduling algorithms (see Figure 10.5). The delay of IP datagrams carrying Streaming traffic starts to increase significantly with 25 stations with DPARR and with 15 MSs without this feature. For WWW a delay of less than 2 s for 40 MSs is achieved using DPARR, whereas the IP datagram delay without the DPARR feature exceeds 2 s already with 30 MSs.

In Figure 10.6 the system measures under the influence of the proposed best-effort scheduling algorithms are presented. The system throughput per cell can be increased from 80 to more than



(a) Downlink IP system throughput per cell (b) Downlink PDCH utilization

Figure 10.6: System measures for different best-effort scheduling algorithms

100 kbit/s with DPARR. The PDCH utilization is not influenced by the scheduling algorithm, since channels are utilized as long as TBFs are active and the number of stations with active TBFs does not heavily depend on the scheduling algorithm.

10.2.2 QoS Support for Traffic Classes

For traffic class scheduling a priority scheduling algorithm (*Priority Queueing* (PQ)) is compared with bandwidth sharing (*Deficit Weighted Round Robin* (DWRR)) [Stuckmann and Möller (2003)] in a scenario with 4 fixed PDCHs. Using the priority algorithm a traffic class queue is only served if all queues of higher priority are empty. This can lead to poor performance and high session blocking rates for low priority classes, if the traffic load for higher priority classes is too high. To guarantee a certain minimum capacity for all classes the DWRR algorithm can be used [Shreedhar and Varghese (1995)]. In this scenario the traffic class weights have been assigned to 58% for Streaming (Video Streaming), 35% for Interactive (WWW) and 7% for Background (e-mail), which is related to the predicted offered traffic per traffic class and the QoS target to be achieved for the classes.

For *Connection Admission Control* (CAC) at the SGSN two strategies are considered: *Soft CAC*, where a larger number of sessions with higher priorities are admitted, and a *Hard CAC*, where only a smaller number of sessions with high priorities are admitted to be able to guarantee the QoS for all traffic classes. The maximum number of allowed sessions per available PDCH are shown in Table 10.1.

Table 10.1: CAC Parameterization

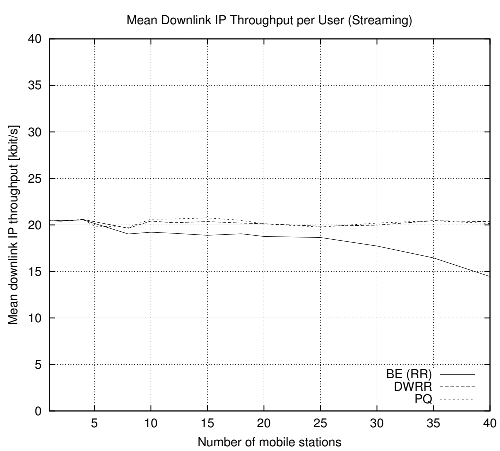
Traffic class	Hard CAC	Soft CAC
Streaming sessions	1 per PDCH	2 per PDCH
Interactive sessions	2 per PDCH	3 per PDCH

Figure 10.7(a) and Figure 10.7(b) show the throughput and delay performance of Streaming applications over the number of mobile stations for the same scenario as of Section 10.2.1.3. Both traffic class scheduling algorithms DWRR and PQ are able to serve Streaming traffic sufficiently for the whole load range, contrary to the best-effort (BE) service. The mean Streaming data rate can be realized over the whole load range, even if the IP datagram delay for Streaming starts

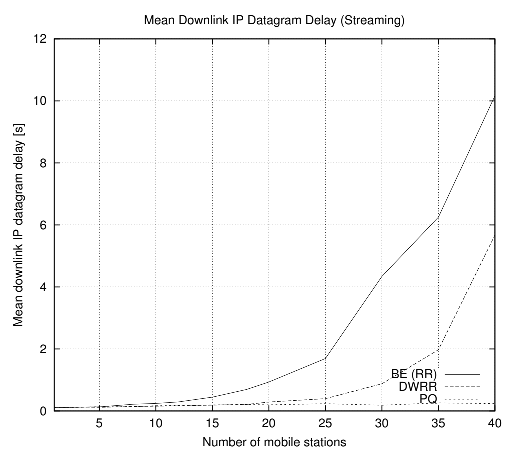
to increase with 30 active MSs using DWRR. While PQ serves Streaming traffic always ahead of WWW and e-mail traffic, DWRR still serves WWW and e-mail in some scheduling periods that are reserved for these lower prioritized classes.

As already explained in Section 10.2.1.3 the throughput performance of Streaming applications for best-effort scheduling starts to decrease with 10 active stations, because WWW and e-mail sessions have the same share of resources as Streaming. Therefore the IP datagram delay for Streaming starts to increase with 10 stations and is exceeding 4 s with 30 stations.

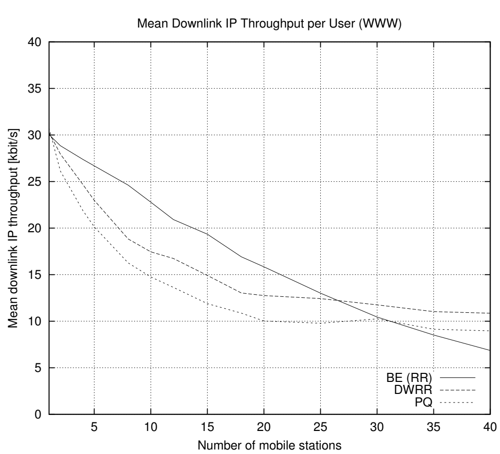
The throughput performance of the traffic classes with lower priority is shown in Figure 10.7(c) and Figure 10.7(d). In situations with low and medium traffic load the mean IP throughput per user for both WWW and e-mail is decreasing faster for PQ and DWRR than for best-effort scheduling, because traffic classes with higher priority are served ahead. For e-mail the throughput already reaches a minimum performance of around 10 kbit/s with 10 stations for PQ. This decrease for relatively low traffic load occurs, because both WWW and Streaming are higher prioritized than e-mail. With DWRR this minimum can be shifted to a higher traffic load of 18 stations, since e-mail still gets a small part of the capacity, even if Streaming and WWW users are active.



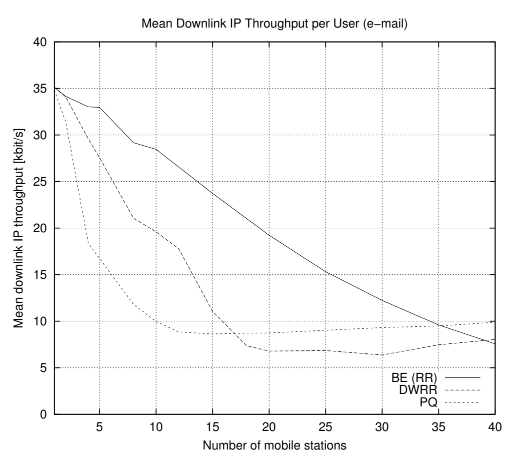
(a) Mean downlink IP throughput per user (Streaming)



(b) Mean downlink IP datagram delay (Streaming)

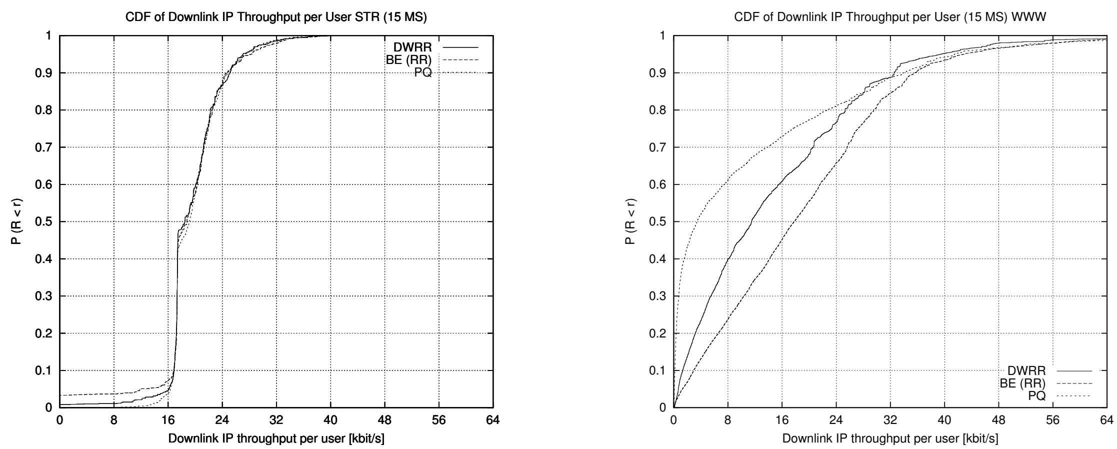


(c) Mean downlink IP throughput per user (WWW)



(d) Mean downlink IP throughput per user (e-mail)

Figure 10.7: Performance of different traffic class scheduling algorithms



(a) CDF of downlink IP throughput per user (15 MS, Streaming)

(b) CDF of downlink IP throughput per user (15 MS, WWW)

Figure 10.8: Throughput distributions for different traffic class scheduling algorithms

In Figure 10.8 the throughput distributions for Streaming and WWW under the influence of the different traffic class scheduling algorithms are shown. For Streaming applications no influence can be observed in the higher throughput range. Some more throughput values are occurring in the smaller throughput range for RR compared to DWRR and PQ. Since for RR Streaming applications are not prioritized there is a certain number of time periods where a lot of WWW and e-mail traffic is offered and the offered Streaming traffic can not be served sufficiently. This effect can be prevented completely with PQ and partly with DWRR.

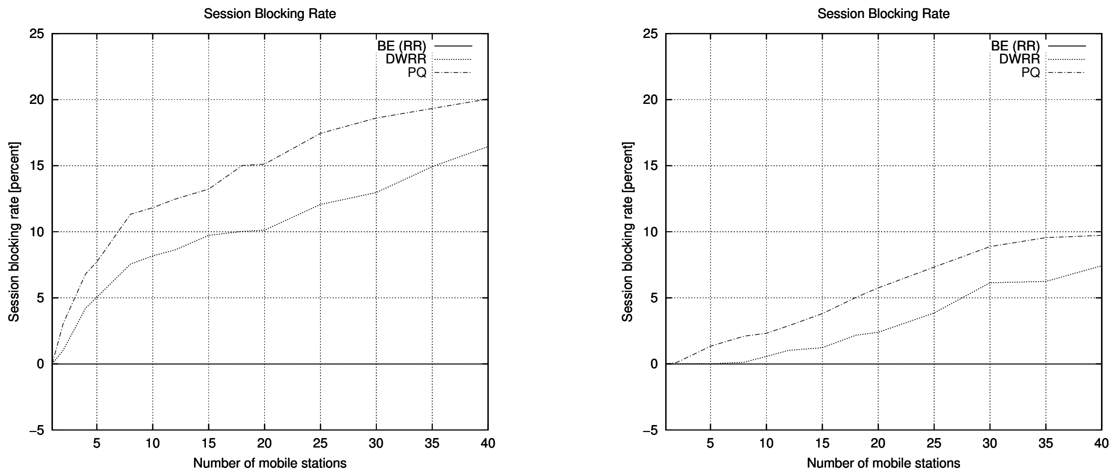
The influence of the scheduling algorithm becomes more significant, when WWW traffic is regarded (see Figure 10.8(b)). While the distribution function for RR is increasing very fast already in the range of very small throughput values below 5 kbit/s to a probability of 0.5, a better performance can be achieved with DWRR. However, 50% of the throughput values are still below 12 kbit/s for DWRR, whereas for a best-effort service with simple RR scheduling without any prioritization of Streaming traffic at least 18 kbit/s can be reached for WWW during 50% of the time.

An interesting effect occurs in situations with high traffic load. Here, the throughput performance for both PQ and DWRR does not continue to decrease and even reaches a better performance than the best-effort scheduling algorithm. This can be explained with high blocking rates (see Figure 10.9). With an increasing number of stations the probability that WWW and e-mail sessions are terminated at an early stage because of poor performance increases. If a notable portion of sessions are terminated, the offered traffic per station is decreasing. Therefore sessions with good performance still contribute to the evaluated throughput, while sessions with bad performance become shorter and are contributing to a smaller extent.

Although e-mail is the traffic class with the lowest priority, the blocking rates for e-mail sessions is smaller compared to the blocking rate for WWW sessions. The reason is that e-mail sessions are much shorter and therefore the probability for blocking in an e-mail session is lower than for WWW. For WWW the blocking rate reaches 20% for PQ and 16% for DWRR, whereas the blocking rate for e-mail does not exceed 10%.

Even if the hard CAC policy is used blocking can not be avoided, since sessions that are not admitted for the requested QoS are degraded to lower traffic classes. However, the blocking rate can be slightly reduced, since traffic classes with lower priority are served more often [Stuckmann and Möller (2003)]. For this scenario no effect of the CAC policy could be observed, because the traffic offered by the high-priority traffic class is relatively low.

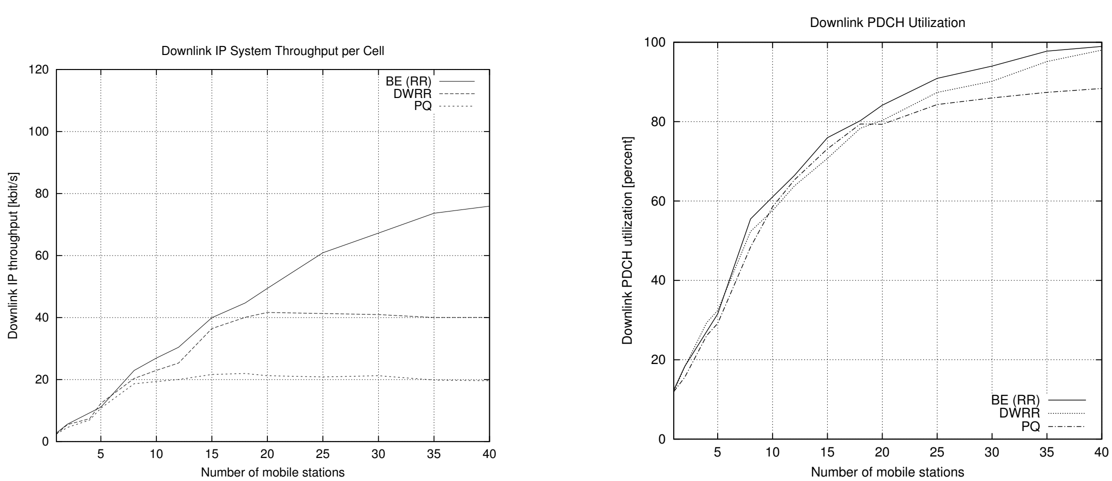
Session blocking is one reason for the decrease in system throughput, when traffic class scheduling is applied (see Figure 10.10(a)). While the system throughput for best-effort scheduling reaches



(a) WWW session blocking rate

(b) e-mail session blocking rate

Figure 10.9: Session blocking rates for different traffic class scheduling algorithms



(a) Downlink IP system throughput per cell

(b) Downlink PDCH utilization

Figure 10.10: System measures for different traffic class scheduling algorithms

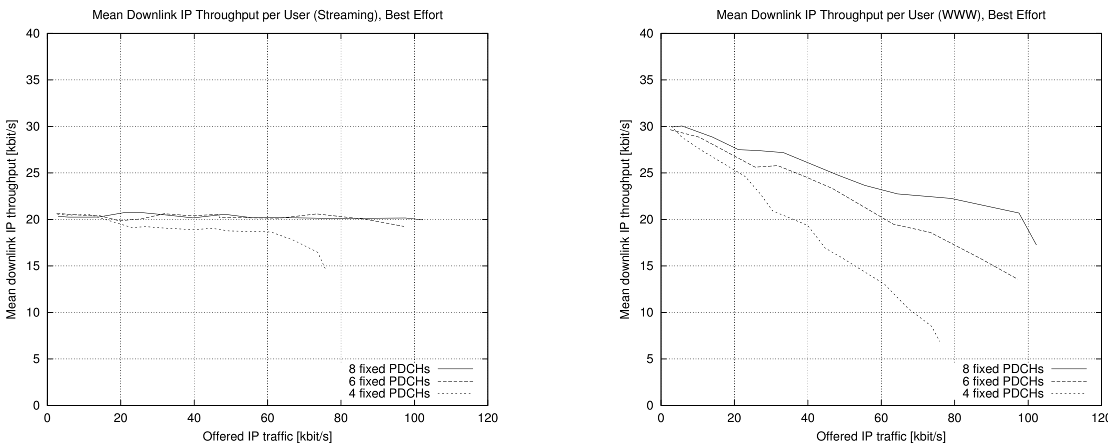
nearly 80 kbit/s, only 50% is realized with DWRR and only 25% with PQ. For PQ the PDCHs are not completely utilized (86% for 40 stations), which also can be explained by high blocking rates leading to fewer active TBFs (see Figure 10.10(b)).

Another reason for the low maximum system throughput is the TCP-unfriendly behavior of the scheduling algorithms. TCP reacts with low throughput performance to interruptions caused by unfair scheduling. When using PQ, frequent changes in the available bandwidth for TCP connections occur resulting in frequent TCP timeouts. DWRR increases the fairness and the TCP-friendliness of the scheduling algorithm and therefore is able to double the maximum system throughput compared to PQ. If all flows are treated equally as in the best-effort algorithm, the system throughput is maximized.

10.3 Mutual Dependency of Traffic Management and Traffic Engineering

In Section 10.1 traffic engineering rules for best-effort services, as they are realized in today's cellular packet radio networks, are presented. If these rules are applied for several applications all QoS targets of all applications have to be met [Hoymann and Stuckmann (2002b)]. One application determines the acceptable traffic or the resources that are needed representing the critical application (see Section 2.2.3).

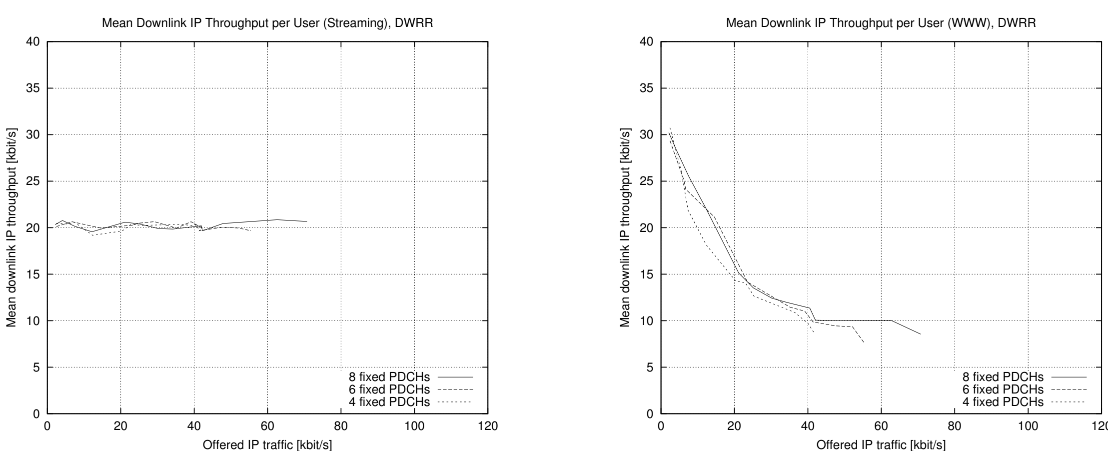
To visualize this phenomenon the EDGE introduction scenario, which has already been regarded in Section 9.2 and Section 10.2, is discussed. Figure 10.11 shows the dimensioning graphs of Streaming and WWW for a best-effort service. For Streaming applications the capacity limit is reached once the throughput or delay performance significantly decreases. Without QoS support this is the case with an offered traffic of 20 kbit/s with 4 PDCHs (see Figure 10.11(a)), whereas for 6 and 8 PDCHs the performance is acceptable over the whole load range. Assuming that the QoS



(a) Throughput performance without QoS support (Streaming)

(b) Throughput performance without QoS support (WWW)

Figure 10.11: Dimensioning graphs for a best-effort service



(a) Throughput performance with QoS support (Streaming)

(b) Throughput performance with QoS support (WWW)

Figure 10.12: Dimensioning graphs for QoS support (DWRR)

target for WWW applications is significantly lower than a mean throughput of 25 kbit/s, Streaming represents the critical application for the scenario with 4 PDCHs (see Figure 10.11(b)).

The purpose of traffic management mechanisms supporting different service classes is the performance improvement for the critical application, while taking a loss for applications with lower priorities.

In Section 10.2.2 it has been shown for an example with Streaming, WWW and e-mail that the performance of the high-priority class Streaming can be improved at the expense of WWW and e-mail. If operators decide to introduce functions for QoS support in their networks, this effect of traffic management on traffic engineering rules should be taken into account for the network planning process.

The effect is visualized in Figure 10.12, where the dimensioning graphs under the influence of the DWRR traffic class scheduling algorithm are shown for Streaming and WWW.

With the DWRR algorithm Streaming can be realized over the whole load range up to an offered traffic of 40 kbit/s with 4 PDCHs. On the other hand the throughput performance for WWW increases significantly. For an offered traffic of 20 kbit/s the mean throughput falls below 15 kbit/s. Additionally the blocking rate increases for lower-prioritized applications as discussed in Section 10.2.2. Even with a fair scheduling algorithm like DWRR, frequent changes in capacity that is available for TCP connections used by WWW and e-mail are occurring. The reason for this is that Streaming applications are assigned a high capacity share and can preempt WWW and e-mail traffic up to certain extent. This increasingly leads to TCP timeouts and to session termination carried out by unsatisfied users. Therefore it is difficult to realize reasonable QoS targets for lower-prioritized classes in addition to the stringent QoS targets of high-prioritized classes.

Since variable prices can be established for different applications and user groups, the decision for or against the prioritization of applications and users depends on the operator's strategy that can be aimed at the satisfactions of as many customers as possible or at the prioritized service of costumers contributing most to the revenue.

Between these distinct strategies it is possible to configure the functions for QoS management considering the pricing for applications or customer groups. With this configuration a large number of high-priority customers can be satisfied and simultaneously the demands of a number of customers with low priorities can be met. This configuration can be done by adapting the weights in the traffic class scheduler and the threshold for the number of admissible sessions per PDCH in the CAC entity (see Section 10.2.2).

CHAPTER 11

Conclusions

The goal of this thesis is to develop dimensioning concepts for cellular packet radio networks that will remain valid during network evolution. Two important requirements for these concepts are practical applicability and accuracy. While the traffic engineering rules themselves should be simple and only consider the user number and traffic volume during the busy hour, they should be based on accurate models for the protocol stacks, the traffic pattern and the radio channel that ought to be close to reality. The key tasks to achieve this aim are the identification and development of adequate traffic models for existing and future mobile applications, the prototypical implementation of the GPRS/EDGE protocol stack as well as the integration of optimized methods for QoS support, so that the presented concepts will be inline with the evolution of the radio interface protocols.

For existing and future mobile applications, which are predicted for the next years, traffic models are proposed. For WWW and e-mail applications, which are already popular today in fixed networks, adequate traffic models from literature have been identified and adapted for the applicability in mobile environments. Traffic models for applications that are emerging for mobile networks like WAP and MMS have been derived from measurement in the framework of this thesis. Finally traffic models for Streaming applications have been identified in following the recommendations of standardization bodies that are defining the standards for mobile Audio and Video applications. From these models an integrated Internet and Multimedia load generator was implemented, which enables accurate modeling of all relevant traffic types in configurable traffic mixes.

To be able to study the existing and evolved systems in their natural environments with the appropriate radio coverage, mobility and typical traffic volumes and to analyze own approaches in the improvements and introduction of new features, the emulation tool GPRSIM has been developed by the author. The complete GPRS/EDGE protocol stack based on the actual standard has been formally specified. Additionally the load generators and the underlying TCP/IP protocol stacks have been integrated and finally adequate channel and mobility models have been implemented and integrated. To ensure that the results of the GPRSIM can be regarded as representative a validation by analysis and measurement for simplified scenarios has been carried out.

In addition to this simulative approach the GPRS system has also been evaluated analytically. With the analytical results a fast estimation of the general system capabilities can be performed. On the other hand the examinations show the limitations of state-of-the-art analytical approaches, when complex scenarios with heavy-tailed traffic sources, the dynamic behavior of TCP and radio interface-specific mechanisms like scheduling and link adaptation have to be considered.

Considering these limitations in analytical modeling the work was concentrated on a comprehensive simulative performance evaluation of GPRS and EGPRS for relevant scenarios, protocol options and predicted applications. The following main insights have been gained in this analysis.

With GPRS and fixed PDCHs the throughput performance is limited to 22 kbit/s for realistic file sizes, while a maximum system throughput of 56 kbit/s can be reached. Besides the block errors on the radio channel, the protocol overhead of the layers RLC/MAC, LLC and SNDCP and the delay arising with TBF setup and RLC/MAC acknowledgement messages, further limitations arise with the TCP flow control mechanisms. While in situations with low traffic load the performance is limited by TCP slow start, the system throughput is limited by TCP congestion avoidance, since the offered traffic per TCP connection is adapted to the available capacity per connection.

GPRS performance shows low sensitivity to block errors. Compared to typical radio coverage in GSM networks a performance difference of between 10% and 15% in throughput was observed in worst-case and best-case scenarios. The reason for the low sensitivity is the use of CS-2, that is a relatively robust coding scheme, and the robust RLC scheme in GPRS which is efficient for both uncorrelated errors and error bursts up to a burst length in the order of the RLC window size.

The performance loss for GPRS if on-demand PDCHs are used, depends on the coexisting circuit-switched traffic in the regarded radio cell. The influence of coexisting voice traffic becomes only significant, if very high voice traffic, e.g., corresponding to a P_b of 10%, is regarded. The main result is that GPRS remains available with small QoS losses even if there are interceptions enforced by speech connections and the PDCHs are only available on demand. Besides the reason of a high probability that a number of channels remains available for GPRS, the flexibility of GPRS has to be mentioned. TBFs can be redistributed on other PDCHs that are still available. If PDCHs are detracted for a certain period, TBFs can share remaining resources and can utilize resources again once they are reallocated after the circuit-switched connection release.

For the regarded traffic mix, a throughput performance gain of up to 100% compared to GPRS is reachable with the new *Modulation and Coding Schemes* (MCSs) in EGPRS. A maximum throughput performance of about 47 kbit/s is reachable for the traffic mix considered. While this gain in user throughput performance is not dramatic, the advantage of EDGE is mainly the higher capacity per PDCH compared to GPRS, since more than 40 active MSs can easily be served with 8 PDCHs maintaining good performance. Additionally *Link Adaptation* (LA) and *Incremental Redundancy* (IR) are enabling performance gains in comparison to a fixed chosen MCS. Although the gain through LA, in comparison with a well chosen fixed MCS, is not dramatic, LA should be used in EGPRS networks. The reason is that without LA, a fixed MCS would have to be chosen for every site. This cannot simply be realized in realistic radio coverage planning processes.

Further insights were gained by the examination of the most important mobile applications over GPRS and EGPRS. It has been shown that WAP traffic can be multiplexed seamlessly with Internet traffic because of the small and limited WAP deck size, while Internet traffic slightly slows down WAP traffic in situations with high traffic load. Regarding Streaming applications over EGPRS it has been found that the throughput performance remains acceptable for up to 20 stations offering Streaming traffic. In traffic mix scenarios with Streaming traffic and TCP-based applications, EGPRS can easily serve 40 stations. In GPRS scenarios, Streaming applications can only be realized in situations with low traffic load.

In addition to the comprehensive performance analysis, traffic management techniques for best-effort services and traffic class support have been developed and evaluated. For a best-effort scheduler design the DPARR algorithm can be recommended, since it is easy to implement and has a great effect on the throughput performance especially in situations with medium and high traffic load. Adaptive scheduling can be implemented in addition to realize an optimized scheduler that has no great effect in normal conditions where all stations have sufficient coverage, but can avoid the waste of capacity by serving stations with very bad channels. For the introduction of quality of service support, several scheduling and admission control strategies are introduced. If operators want to maximize the performance of prioritized subscribers or applications and session blocking for lower traffic classes is introduced, a priority scheduling algorithm should be applied together with hard CAC. If a certain capacity should be guaranteed also for lower traffic classes, a bandwidth sharing scheme should be implemented. The CAC strategy should then be chosen depending on the guarantee the operator wants to offer for prioritized users and on the offered traffic predicted for each service class.

To enable the practical applicability of these results, dimensioning rules have been derived from the simulation output data and examples for the application of these rules are presented for GPRS introduction and evolution scenarios. The dimensioning graph for configurations with fixed PDCHs is based on simulation results for different GPRS loads and different radio capacities. With this dimensioning graph the relationship between desired QoS, offered traffic and required radio capacity can be estimated. For on-demand channel configurations dimensioning graphs comprising the simulation results for different circuit-switched loads arising from GSM speech traffic can be used to determine, if a new TRX has to be installed in a given radio cell to serve the GPRS traffic with a certain QoS target. Finally the mutual dependency of traffic management and traffic engineering is outlined by comparing dimensioning graphs for best-effort services with graphs for scenarios in which traffic management techniques are applied.

Summarizing the contribution of this thesis traffic engineering guidelines have been developed that are accurate and usable in practice. The most important effects that occur for different traffic characteristics are covered in the performance analysis. While the simulation results can be used as a basis for accurate GPRS/EDGE capacity planning, the general concepts and effects presented in this thesis are valid for all cellular packet radio networks. The concepts have been developed in collaboration with several network operators and have been integrated in the planning process by many leading GSM operators world-wide. Further, a comprehensive capacity planning tool has been established that has already been applied by several GSM operators and that presently is further developed to a commercial planning tool. With this tool additional scenarios can be examined in detail und the results can be further integrated in the planning process following the concepts presented in this thesis. The algorithms developed for performance optimization and for QoS management, especially the DPARR algorithm, have proven to be very promising. Since the algorithms are easy to implement, they represent interesting proposals for the further development of network components for all manufacturers. Based on the results provided in this thesis both operators and manufacturers are in a position to estimate the performance and capacity gain of these or similar algorithms.

LIST OF ABBREVIATIONS

2G	*Second-Generation*
3G	*Third-Generation*
3GPP	*3rd Generation Partnership Project*
3GPP2	*3rd Generation Partnership Project 2*
8-PSK	*8-Phase-Shift-Keying*
AAC	*Advanced Audio Coding*
ABM	*Asynchronous Balanced Mode*
ABQP	*Aggregate BSS QoS Profile*
AC	*Admission Control*
ACELP	*Algebraic Codebook Excited Linear Predictive*
ACK	*Acknowledgement*
ADM	*Asynchronous Disconnected Mode*
ARIB	*Association of Radio Industries and Businesses*
ARQ	*Automatic Repeat Request*
ATM	*Asynchronous Transfer Mode*
BCCH	*Broadcast Control Channel*
BCS	*Block Check Sequence*
BEC	*Backward Error Correction*
BER	*Bit Error Ratio*
BLEP	*Block Error Probability*
BLER	*Block Error Ratio*
BS	*Base Station*
BSC	*Base Station Controller*
BSN	*Block Sequence Number*
BSS	*Base Station Subsystem*
BSSGP	*Base Station Subsystem GPRS Protocol*
BTS	*Base Transceiver Station*
BVC	*BSSGP Virtual Connection*
BVCI	*BSSGP Virtual Connection Identifier*
CAC	*Connection Admission Control*
CBR	*Constant Bit Rate*
CCCH	*Common Control Channel*
CCDF	*Complementary Cumulative Distribution Function*
CDF	*Cumulative Distribution Function*
CDMA	*Code Division Multiple Access*
C/I	*Carrier-to-Interference*
CN	*Core Network*
CNCL	*Communication Networks Class Library*
CNG	*Comfort Noise Generator*
CRC	*Cyclic Redundancy Check*
CRTP	*Compressed RTP*
CS	*Circuit-Switched*
CS	*Coding Scheme*
CSD	*Circuit Switched Data*
CWTS	*Chinese Wireless Telecommunications Standards*
DECT	*Digital Enhanced Cordless Telecommunications*
DHCP	*Dynamic Host Configuration Protocol*
DLC	*Data Link Control*
DLCI	*Data Link Connection Identifier*
DPARR	*Displaced Pending Acknowledge Round Robin*
DRX	*Discontinuous Reception*
DTP	*Data Transfer Process*
DTX	*Discontinuous Transmission*
DWRR	*Deficit Weighted Round Robin*
ECSD	*Enhanced Circuit-Switched Data*
EDGE	*Enhanced Data rates for GSM Evolution*
EGPRS	*Enhanced General Packet Radio Service*
ETSI	*European Telecommunications Standards Institute*

FBI *Final Block Indicator*
FCS *Frame Check Sequence*
FDD *Frequency Division Duplex*
FDM *Frequency Division Multiplex*
FEC *Forward Error Correction*
FFM *Fluid-flow Model*
FPB *First Partial Bitmap*
FR *Frame Relay*
FSM *Finite State Machine*
FTP *File Transfer Protocol*
FUNET *Finnish University and Research Network*
GERAN *GSM/EDGE Radio Access Network*
GGSN *Gateway GPRS Support Node*
GIST *Graphical Interactive Simulation result Tool*
GMM *GPRS Mobility Management*
GMM *Global Multimedia Mobility*
GMSK *Gaussian Minimum Shift Keying*
GPRS *General Packet Radio Service*
GR *GPRS Register*
GSM *Global System for Mobile communication*
GSM *Groupe Spéciale Mobile*
GSN *GPRS Support Node*
GTP *GPRS Tunneling Protocol*
GUI *Graphical User Interface*
HDLC *High Level Data Link Control*
HLR *Home Location Register*
HSCSD *High-Speed Circuit-Switched Data*
HTML *Hypertext Markup Language*
HTTP *Hypertext Transfer Protocol*
IDU *Interface Data Unit*
IE *Information Element*
IEI *Information Element Identifier*
IETF *Internet Engineering Task Force*
IMAP *Internet Message Access Protocol*
IMSI *International Mobile Subscriber Identity*
IMT-2000 *International Mobile Telecommunications at 2000 MHz*
IP *Internet Protocol*
IPv4 *IP Version 4*
IPv6 *IP Version 6*
IR *Incremental Redundancy*
IRP *Interrupted Rate Process*
ISDN *Integrated Services Digital Network*
ISO *International Standards Organization*
ISP *Internet Service Provider*
ITU *International Telecommunication Union*
LA *Location Area*
LA *Link Adaptation*
LAN *Local Area Network*
LAPD *Link Access Procedure on the D-channel*
LDT *Large Data Transfer*
LLC *Logical Link Control*
LM *Location Management*
LPC *Linear Predictive Coding*
LQC *Link Quality Control*
LQDWRR *Link Quality-based Deficit Weighted Round Robin*
LSP *Linear Spectral Pair*
MAC *Medium Access Control*
MC *Markov Chain*
MC *MultiCommunicator*
MCS *Modulation and Coding Scheme*
MGCP *Media Gateway Control Protocol*
MM *Mobility Management*
MMAP *Marked Markovian Arrival Processes*
MMRP *Markov Modulated Rate Process*
MMS *Multimedia Messaging Service*
MP3 *MPEG Audio Layer-3*
MPDCH *Master Packet Data Channel*
MPEG *Moving Pictures Expert Group*

MP-MLQ	*Multi-Pulse Maximum Likelihood Quantization*
MS	*Mobile Station*
MSC	*Mobile-services Switching Center*
MSS	*Maximum Segment Size*
MTU	*Maximum Transfer Unit*
NACK	*Negative Acknowledgement*
N-CDMA	*Narrowband-CDMA*
NGN	*Next Generation Network*
NPB	*Next Partial Bitmap*
NRT	*Non Real-Time*
N-SAPI	*Network layer Service Access Point Identifier*
OSI	*Open Systems Interconnection*
PACCH	*Packet Associated Control Channel*
PAGCH	*Packet Access Grant Channel*
PBCCH	*Packet Broadcast Control Channel*
PBSN	*Partial Bitmap Sequence Number*
PC	*Personal Computer*
PCCCH	*Packet Common Control Channel*
PCM	*Pulse Code Modulation*
PDA	*Personal Digital Assistant*
PDC	*Personal Digital Cellular*
PDCH	*Packet Data Channel*
PDF	*Probability Density Function*
PDN	*Packet Data Network*
PDP	*Packet Data Protocol*
PDTCH	*Packet Data Traffic Channel*
PDU	*Protocol Data Unit*
PFC	*Packet Flow Context*
PFI	*Packet Flow Identifier*
PI	*Protocol Interpreter*
PILC	*Performance Implications of Link Characteristics*
PL	*Physical Layer*
PLL	*Physical Link Layer*
PLMN	*Public Land Mobile Network*
PNCH	*Packet Notification Channel*
POP3	*Post Office Protocol version 3*
PQ	*Priority Queueing*
PPCH	*Packet Paging Channel*
PRACH	*Packet Random Access Channel*
PS	*Packet-Switched*
PSDN	*Public Switched Data Network*
PTCH	*Packet Traffic Channel*
PTM	*Point-to-Multipoint*
PTM-M	*PTM Multicast*
Q10	*Quantization Level 10*
QCIF	*Quarter Common Intermediate Format*
QoS	*Quality of Service*
RA	*Routing Area*
RADIUS	*Remote Authentication Dial In User Service*
RAN	*Radio Access Network*
RB	*Reported Bitmap*
RBB	*Received Block Bitmap*
RBP	*Radio Block Period*
RED	*Random Early Detection*
RF	*Radio Frequency*
RLC	*Radio Link Control*
ROHC	*Robust Header Compression*
RR	*Round Robin*
RT	*Real-Time*
RTP	*Real-time Transmission Protocol*
RTSP	*Real-Time Streaming Protocol*
SABM	*Set Asynchronous Balance Mode*
SACK	*Selective Acknowledgement*
SAP	*Service Access Point*
SAPI	*Service Access Point Identifier*
SB	*Synchronization downlink Burst*
SDL	*Specification and Description Language*
SDU	*Service Data Unit*
SGSN	*Serving GPRS Support Node*
SID	*Silence Insertion Descriptor*
SIP	*Session Initiation Protocol*
SM	*Session Management*

SMIL	*Synchronised Multimedia Integration Language*
SMS	*Short Message Service*
SMTP	*Simple Mail Transfer Protocol*
SNDCP	*Sub-Network Dependent Convergence Protocol*
SNS	*Sequence Number Space*
SPEETCL	*SDL Performance Evaluation Tool Class Library*
SSA	*Services and System Aspects*
SSL	*Secure Socket Layer*
TA	*Terminal Aspects*
TA	*Timing Advance*
TBF	*Temporary Block Flow*
TCP	*Transport Control Protocol*
TD-SCDMA	*Time Division—Synchronous Code Division Multiple Access*
TDD	*Time Division Duplex*
TDM	*Time Division Multiplex*
TDMA	*Time Division Multiple Access*
TFI	*Temporary Flow Identity*
TFT	*Traffic Flow Template*
TIA	*Telecommunication Industry Association*
TID	*Tunnel Identifier*
TLLI	*Temporary Logical Link Identifier*
TLS	*Transport Layer Security*
TN	*Timeslot Number*
TRX	*Transmitter Receiver Unit*
TSG	*Technical Specification Group*
TTA	*Telecommunication Technology Association*
TTC	*Telecommunication Technology Committee*
UDP	*User Datagram Protocol*
UMTS	*Universal Mobile Telecommunication System*
URL	*Unified Resource Locator*
USF	*Uplink State Flag*
USSD	*Unstructured Supplementary Service Data*
UTRAN	*UMTS Terrestrial Radio Access Network*
UWC	*Universal Wireless Communications*
VAD	*Voice Activity Detection*
VBR	*Variable Bit Rate*
VoIP	*Voice over IP*
VQEG	*Video Quality Expert Group*
WAE	*Wireless Application Environment*
WAN	*Wide Area Network*
WAP	*Wireless Application Protocol*
WBXML	*WAP Binary Extensible Markup Language*
W-CDMA	*Wideband Code Division Multiple Access*
WDP	*Wireless Datagram Protocol*
WFQ	*Weighted Fair Queueing*
WP-TCP	*Wireless Profiled TCP*
WS	*Window Size*
WSP	*Wireless Session Protocol*
WTA	*Wireless Telephony Application*
WTLS	*Wireless Transport Layer Security*
WTP	*Wireless Transaction Protocol*
WML	*Wireless Markup Language*
WWW	*World Wide Web*
XML	*Extensible Markup Language*

3GPP TSG CN. *Handover Procedures (3GPP TS 23.009 version 5.0.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, December 2001. Cited on p. 39.

3GPP TSG CN. *Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Mobile Station (MS) – Serving GPRS Support Node (SGSN); Logical Link Control (LLC) layer specification (3GPP TS 44.064 version 5.1.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, March 2002a. Cited on p. 25.

3GPP TSG CN. *Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Mobile Station (MS) – Serving GPRS Support Node (SGSN); Subnetwork Dependent Convergence Protocol (SNDCP) (3GPP TS 44.065 version 5.0.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, June 2002b. Cited on p. 25.

3GPP TSG CN. *General Packet Radio Service (GPRS); GPRS Tunneling Protocol (GTP) across the Gn and Gp Interface (3GPP TS 29.060 version 5.0.1 Release 5)*. Technical Specification, 3rd Generation Partnership Project, January 2002c. Cited on p. 23, 45.

3GPP TSG CN. *General Packet Radio Service (GPRS); Interworking between the Public Land Mobile Network (PLMN) supporting GPRS and Packet Data Networks (3GPP TS 29.061 version 5.1.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, January 2002d. Cited on p. 23, 44.

3GPP TSG CN. *Mobile radio interface layer 3 specification; Core Network Protocols (3GPP TS 24.008 version 5.2.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, June 2002e. Cited on p. 45, 51.

3GPP TSG GERAN. *Channel Coding (3GPP TS 45.003 version 5.6.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, June 2002a. Cited on p. 39.

3GPP TSG GERAN. *Digital cellular telecommunications system (Phase 2+); Physical layer on the radio path; General description (3GPP TS 45.001 version 5.4.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, June 2002b. Cited on p. 39.

3GPP TSG GERAN. *General Packet Radio Service (GPRS); Base Station System (BSS) – Serving GPRS Support Node (SGSN); BSS GPRS Protocol (BSSGP) (3GPP TS 48.018 version 5.4.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, July 2002c. Cited on p. 23, 44.

3GPP TSG GERAN. *General Packet Radio Service (GPRS); Mobile Station (MS) – Base Station System (BSS) interface; Radio Link Control / Medium Access Control (RLC/MAC) protocol (3GPP TS 44.060 version 5.2.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, July 2002d. Cited on p. 20, 21, 37, 58, 95.

3GPP TSG GERAN. *General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2 (3GPP TS 43.064 version 5.0.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, April 2002e. Cited on p. 17, 38, 58, 89.

3GPP TSG GERAN. *Radio subsystem link control (3GPP TS 45.008 version 5.4.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, June 2002f. Cited on p. 58, 100.

3GPP TSG GERAN. *Radio Transmission and Reception (3GPP TS 45.005 version 5.4.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, June 2002g. Cited on p. 98.

3GPP TSG SSA. *General Packet Radio Service (GPRS); Service description, Stage 1 (3G TS 22.060 version 5.0.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, October 2001a. Cited on p. 48.

3GPP TSG SSA. *Services and Service Capabilities (3G TS 22.105 version 5.0.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, October 2001b. Cited on p. 52, 183.

3GPP TSG SSA. *General Packet Radio Service (GPRS); Service description; Stage 2 (3GPP TS 23.060 version 5.0.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, January 2002a. Cited on p. 39, 42, 45, 48.

3GPP TSG SSA. *QoS Concept and Architecture (3GPP TS 23.107 version 5.3.0 Release 5)*. Technical Specification, 3rd Generation Partnership Project, January 2002b. Cited on p. 51.

D. Anick, D. Mitra and M. Sondhi. *Stochastic Theory of a Data-Handling System with Multiple Sources. The Bell System Technical Journal*, Vol. 61, No. 8, pp. 1871–1894, October 1982. Cited on p. 114, 115, 116, 118, 119, 120.

M. Arango, A. Dugan, I. Elliott, C. Huitema and S. Pickett. *Media Gateway Control Protocol, Request for Comments 2705*. Technical Report, Internet Engineering Task Force (IETF), October 1999. Cited on p. 79.

M. Arlitt and C. Williamson. *A Synthetic Workload Model for Internet Mosaic Traffic*. In *Proceedings of the 1995 Summer Computer Simulation Conference*, pp. 24–26, Ottawa, Canada, 1995. Cited on p. 63.

S. Asmussen, O. Nerman and M. Olsson. *Fitting Phase Type Distributions via the EM Algorithm. Scand. J. Statist*, Vol. 23, pp. 419–441, 1996. Cited on p. 132, 133.

M.S. Baig, V. Sinha and P. Stuckmann. *Differentiated Services for Mobile Core Networks*. In *Proceedings of the 10th Symposium on Wireless Personal Multimedia Communications, ISBN 87-988568-0-4*, pp. 675–680, Aalborg, Denmark, September 2001. Cited on p. 47.

U. Black. *QoS in Wide Area Networks*. Prentice Hall, Upper Saddle River, N.J., 2000. Cited on p. 10.

C. Bormann, C. Burmeiste, M. Degermark, H. Fukushima, H. Hannu, L-E. Jonsson, R. Hakenberg, T. Koren, K. Le, Z. Liu, A. Martensson, A. Miyazak, K. Svanbro, T. Wiebke, T. Yoshimura and H. Zheng. *Robust Header Compression: Framework and four Profiles: RTP, UDP, ESP, and uncompressed, Request for Comments 3095*. Technical Report, Internet Engineering Task Force (IETF), July 2001. Cited on p. 82.

P. T. Brady. *A Model for Generating On-Off Speech Patterns in Two-Way Conversation. The Bell System Technical Journal*, No. 9, Vol. 48, pp. 2445–2472, September 1969. Cited on p. 82, 84.

G. Brasche. *Prototypical Evaluation and Implementation of new Packet Data Services for GSM (in German)*. Dissertation, RWTH Aachen, Lehrstuhl für Kommunikationsnetze, 1999. Cited on p. 15.

G. Brasche and B. Walke. *Concepts, Services and Protocols of the New GSM Phase 2+ General Packet Radio Service. IEEE Communications Magazine*, Vol. 35, 1997. Cited on p. 8, 15, 86.

K. C. Budka. *Cellular Digital Packet Data: Channel Availability. IEEE Transactions on Vehicular Technology*, Vol. 46, February 1997. Cited on p. 8, 9.

J. Cai and D. J. Goodman. *General Packet Radio Service in GSM. IEEE Communications Magazine*, Vol. 10, 1997. Cited on p. 8.

N. Cardwell, S. Savage and T. Anderson. *Modeling TCP Latency*. In *Proceedings of Infocom*, pp. 1742–1751, 2000. Cited on p. 107.

A. Chockalingham, M. Zorzi and V. Tralli. *Wireless TCP Performance with Link Layer FEC/ARQ*. In *Proceedings of the IEEE ICC'99)*, pp. 1212–1216, June 1999. Cited on p. 106.

H. K. Choi and J. O. Limb. *A Behavioral Model of Web Traffic*. In *Proceedings of the 7th International Conference on Network Protocols (ICNP 99)*, Ontario, Canada, October 1999. Cited on p. 64, 132.

M. Crispin. *Internet Message Access Protocol – Version 4, Request for Comments 2060*. Technical Report, Internet Engineering Task Force (IETF), December 1996. Cited on p. 65.

P. Decker. *A Packet Radio Protocol Proposed for the GSM Mobile Radio Network*. In *Workshop on Mobile Multimedia Communication (MoMuC-1)*, Tokyo, December 1993. IEEE. Cited on p. 15.

P. Decker. *Design and Performance Evaluation of Hybrid Error Control Protocols for Packet-based Speech and Data Services for GSM Networks (in German)*. Dissertation, RWTH Aachen, Lehrstuhl für Kommunikationsnetze, 1995. Cited on p. 15.

S. Deng. *Traffic Characteristics of Packet Voice. IEEE International Conference on Communications*, pp. 1369–1374, June 1995. Cited on p. 83, 84.

A. Dutta-Roy. *The Cost of Quality in Internet-style Networks. IEEE Spectrum*, No. 0018-9235/00, pp. 57–62, September 2000. Cited on p. 10, 12.

E. O. Elliot. *Estimates of Error Rates for Codes on Burst-noise Channels. Bell System Technical Journal, Vol. 42*, pp. 1977–1997, September 1963. Cited on p. 98.

I. Elsen, F. Hartung, U. Horn, M. Kampmann and L. Peters. *Streaming Technology in 3G*

Mobile Communication Systems. IEEE Computer, Vol. 34, No. 9, pp. 46–52, 2001. Cited on p. 61, 73.

ETSI. *Overall Requirements on the Radio Interface of the Universal Mobile Telecommunication System (UMTS) (ETSI TR 04-01/SMG-050401)*. Technical Report, European Telecommunications Standards Institute, September 1996. Cited on p. 5.

ETSI 3GPP. *Selection Procedures for the Choice of Radio Transmission Technologies of the Universal Mobile Telecommunication System UMTS (3G TR 101 112)*. Technical Report ETSI TR 101 112, European Telecommunications Standards Institute, April 1998. Cited on p. 63, 77.

A. Feldmann and W. Whitt. *Fitting Mixtures of Exponentials to Long-tailed Distributions to Analyze Network Performance Models. Performance Evaluation*, Vol. 31, pp. 245–279, 1998. Cited on p. 132.

M. Fiedler and H. Voos. *Fluid-flow Modelling of ATM-Multiplexers (in German)*. Herbert Utz Verlag, München, 1997. ISBN 3-89675-251-0. Cited on p. 113, 114, 115, 116, 118, 119, 120.

R. Fielding, J. Gettys, J. Mogul, H. Frystyk and T. Berners-Lee. *Hypertext Transfer Protocol – HTTP/1.1, Request for Comments 2068*. Technical Report, Internet Engineering Task Force (IETF), January 1997. Cited on p. 62.

A. Furuskär, M. Höök, S. Jäverbring, H. Olofsson and J. Sköld. *Capacity Evaluation of the EDGE Concept for Enhanced Data Rates in GSM and TDMA/136*. In *Proceedings of the Vehicular Technology Conference (VTC)*, pp. 1648–1652, 1999a. Cited on p. 98, 100.

A. Furuskär, S. Mazur, F. Müller and H. Olofsson. *EDGE: Enhanced Data Rates for GSM and TDMA/136 Evolution. IEEE Personal Communications*, pp. 56–65, June 1999b. Cited on p. 53, 54.

A. Furuskär and H. Olofsson. *Aspects of Introducing EDGE in Existing GSM Networks*. No. 1, pp. 28–37, Stockholm, Sweden, 1999. Ericsson Radio Systems AB. Cited on p. 58, 179.

E. N. Gilbert. *Capacity of a Burst-noise Channel. Bell System Technical Journal, Vol. 39*, pp. 1253–1266, September 1960. Cited on p. 98.

D. J. Goodman and S. X. Wei. *Efficiency of Packet Reservation Multiple Access. IEEE Transactions on Vehicular Technology, Vol. 40, No. 1*, pp. 170–176, February 1991. Cited on p. 9.

H. Gudding. *Capacity Analysis of GPRS*. Master thesis, NTNU Norwegian University of Science and Technology, Trondheim, Norway, March 2000. Cited on p. 51.

T. Halonen, J. Romero and J. Melero. *GSM, GPRS and EDGE Performance—Evolution Towards 3G/UMTS*. John Wiley & Sons, Chichester, 2002. Cited on p. 6.

M. Handley, H. Schulzrinne, E. Schooler and J. Rosenberg. *SIP: Session Initiation Protocol, Request for Comments 2543*. Technical Report, Internet Engineering Task Force (IETF), March 1999. Cited on p. 78.

Q. M. He. *Queues with Marked Customers. Adv. Appl. Prob., 28*, Vol. 28, pp. 567–587, 1996. Cited on p. 113, 132.

F. Hillebrandt. *The Creation of Global Mobile Communication*. John Wiley & Sons, Chichester, 2002. Cited on p. 15.

C. Hoymann and P. Stuckmann. *On the Feasibility of Video Streaming Applications over GPRS/EGPRS*. In *Proceedings of the IEEE Global Communications Conference (Globecom '02)*, Taipei, Taiwan, November 2002a. Cited on p. 155.

C. Hoymann and P. Stuckmann. *Traffic Engineering for the Evolution of GPRS/EDGE Networks*. In *Proceedings of the IEEE Vehicular Technology Conference (VTC '02 fall)*, September 2002b. Cited on p. 173.

T. Irnich and P. Stuckmann. *Analytical Performance Evaluation of Internet Access over GPRS and its Comparison with Simulation Results*. In *Proceedings of the 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Lisbon, Portugal, September 2002. Cited on p. 135.

T. Irnich and P. Stuckmann. *Fluid-flow Modeling of Internet Traffic in GSM/GPRS Networks. Computer Communications, Elsevier Science*, May 2003. Cited on p. 121.

ISO/IEC JTC1. *Open Systems Interconnection – Basic Reference Model: The Basic Model*. International Standard ISO/IEC 7498-1, International Organization for Standardization, Geneva, Switzerland, 1994. Cited on p. 22.

ISO/IEC JTC1/SC29. *Generic coding of moving pictures and associated audio information –*

Part 7: Advanced Audio Coding (MPEG 2 – AAC). International Standard ISO/IEC 13818-7, International Organization for Standardization, Geneva, Switzerland, 1997. Cited on p. 78.

ITU-T. *Data Compression Procedures for Data Circuit Terminating Equipment (DCE) Using Error Correction Procedures.* ITU-T Recommendation V.42bis, International Telecommunication Union – Telecommunication Standardization Sector, Geneva, Switzerland, 1990. Cited on p. 25.

ITU-T. *Recommendation H.261, Series H: Audiovisual and Multimedia Systems - Video codec for audiovisual services at p x 64 kbit/s.* Technical Report H.261, March 1993. Cited on p. 74.

ITU-T. *Packet-switched Signalling System between Public Networks and Providing Data Transmission Services.* ITU-T Recommendation X.75, International Telecommunication Union – Telecommunication Standardization Sector, Geneva, Switzerland, October 1996a. Cited on p. 44.

ITU-T. *Speech coders: Dual Rate Speech Coder for Multimedia Communications transmitting at 5.3 and 6.3 kbit/s.* Technical Report G.723, March 1996b. Cited on p. 79.

ITU-T. *Draft Text of Recommendation H.263 Version 2 (H.263+) for Decision.* Technical Report H.263, September 1997. Cited on p. 74.

ITU-T. *Packet-based Multimedia Communications System.* Technical Report H.323, November 2000. Cited on p. 78.

ITU-T SG 2. *Terms and Definitions Related to Quality of Service and Network Performance Including Dependability.* ITU-T Recommendation E.800, International Telecommunication Union – Telecommunication Standardization Sector, Geneva, Switzerland, 1994. Cited on p. 10.

V. Jacobson. *Compressing TCP/IP Headers for Low-Speed Serial Links, Request for Comments 1144.* Technical Report, Internet Engineering Task Force (IETF), February 1990. Cited on p. 25.

M. R. Jeong, H. Morikawa and T. Aoyama. *Fair Scheduling Algorithm for Wireless Packet Networks.* In *Proceedings of the 1999 International Workshops on Parallel Processing*, 1999. Cited on p. 13.

M. Junius. *CNCL: a C++ Library for Event Driven Simulation, Statistical Evaluation and Random Number Generators and Distributions.* Technical Report, Communication Networks, Aachen University of Technology, 1993. Cited on p. 85.

R. Kalden, I. Meirick and M. Meyer. *Wireless Internet Access Based on GPRS. IEEE Personal Communications*, Vol. 4, 2000. Cited on p. 8, 138.

I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus.* Second Edition, Springer-Verlag New York, 1991. Cited on p. 102.

L. Kleinrock. *Queueing Systems*, Vol. I: *Theory*. John Wiley & Sons, New York/London/Sydney/Toronto, 1975. Cited on p. 7, 121, 130.

L. Kosten. *Stochastic Theory of a Multi Entry Buffer.* In *Delft Progress Report*, Vol. 1 der *F*, pp. 10–18, 1974. Cited on p. 113, 114.

T. Laksman and U. Madhow. *The Performance of TCP/IP for Networks with High Bandwidth-delay Products and Random Loss. IEEE/ACM Trans.Networking*, Vol. 7, No. 5, pp. 36–50, 1999. Cited on p. 106.

S. S. Lam. *Packet Broadcast Networks—A Performance Analysis of the R-ALOHA Protocol. IEEE Transactions on Computers, Vol. 29, No. 7*, pp. 596–603, July 1980. Cited on p. 9.

S. Lin and D.J. Costello. *Error Control Coding—Fundamentals and Applications.* Prentice-Hall, 1983. Cited on p. 109.

C. Lindemann and A. Thümmler. *Performance Analysis of the General Packet Radio Service.* In *Proceedings of the 21st International Conference on Distributed Computing Systems (ICDCS '01)*, pp. 673–680, Phoenix, USA, April 2001. Cited on p. 8.

R. Litjens and R. J. Boucherie. *Radio Resource Sharing in a GSM/GPRS Network.* In *Proceedings of the 12th ITC Specialist Seminar on Mobile Systems and Mobility*, pp. 261–274, Lillehammer, Norway, 2000. Cited on p. 8, 9.

M. Mathis, J. Semke, J. Madhavi and T. Ott. *The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. Computer Communications Review*, Vol. 27, No. 3, pp. 67–82, 1997. Cited on p. 108.

D. Mitra. *Stochastic Theory of a Fluid Model of Producers and Consumers Coupled by a Buffer. Advances of Applied Probability*, Vol. 20, pp. 646–676, 1988. Cited on p. 113.

J. Myers and M. Rose. *Post Office Protocol — Version 3, Request for Comments 1939.* Technical Report, Internet Engineering Task Force (IETF), May 1996. Cited on p. 65.

M. F. Neuts. *A Versatile Markovian Point Process*. *Journal of Applied Probability*, Vol. 16, pp. 764–779, 1979. Cited on p. 132.

S. Ni and S. G. Häggmann. *GPRS Performance Estimation in a GSM Voice and GPRS Shared Resource System*. In *Proceedings of the IEEE Wireless Communication and Networking Conference (WCNC '99)*, pp. 1417–1421, 1999. Cited on p. 8, 9.

S. Ni, Y. Liang and S. G. Häggmann. *Outage Probability in GSM-GPRS Cellular Systems with and without Frequency Hopping*. *Wireless Personal Communications*, Vol. 14, 1999. Cited on p. 8, 9.

W. Olzem. *Integration of Protocols Specified in SDL into an Event-driven Simulation Environment*. Technical Report, RWTH Aachen, Lehrstuhl für Kommunikationsnetze, Kopernikusstr. 16, D-52074 Aachen, Germany, January 1995. Cited on p. 85.

V. Paxson. *Empirically-Derived Analytic Models of Wide-Area TCP Connections*. *IEEE/ACM Transactions on Networking*, Vol. 2, No. 4, pp. 316–336, August 1994. Cited on p. 66, 67, 68.

V. Paxson and S. Floyd. *Wide Area Traffic: The Failure of Poisson Modeling*. *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp. 226–244, June 1995. Cited on p. 132.

J. Peisa and M. Meyer. *Analytical Model for TCP File Transfers over UMTS*. In *Proceedings of the IEEE International Conference on Third Generation Wireless and Beyond (3Gwireless '01)*, San Francisco, USA, June 2001. Cited on p. 106.

C. Perkins. *IP Mobility Support, Request for Comments 2002*. Technical Report, Internet Engineering Task Force (IETF), October 1996. Cited on p. 23.

J. Postel. *Transmission Control Protocol, Request for Comments 793*. Technical Report, Internet Engineering Task Force (IETF), September 1981. Cited on p. 62.

J. Postel. *Simple Mail Transfer Protocol, Request for Comments 821*. Technical Report, Internet Engineering Task Force (IETF), August 1982. Cited on p. 65.

J. Postel and J. Reynolds. *File Transfer Protocol, Request for Comments 959*. Technical Report, Internet Engineering Task Force (IETF), October 1985. Cited on p. 67.

J. W. Roberts. *Traffic Theory and the Internet*. *IEEE Communications Magazine*, pp. 94–99, January 2001. Cited on p. 7, 9, 134.

O. Rose and M. Frater. *A Comparison of Models for VBR Video Traffic Sources in B-ISDN*. *IFIP Transactions C-24: Broadband Communications II*, pp. 275–287, 1994. Cited on p. 75.

J. Sachs and M. Meyer. *Mobile Internet—Performance Issues Beyond the Radio Interface*. In *Proceedings of the 10th Aachen Symposium on Signal Theory*, pp. 163–168, Aachen, Germany, September 2001. VDE Verlag. Cited on p. 110.

A. Schieder. *Real-time Services in Packet-switched Mobile Radio Networks (in German)*. Aachener Beiträge zur Mobil- und Telekommunikation. to appear in Verlag der Augustinus Buchhandlung, Aachen, 2003. Dissertation. Cited on p. 100.

F. Schreiber. *Effective Control of Simulation Runs by a New Evaluation Algorithm for Correlated Random Sequences*. In *Proceedings of the 12th International Teletraffic Congress (ITC)*, pp. 4.3B.1.1–9, 1988. Cited on p. 9, 104.

F. Schreiber and C. Görg. *Stochastic Simulation: A Simplified LRE-Algorithm for Discrete Random Sequences*. *AEÜ, Int. J. Electron. and Commun.*, Vol. 50, No. 4, pp. 233–239, 1996. Cited on p. 9, 104.

H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson. *RTP: A Transport Protocol for Real-Time Applications, Request for Comments 1889*. Technical Report, Internet Engineering Task Force (IETF), March 2001. Cited on p. 73.

M. Shreedhar and G. Varghese. *Efficient Fair Queueing Using Deficit Round Robin*. In *Proceedings of SIGCOMM 95*, pp. 636–646, August 1995. Cited on p. 13, 166, 169.

N. Sollenberger, N. Seshadri and R. Cox. *The Evolution of IS-136 TDMA for Third-Generation Wireless Services*. *IEEE Personal Communications*, Vol. 6, 1999. Cited on p. 54.

M. Steppler. *Performance Analysis of Communication Systems Formally Specified in SDL*. In *Proceedings of the First International Workshop on Simulation and Performance '98 (WOSP '98)*, pp. 49–62, 1998. Cited on p. 85.

D. Stern and A. Elwalid. *Analysis of Separable Markov-Modulated Rate Models for Information-Handling Systems*. *Advances in Applied Probability*, Vol. 23, pp. 105–139, 1991. Cited on p. 117.

R. Stevens. *TCP/IP Illustrated*, Vol. 1. Addison-Wesley, Massachusetts, October 1996. Cited on p. 124, 134.

P. Stuckmann. *Quality of Service Management in GPRS-based Radio Access Networks. Telecommunication Systems 19:3, Kluwer Academic Publishers*, pp. 515–546, March 2002a. Cited on p. 47, 121, 123.

P. Stuckmann. *The GSM Evolution—Mobile Packet Data Services.* John Wiley & Sons, Chichester, 2002b. Cited on p. 6, 15, 38, 146, 151.

P. Stuckmann, N. Ehlers and B. Wouters. *GPRS Traffic Performance Measurements.* In *Proceedings of the IEEE Vehicular Technology Conference (VTC '02 fall)*, Vancouver, Canada, September 2002. Cited on p. 106, 111, 137.

P. Stuckmann, H. Finck and T. Bahls. *A WAP Traffic Model and its Appliance for the Performance Analysis of WAP over GPRS.* In *Proceedings of the IEEE International Conference on Third Generation Wireless and Beyond (3Gwireless '01)*, pp. 338–343, San Francisco, USA, June 2001. Cited on p. 72, 73.

P. Stuckmann and J. Franke. *Performance Characteristics of the Enhanced General Packet Radio Service for the Mobile Internet Access.* In *Proceedings of the Second International Conference on 3G Mobile Communication Technologies*, pp. 287–291, London, UK, March 2001. Cited on p. 153.

P. Stuckmann and C. Hoymann. *Dimensioning GPRS Networks for Coexisting Applications based on WAP and Conventional Internet Protocols.* In *Proceedings of the European Wireless 2002 (EW '02)*, Florence, Italy, February 2002a. Cited on p. 155.

P. Stuckmann and C. Hoymann. *Performance Evaluation of WAP-based Applications over GPRS.* In *Proceedings of the IEEE International Conference on Communications (ICC '02)*, New York, USA, April 2002b. Cited on p. 155.

P. Stuckmann and C. Hoymann. *Traffic Performance Analysis of Multimedia Applications in Evolved GSM Networks.* In *Multiaccess, Mobility and Teletraffic for Wireless Communications, Volume 6*, pp. 279–293. Kluwer Academic Publishers, June 2002c. Cited on p. 155.

P. Stuckmann and O. Möller. *Advanced Scheduling and Admission Control Techniques for Cellular Packet Radio Networks.* In *Proceedings of the 5th European Personal Mobile Communications Conference (EPMCC 2003)*, Glasgow, Scotland, April 2003. Cited on p. 121, 169, 171.

P. Stuckmann and F. Müller. *GPRS Radio Network Capacity Considering Coexisting Circuit-Switched Traffic Sources.* In *Proceedings of the European Conference on Wireless Technology (ECWT 2000)*, pp. 66–69, Paris, France, October 2000. Miller Freeman. Cited on p. 165.

P. Stuckmann and F. Müller. *Quality of Service Management in GPRS Networks.* In *Proc. of the IEEE International Conference on Networking (ICN '01), LNCS 2093*, Vol. 1, pp. 276–285, July 2001. Cited on p. 47, 51, 90.

P. Stuckmann and O. Paul. *Dimensioning GSM/GPRS Networks for Circuit- and Packet-Switched Services.* In *Proceedings of the 10th Symposium on Wireless Personal Multimedia Communications, ISBN 87-988568-0-4*, pp. 597–602, Aalborg, Denmark, September 2001. Cited on p. 163.

U. Vornefeld. *Analytical Concepts for Packet Switching Mobile Radio Network Dimensioning.* In *Proceedings of the 10th Aachen Symposium on Signal Theory*, pp. 175–182, Aachen, Germany, September 2001. VDE Verlag. Cited on p. 9.

U. Vornefeld. *Analytical Concepts for GPRS Network Dimensioning.* In *Proceedings of the International Communications Conference '02*, New York, US, April 2002a. Cited on p. 113, 124, 132, 133, 134, 180.

U. Vornefeld. *Analytical Performance Evaluation of Mobile Internet Access via GPRS Networks.* In *Proceedings of the European Wireless 2002*, Florence, Italy, 2002b. Cited on p. 113, 124, 132, 133.

W3C. *Synchronized Multimedia Integration Language (SMIL) Boston Specification.* Working Draft SMIL 2.0, September 2001. Cited on p. 71.

B. Walke. *Mobile Radio Networks—Networking, Protocols and Traffic Performance.* 2nd Edition, John Wiley & Sons, Chichester, 2001. Cited on p. 8, 151.

B. Walke, W. Mende, P. Decker and R. Crumbach. *The Performance of CELLPAC: A Packet Radio Protocol Proposed for the GSM Mobile Radio Network.* In *Mobile Radio Conference*, pp. 57–63, Nice, France, November 1991a. Cited on p. 15.

B. Walke, W. Mende and G. Hatziliadis. *CELLPAC, a Packet Radio Protocol Applied to the Cellular GSM Mobile Radio Network*. In *Proceedings of the 41st IEEE Vehicular Technology Conference*, pp. 408–413, St. Louis, Missouri, USA, May 1991b. Cited on p. 15.

J. Wigard and P. Mogensen. *A Simple Mapping from C/I to FER and BER for a GSM Type of Air-Interface*. In *Proceedings of the International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 78–82, 1996. Cited on p. 98.

J. Wigard, T. T. Nielsen, P. H. Michaelsen and P. Mogensen. *BER and FER Prediction of Control and Traffic Channels for a GSM Type of Air-Interface*. In *Proceedings of the Vehicular Technology Conference (VTC)*, pp. 1588–1592, 1998. Cited on p. 98.

W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson. *Self-similarity through High-variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level*. *IEEE/ACM Transactions on Networking*, Vol. 5, pp. 71–86, February 1997. Cited on p. 75.

Wireless Application Protocol Forum. *Wireless Application Protocol - Wireless Application Environment Overview*. Technical Report, 1999a. Cited on p. 61, 68.

Wireless Application Protocol Forum. *Wireless Application Protocol Architecture Specification*. Technical Report, 1999b. Cited on p. 68.

Wireless Application Protocol Forum. *Wireless Application Protocol - Multimedia Messaging Service Architecture Overview Specification, Version 25*. Technical Report, 2001a. Cited on p. 71.

Wireless Application Protocol Forum. *Wireless Application Protocol - WAP 2.0 Technical White Paper*. Technical Report, 2001b. Cited on p. 70.

Wireless Application Protocol Forum. *Wireless Application Protocol 2.0 - Wireless Application Protocol Architecture Specification*. Technical Report, 2001c. Cited on p. 61, 70.

Wireless Application Protocol Forum. *Wireless Profiled TCP*. Technical Report WAP-225-TCP-200103331-a, WAP Forum, March 2001d. Cited on p. 70.

Y. Yatsuzuka. *Highly Sensitive Speech Detector and High-Speed Voiceband Discrimination in SDI-ADPCM system*. *IEEE Transactions on Communications*, No. 4, Vol. 30, pp. 739–750, April 1982. Cited on p. 83.

T. Yoshimura, M.R. Jeong, H. Morikawa and T. Aoyama. *Wireless Packet Scheduling for Adaptive Service over IEEE802.11*. In *The First Int. Symp. on Wireless Personal Multimedia Communications, Yokosuka*, pp. 314–319, 1998. Cited on p. 13.

M. Zaddach and K. Heidtmann. *Measurement and Traffic Characterization of H.26x-coded Video Streams*. In *Proceedings of the 11th GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems*. VDE-Verlag, Berlin, 2001. Cited on p. 75.

Symbols

A

B

C

G

H

I

J

L

M

BIOGRAPHY

24th of April 1973	Born in Eslohe
1979–1983	Primary School in Schalksmühle
1983–1992	Bergstadt-Gymnasium Lüdenscheid
07/1992–10/1993	Military Service
10/1993–02/1999	Studies of Electrical Engineering at Aachen University (RWTH)
05/1999–05/2003	Research Assistant at the Chair of Communication Networks of Aachen University (RWTH)

ACKNOWLEDGEMENTS

This thesis has been authored during my activity as a research assistant at the Chair of Communication Networks of Aachen University (RWTH).

Firstly, I want to thank Prof. Dr.-Ing. Bernhard Walke for offering the possibility of my PhD studies at his chair, the active support of my research as well as the constructive review of the thesis text. Prof. Dr. Petri Mähönen I want to thank for the co-evaluation activities and the comments on the manuscript.

Then I want to thank my collegues at the chair for the fun we had together and for the cooperative atmosphere. I also thank all master students and student helpers that have contributed to my research activities.

Last but not least thanks to my family for their love and their great support of my studies.

Aachen, June 2003 *Peter Stuckmann*