Telecommunication Systems 19:3,4, 515–546, 2002

Quality of Service Management in GPRS-Based Radio Access Networks

PETER STUCKMANN pst@comnets.rwth-aachen.de
Communication Networks, Aachen University of Technology, Kopernikusstrasse 16, 52074 Aachen, Germany

Abstract. In this article, the performance and capacity gain achievable with quality of service (QoS) management in packet switched radio networks based on the General Packet Radio Service (GPRS) are examined. Both the functions defined in the GPRS specification for QoS support and implementation-specific strategies for subscriber- and application-based Connection Admission Control (CAC) and scheduling are introduced. The feasibility of QoS provisioning in mobile core networks with use of DiffServ compared to present IP technology realizing a pure Best-effort service is examined in addition. To achieve this, simulation results of GPRS performance and system measures for different load situations are produced with the simulation tool GPRSim that models the realistic traffic behavior of a GPRS network.

Keywords: GPRS, QoS, DiffServ, GPRSim, stochastic simulation

1. Introduction

In the framework of the evolution of the Global System for Mobile Communication (GSM) towards third-generation (3G) mobile communication systems, known as the International Mobile Telecommunications 2000 (IMT-2000) family of systems, new standards are presently integrated into the existing mobile radio networks. The driving force for this development is the predicted user demand for mobile data services that will offer mobile multimedia applications and mobile Internet access.

After High Speed Circuit Switched Data (HSCSD) has been introduced in some countries in 1999, the General Packet Radio Service (GPRS) will be available in 2001 in Europe and many other countries worldwide. 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. To realize mobile real-time applications as the next step the European Standardization Institute (ETSI) and the 3rd Generation Partnership Project (3GPP) are presently developing the Enhanced Data Rates for GSM Evolution (EDGE) standard, which offers a net bit rate of up to 384 kbit/s by means of modified modulation, coding and medium access schemes. The packet-oriented part is the Enhanced General Packet Radio Service (EGPRS). GSM and EDGE networks extended by GPRS capabilities are called GSM/EDGE Radio Access Networks (GERAN) in the latest ETSI/3GPP standardization [Furuskär et al., 18; Stuckmann and Franke, 31].

The next evolution step will be the integration of new air interfaces like Wideband-CDMA realizing UMTS Terrestrial Radio Access Networks (UTRANs) and the

HIPERLAN/2-based infrastructure in the local area realizing Broadband Radio Access Networks (BRAN) [Walke, 36].

For the interconnection of these new radio systems with the information infrastructure, e.g., the public Internet, all these radio access technologies will be based on the same core network architecture. Core networks standardized by ETSI/3GPP are composed of IP routers, which realize the tunneling of user data to gateway IP routers and the interworking functions with subnetworks like external networks or other Public Land Mobile Networks (PLMNs).

The Internet Protocol (IP) is not sufficient to serve traffic with specific latency, variance, packet loss, and throughput requirements. As a result, proposals have been made to improve the Best-effort services of IP. Differentiated Services (DiffServ) [Blake et al., 3] specified by the Internet Engineering Task Force (IETF) is one such proposal, which is at present seen as the future technology for mobile core networks and other Internet networks when the focus is set on the scalability of network resources.

While in the first phase after GPRS introduction only Best-effort data services without differentiating subscribers and applications will be supported, in the second phase quality of service (QoS) management functions will be integrated to be able to guarantee subscriber- and application-specific QoS requirements. These QoS functions will be based on the aggregation of flows belonging to the same service class and the prioritized admission control and scheduling of these aggregate flows in the radio network. No strict resource reservation is performed. This is the motivation to use DiffServ in GPRS/EDGE core networks [Costa and Dell'Uomo, 8; Koodli and Puuskari, 25], since DiffServ is also a scalable approach without need for per-flow state information and signalling at every hop, in contrast to the Integrated Services (IntServ) [Braden et al., 5] approach, which is based on the Resource Reservation Protocol (RSVP) [Braden et al., 4].

For radio network dimensioning and network equipment further development the effect of these QoS management functions on the overall system performance has to be determined. This article does not aim at optimized solutions for QoS management functions, but at an estimation of the performance gain achieved with their introduction compared to a pure Best-effort service.

First, the performance gain through QoS functions in the radio network is examined compared to a pure Best-effort service. Since radio resources are scarce and representing the system bottleneck it is firstly assumed that the core network is well dimensioned.

In the next step an IP-based core network with and without DiffServ capabilities is regarded. The feasability of QoS support in mobile core networks interworking with the QoS functions in the radio network is examined for low traffic load situations and under congestion.

2. Quality of service – general aspects

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" [21].

It is subject to numerous parameters, the most important of which will be described in the following.

2.1. QoS characteristics

Technically, QoS refers to an aggregation of system performance measures. The five most important of these are [Black, 2; Dutta-Roy, 9; 21]:

Availability. The availability of a network, its components, or even a service, should ideally approximate 100%. Even a figure 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, or bandwidth, 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 to adopt 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 owing to the special conditions on the radio interface. Both cases usually result in a retransmission of the packet becoming necessary, 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, especially interactive real-time (RT) applications are affected by it, 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).
- *Queuing delay*: the time a packet must wait in a queue before it is scheduled for transmission.

Table 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

In a mobile radio environment there may be additional delays adding to the delays mentioned, caused by random access (RA) 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 re-assembly 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, and are then called "adaptive" applications. Packets arriving after their playback point has passed are generally not useful to the application, and are, in most cases, discarded.

As mentioned above, applications vary significantly in their QoS requirements (see table 1, [Dutta-Roy, 9]). Interactive RT applications, like VoIP, of course 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 independent, but need as good as possible throughput values. 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, due to packet retransmissions, when regarding NRT applications. Network availability should, of course, be as high as possible, but is not a parameter under the influence of mechanisms for QoS provisioning. Thus, such mechanisms have to primarily concentrate on optimizing delay and throughput measures on behalf of application requirements.

2.2. QoS provisioning

There are two basic mechanisms for providing adequate QoS based on delay and throughput measures:

- plentiful capacity,
- traffic management.

With *plentiful capacity*, the assumption is that there is enough "capacity" available in the network that no explicit mechanisms have to be provided to ensure QoS. This implies that there are enough:

- high-capacity links,
- fast processors,
- plentiful buffers.

This is perhaps a reasonable assumption for a controlled, localized environment, such as a corporate LAN. It is unlikely that such assumptions will hold true across a global network such as the Internet. While the cost of bandwidth, memory, and processing are coming down (MOORE'S law), there are still a high cost to be paid for the high-end equipment assumed by this model. Even today, the cost of high-capacity Wide Area Network (WAN) links are quite high, as are those of high-speed routers and switches.

A variation of the *plentiful capacity* model is that, even if enough bandwidth is not available, applications can adapt to variing bearer quality. Adaptive applications certainly exist, but they are restricted to a certain range of parameters to provide useful service. Therefore, beyond a certain point, it may not be possible to provide adequate QoS without any explicit controls.

The second model of provisioning QoS is *traffic management*. The fundamental idea here is that traffic can be differentiated and classified into different levels 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 be exerted on how much traffic of each class is allowed into the network, based on the available resources – this may be done statically (provisioning) or dynamically (signalling 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 subcategories 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 set up 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 queuing mechanisms, e.g., Weighted Fair Queueing (WFQ), to service packets.

Reservationless. In this model, 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.

Coming to GPRS, ETSI standardization has followed the second approach, defining a set of QoS parameters that are combined to QoS profiles 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 undiminished, and that any kind of traffic will be served at least to a certain degree.

3. General packet radio service (GPRS)

The main intention of integrating the GPRS into the GSM is to increase the number of connections per bearer by utilizing the given physical channels more efficiently than the existing *phase-2* services. Variable data rates that are broadly higher than 9.6 kbit/s are realized by the possibility of multi-slot assignment [Brasche and Walke, 6; Cai and Goodman, 7; Kalden et al., 23].

GPRS has been standardized by the ETSI as part of the GSM *phase 2* + development. It represents the first implementation of packet switching within GSM, which is essentially a circuit-switched 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 concurrently 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 on one or more GSM physical channels, GPRS reaches a flexible use of channel capacity for applications with variable bit rate.

GPRS is extremely efficient in its use of scarce spectrum resources and enables GSM operators to introduce a wide range of value-added services for market differentiation. It is ideal for "bursty" data applications such as e-mail, Internet access or Wireless Application Protocol (WAP)-based applications. It integrates IP infrastructure into the GSM network and enables access to a wide range of public and private data networks using industry standard data protocols such as TCP/IP and X.25.

To identify starting-points for a traffic and QoS management functionality within GPRS, the system's logical architecture as well as the protocol architecture have to be regarded.

3.1. Logical architecture

In order to integrate the functionality for a packet data service, the GSM architecture is extended by several logical entities to realize GPRS:

GGSN. The Gateway GPRS Support Node (GGSN) serves as the interface towards external Packet Data Networks (PDN) or other Public Land Mobile Networks (PLMN).

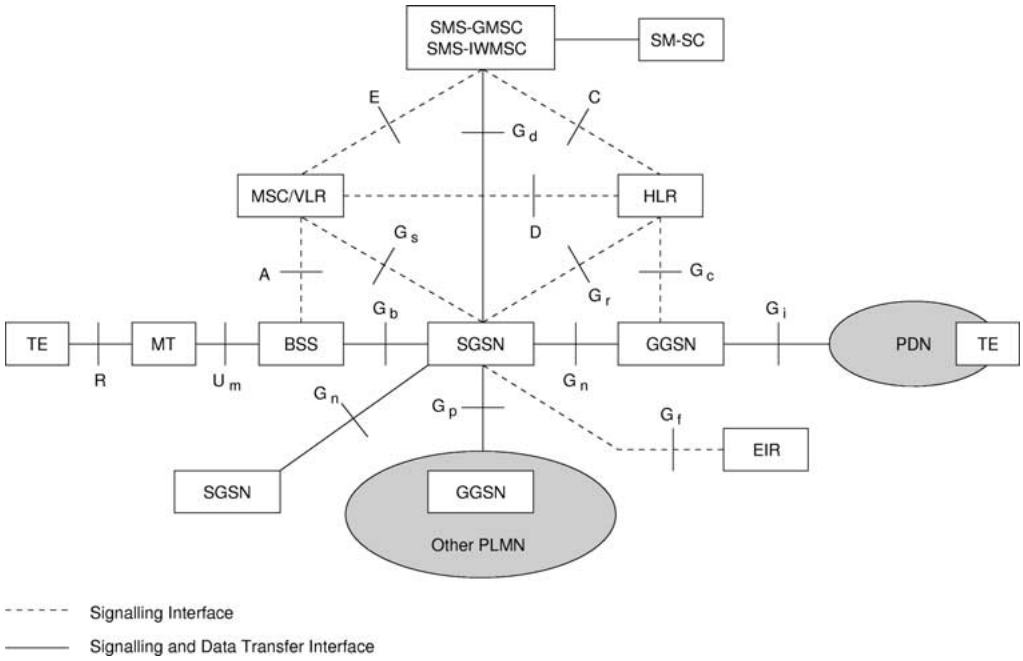

Figure 1. GPRS interfaces and reference points.

Here, switching functions are realized, e.g., the processing of the Packet Data Protocol (PDP) addresses and the routing to mobile subscribers via the SGSN.

SGSN. The Serving GPRS Support Node (SGSN) represents the GPRS switching centre in analogy to the Mobile Services Switching Centre (MSC) in GSM. PDP addresses are evaluated and mapped to the Interim Mobile Subscriber Identity (IMSI). The Serving GPRS Support Node (SGSN) is responsible for the routing inside the packet radio network and for mobility and resource management. Furthermore, it provides authentication and ciphering between GPRS subscribers.

GR. All GPRS subscriber-related information is stored in the GPRS Register (GR) that has to be regarded as part of the Home Location Register (HLR). Particularly, the IMSI is associated to one or several PDP addresses in the GR. In addition, the PDP addresses are associated to one or several GGSN. Subscriber QoS profiles are located in the GR as well.

With the extension of the existing GSM network by GPRS specific units, also the interfaces and reference points had to be redefined [11]. The interfaces defined are displayed in figure 1. The dotted lines represent signalling traffic between the related elements. Solid lines mean that user data can be transmitted additionally at these reference points.

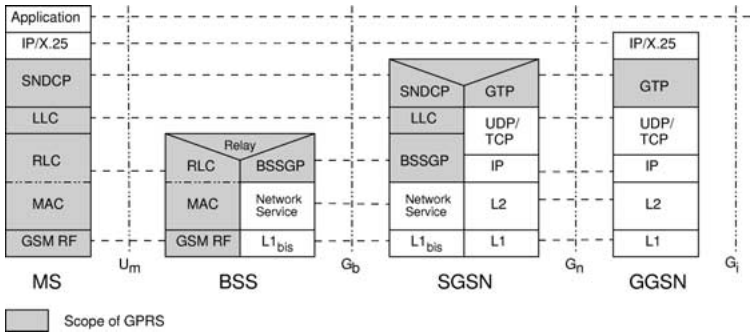

Figure 2. GPRS protocol stack.

3.2. Protocol architecture

The GPRS protocol architecture follows the International Standards Organisation/Open Systems Interconnection (ISO/OSI) reference model. The protocol stack is shown in figure 2. For GPRS, interworking with IP and X.25 based networks is foreseen.

Between SGSN and GGSN the GPRS Tunneling Protocol (GTP) realizes a transport service of IP packets by establishing an IP tunnel through the GPRS core network.

On the G_b interface, there is a connectionless link provided between SGSN and Base Station Subsystem (BSS) by the Base Station Subsystem GPRS Protocol (BSSGP) working on top of the network service. The network service is a link layer protocol and is based on Frame Relay (FR).

Communication between MS and SGSN is handled by the SNDCP and the LLC layer. The Subnetwork Dependent Convergence Protocol (SNDCP) is used to support multiple network layer protocols and multiplexes data generated by different sources. The Logical Link Control (LLC) layer is located between the BSS and the SGSN and provides a reliable ciphered link between MS and the network.

Finally, MS and BSS comprise the RLC/MAC layer that is responsible for reliable transmission of the LLC Protocol Data Units (PDUs) on the shared radio channels between MS and BSS.

4. End-to-end quality of service provisioning in GPRS networks

The performance of different applications experienced by a user is influenced by all network elements located on the path between the client and the server. Depending on the location of the server, also called host, different end-to-end QoS scenarios have to be regarded.

In figure 3, a typical GPRS network (PLMN 1) is shown, with its ingress/egress nodes and neighbouring external networks that may lie in the transmission path of data packets requested by a GPRS terminal TE. Only in case of the requested service being offered by Host 1, the GPRS network operator is given full control of all nodes and links located in the transmission path. Even then, it might be necessary for the operator to rent

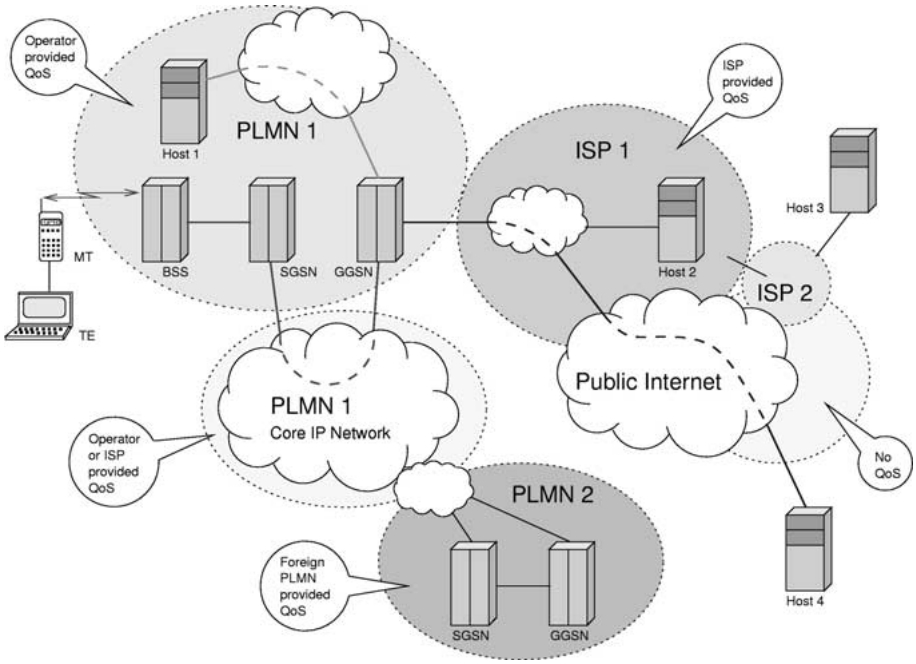

Figure 3. End-to-end QoS for different server locations.

transport resources for his core IP network from a separate network operator (PLMN 1 Core IP Network) and partially lose the QoS control. Any other host (Host 2–Host 3) being the GPRS subscriber's requested target host, withdraws a substantial part of the intermediate network from the GPRS operator's controlled domain.

To be in a position to offer QoS to the subscriber, the operator needs to have contracts with the owners of the external networks, so-called Service Level Agreements (SLAs), that assure an appropriate QoS mapping between the seperate domains. Should the public Internet be part of the transmission path (Host 4) no more than Best-effort treatment can be expected.

4.1. QoS in the radio network

To define a QoS contract between the mobile station (MS) and the network, Packet Data Protocol (PDP) contexts containing QoS profiles are negotiated between the MS and the Serving GPRS Support Node (SGSN) [16]. In ETSI Release 99, the Base Station Subsystem (BSS) is provided with a Packet Flow Context (PFC) containing an Aggregate BSS QoS Profile (ABQP) and is responsible for resource allocation on a Temporary Block Flow (TBF) base and scheduling of packet data traffic with respect to the according QoS profiles negotiated. Moreover, it regularly informs the SGSN about the current load conditions in the radio cell. The tasks of the Gateway GPRS Support Node (GGSN) comprise mapping of PDP addresses as well as classification of incoming traffic from external networks regarding the downlink Traffic Flow Template (TFT). The

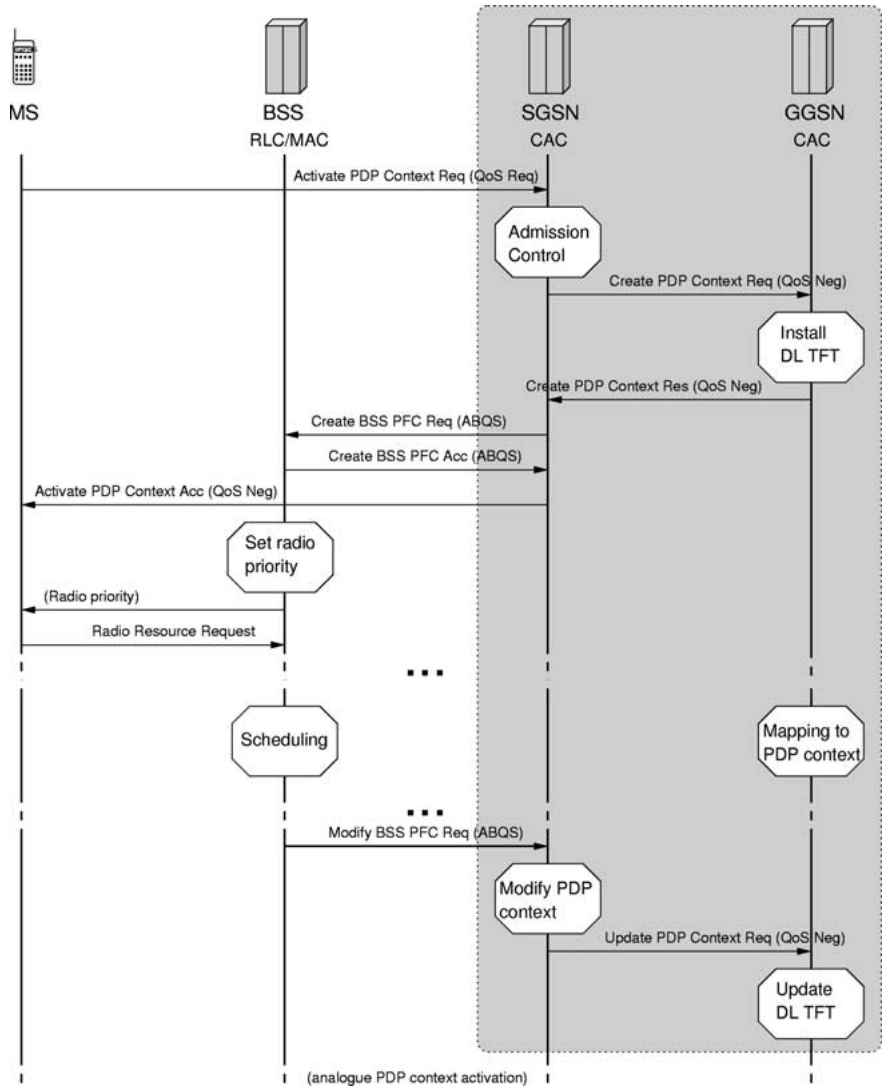

Figure 4. QoS negotiation and renegotiation procedures (example).

GPRS Register (GR) holds the QoS-related subscriber information and delivers it on demand to the SGSN. Figure 4 shows a GPRS session depicting the actions of PDP context (re)negotiation, PFC setup, and TFT installation.

From a time-scale point of view, the mechanisms for QoS management within the GPRS can be regarded as a three-stage model (see figure 5). 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 a set of Packet Data Channels (PDCH) usable for this TBF are assigned

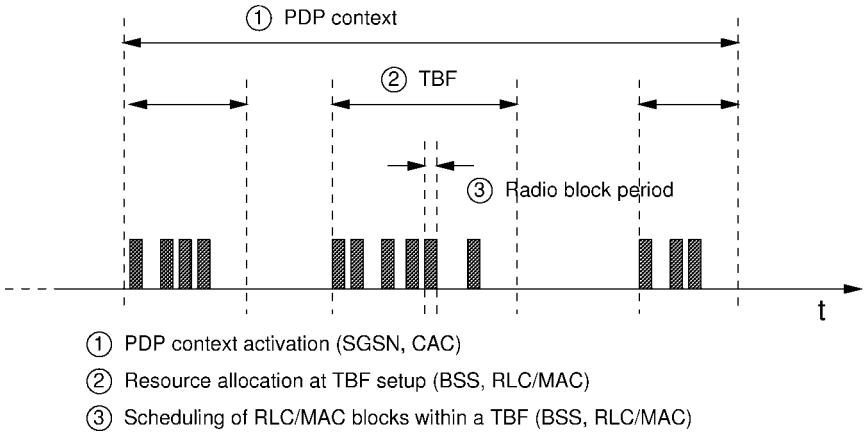

Figure 5. Three-stage model of QoS management.

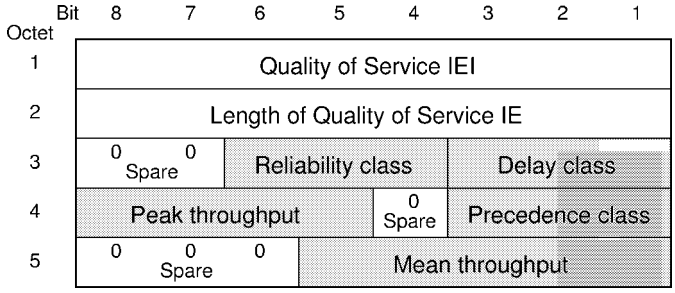

Figure 6. QoS profile information element.

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 performed considering the QoS profiles of the PDP contexts associated with the TBFs.

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 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.

4.1.1. QoS attributes according to GPRS release 97/98

All QoS-related information to be exchanged between MS and SGSN is stored in a QoS profile. The QoS profile Information Element (IE) is shown in figure 6. It consists of an Information Element Identifier (IEI), a length field, five fields that contain the values

Table 2
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
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			

of the GPRS service classes, and three fields filled with spare bits [17]. A QoS profile defines the QoS within the range of the following service classes [10, 16].

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 precendence classes are presented in table 2.

Delay classes. The packet delay is defined by the time needed for transmission from one GPRS Service Access Point to another. Delays outside the system, e.g., in transit networks, are not considered. The technical specification 3GPP 022.060 [10] determines four delay classes (see table 3). Although there is no need for all delay classes to be available, at least *best effort* has to be offered.

Reliability classes. Data services generally require a low residual bit error rate. 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 data loss,
- 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 the services of each layer. By combining different modes of operation of the GPRS specific protocols GTP, LLC, and RLC the reliability requirements of various applications, e.g., real-time (RT) or non-real-time (NRT) are supported. The reliability classes are summarized in table 4.

Table 4
Reliability classes.

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	UPR	ACK	NRT traffic, error sensitive, not loss sensitive
4	UACK	UACK	UPR	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

Table 5
Peak throughput classes.

Peak throughput class	Peak throughput	
	[byte/s]	[kbit/s]
1	up to 1000	8
2	up to 2000	16
3	up to 4000	32
4	up to 8000	64
5	up to 16000	128
6	up to 32000	256
7	up to 64000	512
8	up to 128000	1024
9	up to 256000	2048

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. The peak throughput is measured in byte/s at the reference points G_{i} and R. Peak throughput specifies the maximum rate, at which data is transmitted within a certain PDP context. There is no guarantee given that this data rate is actually achieved at any time during transmission. 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 5.

Table 6
Mean throughput classes.

Mean throughput class	Mean throughput	
	[byte/h]	≈[bit/s]
1	100	0.22
2	200	0.44
3	500	1.11
4	1000	2.2
5	2000	4.4
6	5000	11.1
7	10000	22
8	20000	44
9	50000	111
10	100000	220
11	200000	440
12	500000	1110
13	1000000	2200
14	2000000	4400
15	5000000	11100
16	10000000	22000
17	20000000	44000
18	50000000	111000
31	Best effort	

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 data rate 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 an MS whenever there are resources needed and at disposal. Table 6 summarizes the classes of mean throughput.

4.1.2. QoS in GPRS release 99

The QoS architecture defined in GPRS Release 97/98 shows some major drawbacks (see also [Gudding, 19; Stuckmann and Müller, 33]):

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. Only one QoS profile can specify the requirements of all applications for one PDP address.

In GPRS Release 97/98, the BSS cannot use QoS profile information to schedule resources on a continuous data flow, since there is no mechanism provided to download

the QoS profile from the SGSN. 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. 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. It 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 per MS, so that each application may receive the appropriate QoS requirement. Additionally, GPRS Release 99 defines several further QoS parameters with finer grained properties to meet requirements on different levels of service for applications [13]:

- maximum bitrate,
- delivery order,
- Service Data Unit (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').

Finally, four distinct *traffic classes* are introduced. The following parameters are specifying their QoS requirements (see table 7):

Table 7
End-user performance expectations for selected services belonging to different traffic classes.

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

- *conversational*,
- *streaming*,
- *interactive*,
- *background*.

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

Architectural components, which are conform with GPRS Release 99, will not be available before the year 2002. For interworking purposes with GPRS Release 97/98 network equipment, mapping rules between the Relaese 99 traffic classes and the Release 99 service classes are defined [13]. In the context of the examination presented in this article, only the traffic classes defined in Release 99 together with different subscriber classes are regarded for application-level QoS profile definitions, admission control rules, and application-based scheduling. It is assumed that each new session is classified into a Release 99 service class. The QoS classes of Release 97/98 will not be further regarded in this article.

4.2. QoS in the core network applying DiffServ

Differentiated Services (DiffServ) is an approach to provide quality of service (QoS) within IP networks in a scalable manner. It is a relatively simple and coarse method of providing differentiated classes of service for various application and subscriber requirements in IP networks [Blake et al., 3].

The DiffServ architecture (see figure 7) comprises a small, well-defined set of building blocks from which a variety of aggregate behaviors may be built. The QoS support may be end-to-end or intra-domain. Both quantitative performance requirements (e.g., peak throughput) and requirements based on relative performance (e.g., class differentiation) are supported.

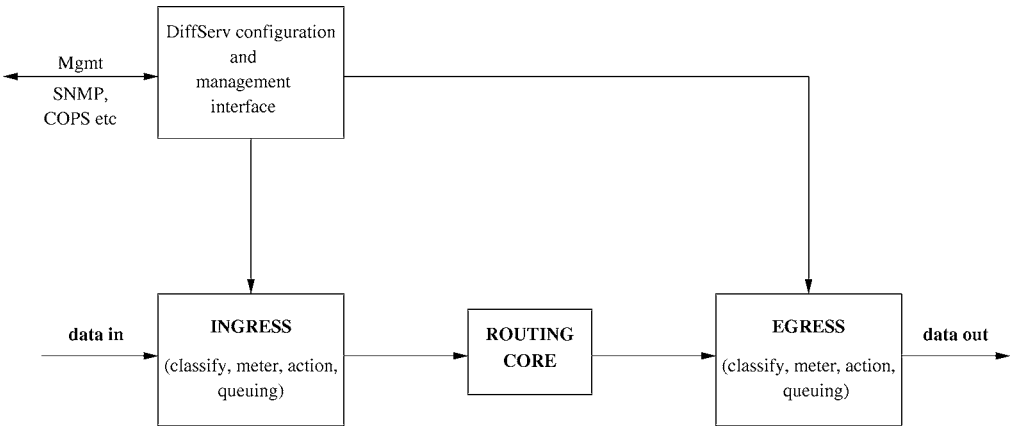

Figure 7. Differentiated services architecture.

In DiffServ, the network allows applications to negotiate one of the several different services through connection admission control (CAC). The packets generated by applications are treated differently by the network. With proper engineering, including boundary policing, DiffServ can provide expedited handling appropriate for a wide class of applications, including delay-critical applications. DiffServ-capable routers need to track a small number of per-hop behaviors and they serve packets based on a single byte.

4.2.1. DiffServ field

A bit pattern in each packet, in the IPv4 Type of Service (TOS) octet or the IPv6 traffic class octet (see figure 8) is used to mark a packet to receive a particular forwarding treatment, or per-hop behavior, at each network node. A common understanding of the use and interpretation of this bit pattern is required for inter-domain use, multi-vendor interoperability, and consistent reasoning about expected aggregate behaviors in a network. Thus, the DiffServ Working Group of the Internet Engineering Task Force (IETF) has standardized a common layout for a six-bit field of both octets, called the DiffServ field (DS field) [Nichols et al., 26].

The first six bits of the DS field are used as a codepoint (DSCP). These determine the per-hop behavior (PHB) the packet sees at each node and usually consists of packet queuing and scheduling. PHBs define how traffic belonging to a particular behavior aggregate is treated at an individual network node. A two-bit field currently unused (CU) is reserved for future use. Depending on the first three bits of DSCP 8 precedence levels (classes) are available in DiffServ as shown in table 8.

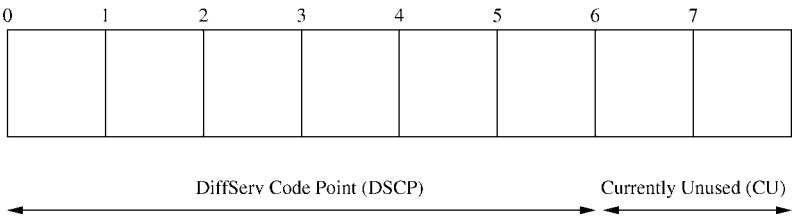

Figure 8. IPv4 type of service (TOS) octet or IPv6 traffic class octet.

Table 8
Precedence levels of DiffServ based on bits 0, 1, 2 of DSCP.

Bits 0, 1, 2 of DSCP	Precedence level	Usage
111	7	Link layer and routing protocol control information
110	6	Routing protocol control information
101	5	Expedited Forwarding class
100	4	Assured Forwarding class 4
011	3	Assured Forwarding class 3
010	2	Assured Forwarding class 2
001	1	Assured Forwarding class 1
000	0	Best Effort class

4.2.2. Per-hop behaviors at interior routers
An interior router is any router that is not at the boundary of a DiffServ network domain. Since interior routers make up the majority of routers through which most IP packets pass, the complexity of the functions performed by interior routers must remain low. The DiffServ architecture recognizes this fact and mandates that only simple PHBs are implemented at interior routers.

4.2.3. Classification and conditioning at boundary routers
The boundary router is located at the edge of a DiffServ network domain. This boundary router must perform sophisticated packet classification, metering, marking, policing, and shaping operations.

4.2.4. Bandwidth broker
To make appropriate internal and external admission control decisions and to configure boundary devices correctly, each DiffServ domain is outfitted with a bandwidth broker. The bandwidth broker performs admission control depending on network resources and configures boundary routers to take or drop a connection request. How the information of the traffic load situation at the network nodes is retrieved is not specified in the DiffServ specifications. The bandwidth broker might request information from the queues of all routers on one or several paths before admitting a request.

4.3. A proposal for mapping GPRS/UMTS classes to DiffServ classes

To integrate the QoS management functions in the radio network and the QoS functions in the IP core network, e.g., DiffServ, mapping rules have to be defined so that these functions can interwork efficiently. A proposal for mapping GPRS/UMTS classes onto DiffServ classes are presented in table 9. In this proposal the Conversational class is mapped to the expedited forwarding class since it is a real-time application and requires both low delay and low delay jitter. The Streaming class is mapped to assured forwarding class 4 since it requires stringent delay jitter requirements. The Interactive class is mapped to the assured forwarding class 3 since it requires low latency but not as low as in the Conversational class. The Background class can be mapped to any lower DiffServ class.

Table 9
Proposal for mapping 3GPP classes to DiffServ classes.

3GPP QoS class	DiffServ class	Reason
Conversational	Expedited Forwarding class	low latency and jitter required
Streaming	Assured Forwarding class 4	low jitter required
Interactive	Assured Forwarding class 3	relatively low latency required
Background	Assured Forwarding class 2 or Assured Forwarding class 1 or Best Effort	only reliability required

5. Simulation environment

Although analytical and algorithmic models for the performance analysis of packet-switched radio networks are under development [Vornefeld, 35], 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 described by formulas usable in practice. Since GPRS networks are presently introduced in the field, traffic engineering rules and related performance results are needed soon, so that capacity and performance estimations become possible for GPRS introduction and evolution scenarios.

Measuring the traffic performance in the existing GPRS network is not possible, since a scenario with a well-defined traffic load is hard to set-up, the evaluation of the performance by measurement is very difficult, and the analysis of different protocol options is not possible in an existing radio network.

Therefore computer simulation based on the prototypical implementation (called emulation) of the GPRS protocols 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.

The (E)GPRS Simulator GPRSim [Stuckmann, 34] is a pure software solution based on the programming language C++. Up to now models of Mobile Station (MS), Base Station (BS), Serving GPRS Support Node (SGSN), and Gateway GPRS Support Node (GGSN) have been implemented. The simulator offers interfaces to be upgraded by additional modules (see figure 9).

For the implementation of the simulation model in C++ the Communication Networks Class Library (CNCL) [Junius et al., 22] is used, a predecessor to the SDL Performance Evaluation Tool Class Library (SPEETCL) [Steppler, 28]. This enforces an object oriented structure of programs and is especially suited for event driven simulations.

Different from usual approaches to establish 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 GSM and (E)GPRS protocols. This enables a realistic study of the behavior of EGPRS and GPRS. The real protocol stacks of (E)GPRS are used during system simulation and are statistically analyzed under a well-defined and reproducable traffic load.

The complex layers of the protocol stacks like SNDCP, LLC, RLC/MAC based on (E)GPRS Release 99, the Internet traffic load generators and TCP/IP itself are specified formally with the Specification and Description Language (SDL) [20], translated to C++ code by means of the Code Generator SDL2CNCL [Steppler, 28] and finally integrated into the simulator.

5.1. Packet traffic generators

The Internet sessions studied consist of the applications World Wide Web (WWW) and electronic mail (e-mail) running on top of the TCP/IP protocol stack.

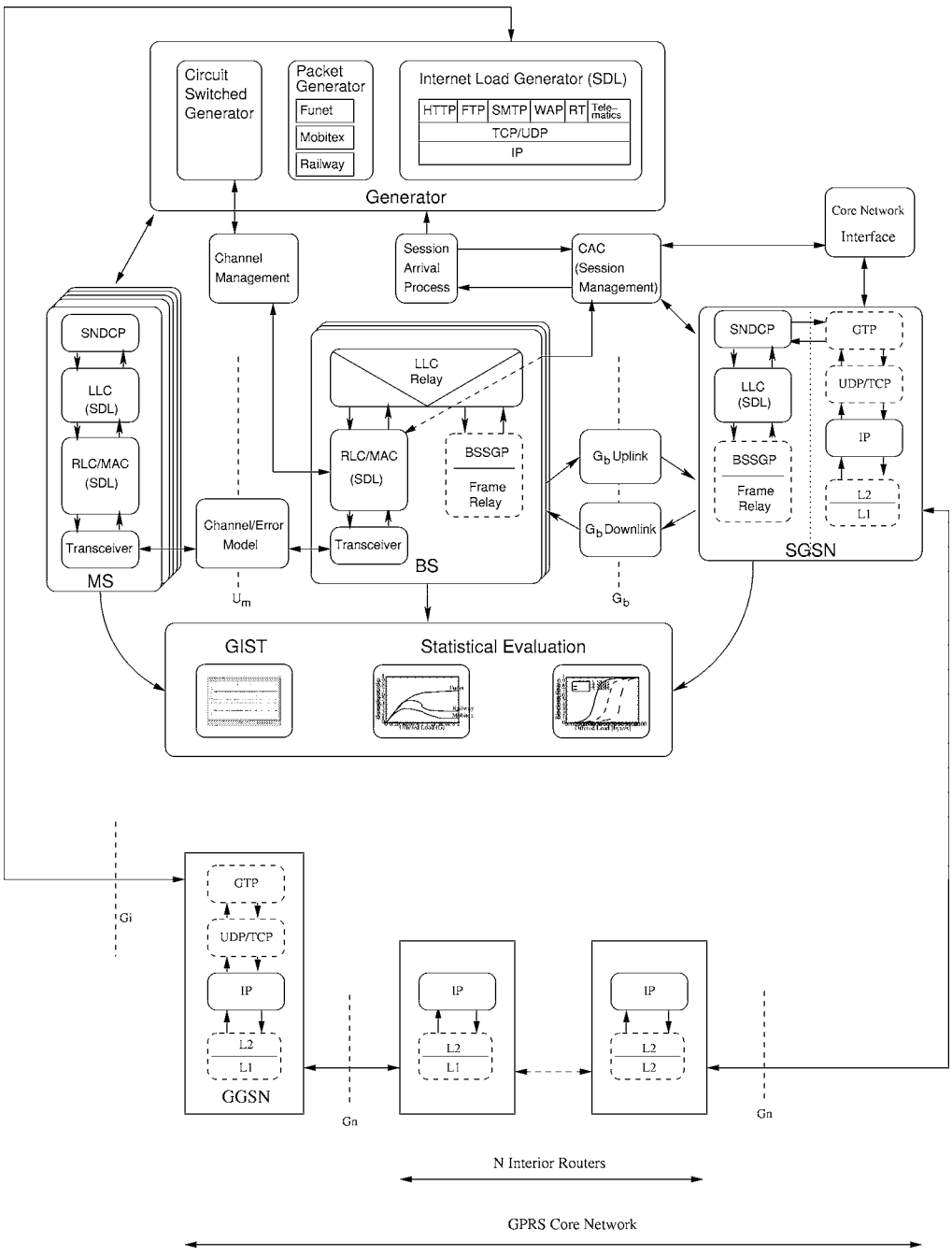

Figure 9. The GPRS Simulator GPRSim.

Table 10
Model parameters of Internet applications (WWW and e-mail).

WWW parameter	Distribution	Mean
Pages per session	geometric	5.0
Intervals between pages [s]	exponential	12.0
Objects per page	geometric	2.5
Object size [byte]	$\log_2$-Erlang-k	3700
e-mail parameter	**Distribution**	**Mean**
e-mail size [byte]	$\log_2$-normal	10000
Base quota [byte]	constant	300

In the following, the parameters of the two applications that specify the characteristic traffic load to the (E)GPRS are presented. Related documents can be found in [Arlitt and Williamson, 1; Paxson, 27]. The parameters of these models have been updated by parameters given by ETSI/3GPP suppositions for the behaviour of mobile Internet users [12] (see table 10).

5.1.1. WWW model

WWW sessions consist of requests for a number of *pages*. These pages consist of a number of *objects* with a dedicated *object size*. Another characteristic parameter is the delay between two pages depending on the user's behaviour to surf around the Web [Arlitt and Williamson, 1; 12]. Table 10 gives an overview of the WWW traffic parameters. The small number of objects per page (2.5 objects), and the small object size (3700 byte) were chosen, 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.

5.1.2. E-mail model

The e-mail model describes the traffic resulting from the download of an e-mail by an e-mail user. The relevant parameters are the amount of data per e-mail and its distribution (see table 10). A constant base quota of 300 byte has been added per e-mail [Paxson, 27]. The value of 10000 byte as the mean e-mail size has been chosen, since e-mails without large attachments have been assumed to be downloaded on mobile terminals.

5.1.3. Wireless application protocol (WAP) model

A WAP traffic model has been developed and applied in [Stuckmann et al., 30]. The main characteristics of the model are a very small mean packet size (511 byte) resulting from a log2-normal distribution with a limited maximum packet size of 1400 byte.

Since one of the main results in [Stuckmann et al., 30] is that one PDCH in GPRS can serve more than 20 WAP users with an acceptable QoS, WAP traffic is less important for traffic engineering of GPRS compared to WWW and e-mail. Owing to the small packet sizes WAP traffic can be multiplexed seamlessly with the other Internet traffic classes.

Therefore, WAP traffic is not further regarded in this article.

5.2. Transmission control protocol (TCP)

TCP has been implemented based on the description in [Stevens, 29] including slow start and congestion avoidance algorithms. According to version 1.0 of the Hypertext Transfer Protocol (HTTP) it is assumed that for each HTTP object a new TCP connection is set-up. Although in HTTP version 1.1 a TCP connection can be reused to transmit the a sequence of HTTP objects, several TCP connections can be set-up in parallel for the first HTTP objects. Since in the WWW model (see section 4.1.1) a small number of objects per page has been assumed, the probability for separate TCP connections for each object is high. Additionally HTTP objects may be located on different servers, i.e., separate TCP connections are needed. These considerations lead to our conclusion that our results are also valid for HTTP version 1.1.

5.3. Traffic generator for circuit-switched services

The circuit-switched (CS) traffic generator generates calls with a negative-exponentially distributed interarrival time. The call duration is also assumed to be negative exponentially distributed. The traffic load results from the mean values of the call interarrival and the call duration times and can be calculated from the Erlang-B formula [Kleinrock, 24].

5.4. Channel management

The Channel Management module in the simulator applies dynamic channel allocation to control the pool of GSM traffic channels (TCH) for GPRS and GSM applications as described in GSM 03.60 [16]. A TCH can be used for both a circuit-switched channel and a PDCH. 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 in the cell is free. All TCHs that are not used by CS connections are available for GPRS, if the number of PDCHs does not exceed the sum of maximum allowed fixed and on-demand PDCHs. It is assumed in the simulations presented that TCHs used for CS connections are placed adjacently beginning with channels not lying on a frequency used for GPRS. If a CS connection is released, the TCHs are shifted in a way so that TCHs and PDCHs are adjacent again after release. This mechanism is known as the repacking algorithm. Transitions, i.e., when the number of PDCHs available for GPRS changes, are immediately indicated to the Radio Resource Management (RRM) of RLC/MAC.

5.5. Air interface transmission error model

Within the air interface transmission error model it is decided whether a received data or control block is error-free or not. For this purpose a set of mapping curves is used gained from link level simulations that allow the mapping of a C/I value to the corresponding

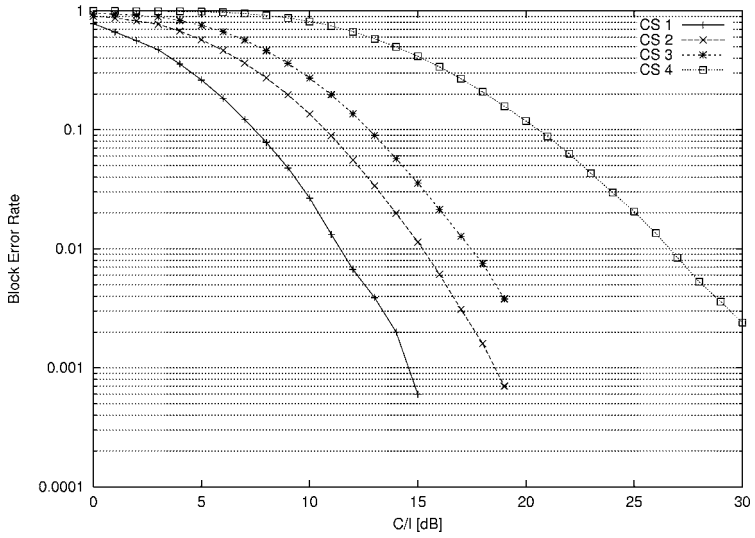

Figure 10. BLEP over C/I reference function used for the air interface error model.

block error rate (BLER) of a radio block [Furuskär et al., 18; Wigard and Mogensen, 37; Wigard et al., 38]. Figure 10 shows the BLER versus C/I results gained from link level simulations. The TU3 (Typical Urban) channel model of GSM 05.05 was assumed there [15].

5.6. Radio network QoS management

The functions not specified in detail in the GPRS specification are the CAC policy and the scheduling strategy. The implementation of these components in the GPRSim is depicted in the following.

5.6.1. Connection admission control

In the simulation model PDP requests are differentiated on subscriber base (Premium, Standard, Best-Effort (BE)) and application base (Conversational, Streaming, Interactive, Background). In this study only Interactive (WWW) and Background (e-mail) are regarded, since these are the applications predicted for GPRS in the next years. To avoid a total withdrawal of resources from the Standard traffic classes with lower QoS requirements, e.g., other than Conversational, there is a share reserved for this kind of traffic from the pool of radio resources in the cell. In general, all resources are open to traffic of any kind. In times of high load, however, 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 (see figure 11), where P and I represent the appropriate limits. When this limit is reached, the requested QoS is not accepted, but rather degraded to the next-lower-prioritized class.

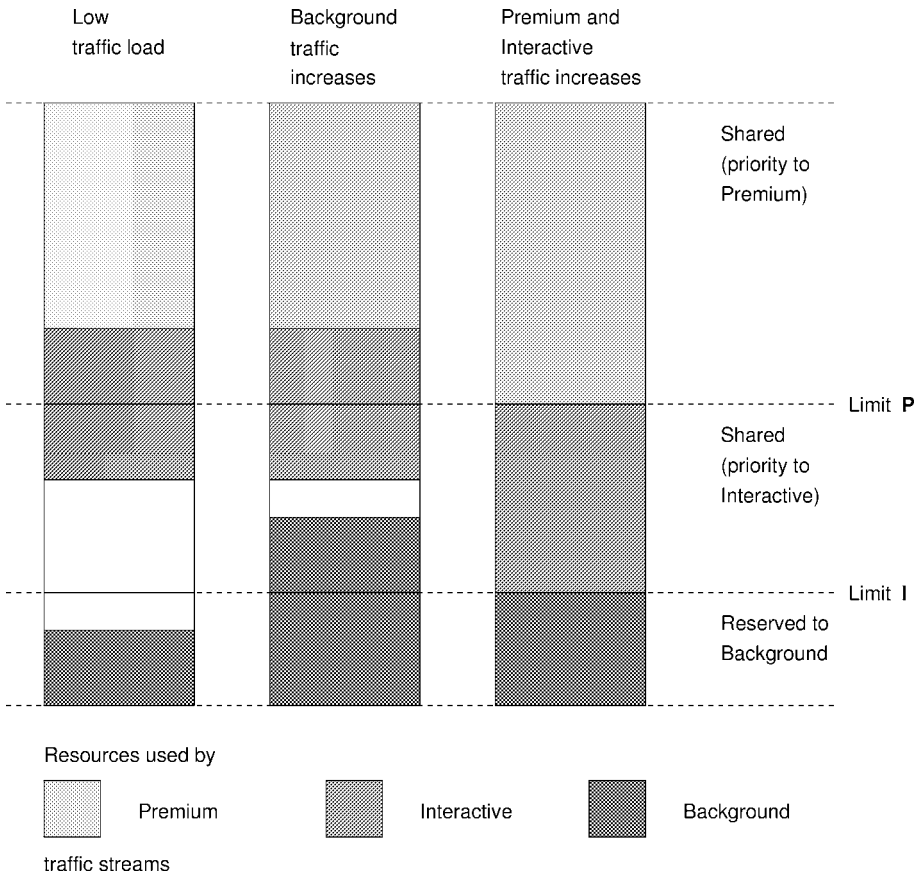

Figure 11. Admission control policy (example).

5.6.2. Scheduling in the BSS

Depending on the QoS profile negotiated the BS RLC/MAC layer performs the scheduling of the radio blocks. The scheduling mechanism implemented for both uplink and downlink direction follows a three-stage principle (see figure 12). First, incoming radio blocks are distributed into one of three queues according to the QoS subscription associated with the respective traffic flow. It is differentiated between Premium ("Gold Card"), Standard and Best-effort. The second stage is only valid for Standard service traffic. Depending on 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 packets are scheduled according to their TBF and a Round Robin (RR) algorithm with the depth of 20 radio blocks per scheduled TBF in the RR cycle. The third stage is built by a simple priority mechanism, serving the traffic class queues in order from highest priority (Premium) to lowest priority (Best-effort).

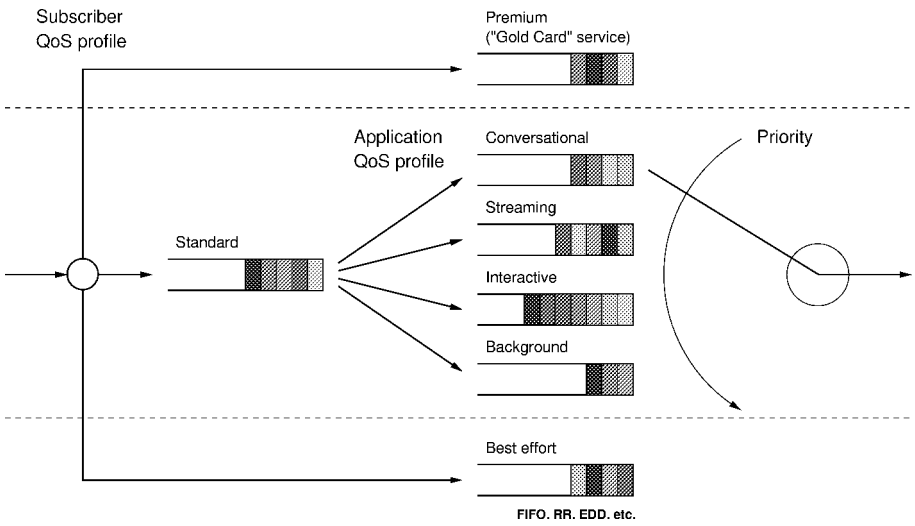

Figure 12. Principle of the scheduling function located in the BS RLC/MAC layer.

5.7. Core network model

An IP core network model consisting of boundary routers and a cascade of three interior routers is developed and integrated into the GPRS simulator. Classification and conditioning is performed by these boundary routers with simple scheduling and queuing performed by interior routers. The cascade model is suitable only as long as we are concerned with traffic on the same path between source and destination. Several different paths are not regarded. This model is sufficient to get statements about the capability of DiffServ to serve traffic even if routers in the core network are congested because of backbone traffic. CAC is realized by a bandwidth broker located at the boundary router, here the first SGSN connected to the BSS and the GGSN. It requests information of queues of all interior routers before admitting a request. In both boundary routers and interior routers the scheduling algorithm used is weighted round robin (WRR) scheduling. The reason for the choice is the DiffServ standard, which specifies that the interior routers should be free from complex methods and work performed by interior routers should be a minimum. Another method of scheduling that can be used is weighted fair queuing (WFQ) but since it would require that each interior router should also know about the flow states, namely, the number of flows passing through the router and their priority class, it is not regarded here.

6. Traffic performance evaluation

6.1. Simulation scenario parameter settings

The cell configuration is defined by the number of transceiver units (TRX) in the radio cell. Here a typical 3-TRX scenario is regarded with 0 and 1 fixed and 8 and 7 on-demand

Packet Data Channels (PDCH) that are shared with circuit switched GSM traffic, which is offered corresponding to an Erlang-blocking probability of 1%. This means that on average around 7 PDCHs are available for GPRS [Stuckmann and Müller, 32].

A constant RLC/MAC block error rate of 13.5% has been assumed throughout the simulations corresponding to a C/I of 12 dB. CS-2 is used as the coding scheme for user data.

LLC and RLC/MAC are operating in acknowledged mode. The multislot capability is one uplink and four downlink slots – a typical value for the first phases of GPRS operation. The MAC protocol instances in the simulation model are operating with three random access subchannels per 52-multiframe. LLC has a window size of 16 frames. TCP/IP header compression in SNDCP is performed. TCP is operating with a maximum congestion window size of 8 kbyte and a TCP Maximum Segment Size (MSS) of 536 byte. The transmission delay in the core network and external networks, i.e., the public Internet, is neglected. This corresponds to a scenario where the server is located in the operator's domain. The session interarrival time is assumed exponentially distributed with a mean of 12 seconds. The Internet traffic (see section 4) is composed of 70% e-mail sessions and 30% WWW sessions (see table 10) not depending on the subscription profile of the regarded MS. 10% of the mobile stations are representing Premium subscribers and 90% Standard subscribers.

6.2. Performance and system measures

As performance measures the downlink IP throughput per user during transmission period and the 95-percentile of the downlink IP packet delay are regarded. These are the QoS measures that are noticed by the user and that can be compared to the ETSI/3GPP QoS classes [14, 16]. For WWW and e-mail applications the throughput per user during transmission periods is the important measure since it mirrors the response time of a requested file.

The system measures comprise the downlink IP system throughput per radio cell and the downlink PDCH utilization, which is calculated by the total number of radio blocks carrying data or control information normalized to the total number of transmitted radio blocks. The measures are presented over the number of mobile stations (MS) offering GPRS traffic.

6.3. Simulation results neglecting the core network

Figure 13(a) shows the mean downlink IP system throughput per radio cell for 0 and 1 fixed PDCHs and with and without QoS management functions. The difference between the curves with 0 and 1 fixed PDCHs is very small since only in 1% of the time all PDCHs are allocated for circuit-switched calls. Since the offerd circuit-switched traffic is lower for the 1-fixed-PDCH scenario, the system throughput is 1–4% higher in the 0-fixed-PDCH scenario. As expected the system throughput for low load situations with less than 20 MS in the cell are nearly the same for the results for a Best-Effort (BE) service and a service with QoS functions. In higher load situations the system throughput

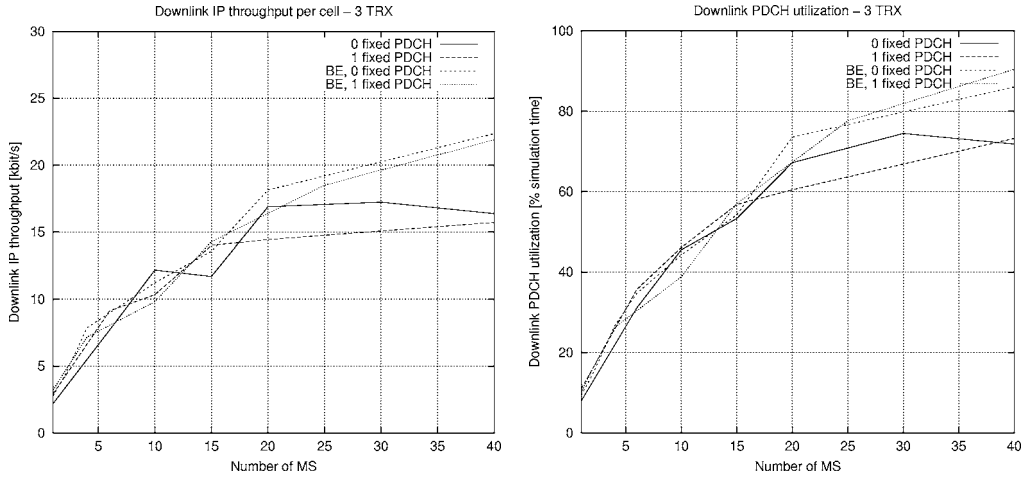

Figure 13. System measures with and without QoS management functions: (a) mean downlink IP system throughput per cell; (b) mean downlink PDCH utilization.

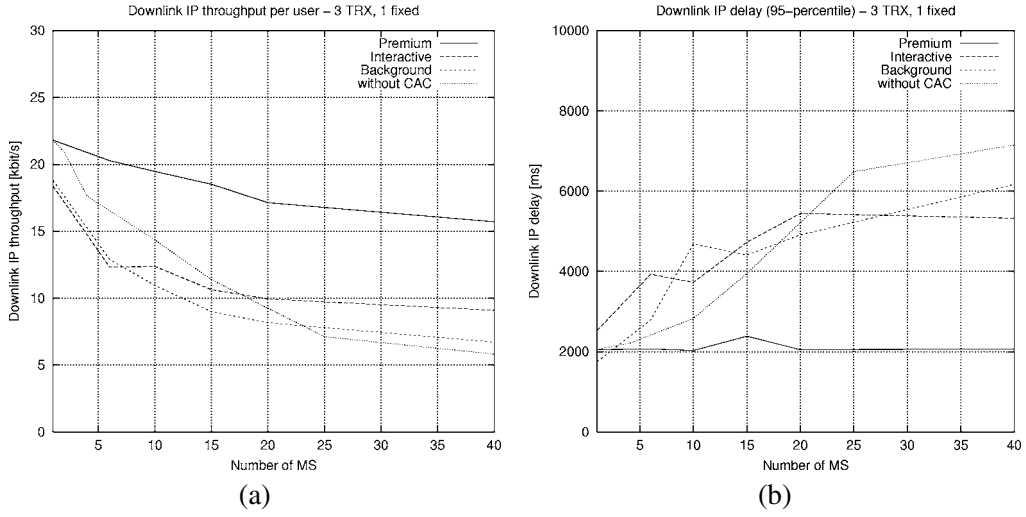

Figure 14. Performance measures for different subscriber and application classes: (a) mean downlink IP throughput per user; (b) 95-percentile of downlink IP packet delay.

comes into saturation. This can be explained by the effect that up to 28% of the Background sessions are terminated, when no IP packets are received for a period of more than 30 seconds (see figure 15). This does not occurr in the BE simulations. The same effect can be seen in figure 13(b), where the channel is not utilized with more than 75% in the results with QoS functions.

In figure 14(a) the downlink IP throughput per user during transmission periods for the different service and subscriber classes Premium, Interactive and Background

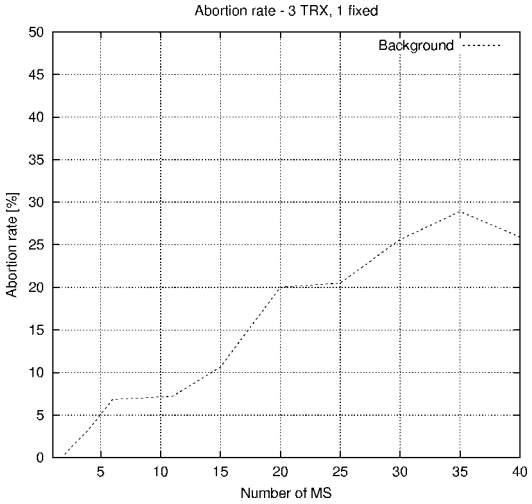

Figure 15. Abortion rate for Background applications.

compared with simulation results for a pure BE service (without CAC) is presented. In situations with low traffic load Standard users are losing 15–20% of performance compared to the BE service while the Premium user performance always remains higher than 15 kbit/s. In higher load situations the service differentiation between Interactive and Background becomes visible. While the throughput performance of the Interactive traffic does not fall below 10 kbit/s, the performance loss for Background applications is not visible in this measure. Nevertheless, more than 20% of the Background sessions are terminated because of poor performance as mentioned above. This can be avoided using fairer scheduling algorithms. The 95-percentile of the IP packet delay in figure 14(b) shows the similar effect.

6.4. Simulation results considering the core network

The core network model is configured with an output rate for each router of 2 Mbit/s and scheduling periods of 40 ms. Although backbone traffic of 1 and 2 Mbit/s is generated to model a low load scenario and a congestion scenario in the core network.

For each background load scenario, one simulation series is performed with QoS management in the radio and core network (interactive and background) and one with a pure Best-effort service without service differentiation in the radio and core network.

6.4.1. Low backbone traffic load

Scenarios were examined for service differentiation in the radio and core network. They are compared with the performance of the same traffic mix of WWW and e-mail, if no service differentiation in the radio and core network is done. Since the resources in the core network are not highly utilized by backbone traffic the core network can be seen as nearly transparent even for the Best-effort GPRS traffic.

Results very similar to the results shown in figures 14(a) and (b) were measured. In this scenario with 2 Mbit/s router output rate and only 1 Mbit/s backbone traffic DiffServ does not have a significant influence on the performance, since there is enough capacity left for the regarded GPRS traffic. However, it can be stated that the prioritization of the Interactive class is supported by DiffServ and the results are nearly equal to the case, where the influence of the core network is neglected.

6.4.2. Congested core network

More significant effects of DiffServ on the performance of different service classes become visible in a scenario with high backbone traffic load. In the regarded scenario the backbone traffic equals the output rate of the routers in the core network. Since the Best-effort GPRS traffic will be served with the same priority as the backbone traffic, the congestion in the core network will effect the GPRS traffic. This results in poor performance for Best-effort GPRS traffic. This is shown in figure 16 with average IP throughput values below 5 kbit/s even in situations with low GPRS traffic load. If WWW and e-mail traffic is not served as Best-effort traffic, but service differentiation both in the radio and the core network is performed, the same performance as in a low utilized core network can be reached through DiffServ QoS functions. Figure 16 shows that average downlink IP throughput values above 20 kbit/s are achieved in situations with low GPRS traffic load. With increasing GPRS traffic the performance decreases similarly as in the scenario without core network congestion (see figure 14(a)). This is only caused by the limited resources of 8 on-demand PDCHs in the GPRS radio network.

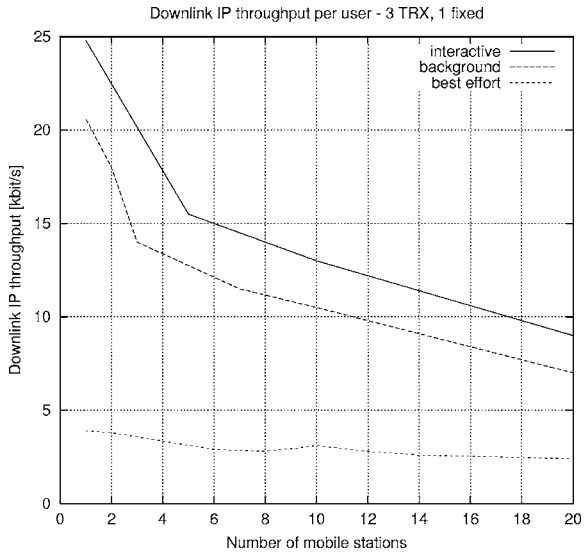

Figure 16. Mean downlink IP throughput for high backbone traffic load.

7. Conclusions

In this article, the capacity and performance gain achievable with quality of service functions in GPRS networks comprising Connection Admission Control (CAC) and scheduling with subscriber and service differentiation is examined. Simulation results show that Premium users can be served with nearly constant throughput and delay performance even if the number of active mobile stations in the radio cell rises to 40. 40 users instead of 12 in the pure best effort case can be served with a throughput performance for Interactive applications of 10 kbit/s, while the performance for Background users remains acceptable even in relatively high load situations. These results show that QoS functions in GPRS networks are increasing the application-specific performance significantly and realize the capability to serve subscribers and applications with respect to their QoS requirements.

Furthermore the capability of DiffServ to interwork with GPRS QoS functions is examined. To achieve this an IP core network model based on the DiffServ architecture, which is composed of two boundary routers and a cascade of three interior routers, was developed and integrated into the GPRS simulation tool GPRSim. With this model it has been shown that DiffServ is able to support service differentiation, which is done in the radio network based on GPRS/UMTS QoS classes, also in the core network. There is no difference to the performance of a scenario, where the core network influence is fully neglected. If the core network is congested, a significant advantage compared to a Best-effort service in the core network is achieved. While a congested core network without IP QoS functions would lead to poor performance for GPRS traffic, DiffServ is able to serve prioritized GPRS traffic similar to the performance without any influence of the core network. As a result DiffServ is capable to interwork with GPRS QoS functions, so that the core network can be seen as transparent and nearly without influence on the GPRS traffic performance even in the congestion case.

References

[1] M.F. Arlitt and C.L. Williamson, A synthetic workload model for internet mosaic traffic, in: *Proc. of the 1995 Summer Computer Simulation Conference*, Ottawa, Canada, July 1995, Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada, pp. 24–26.

[2] U. Black, *QoS in Wide Area Networks* (Prentice-Hall, Upper Saddle River, NJ, 2000).

[3] S. Blake, D. Black, M. Carlson, E. Davis, Z. Wang and W. Weiss, An architecture for differentiated services, Request for comments 2475, Technical Report, Internet Engineering Task Force (IETF) (December 1998).

[4] B. Braden, L. Zhang, S. Berson and S. Herzog, Resource reservation protocol, Request for comments 2205, Technical Report, Internet Engineering Task Force (IETF) (September 1997).

[5] R. Braden, D. Clark and S. Shenker, Integrated services in the Internet architecture, Request for comments 1633, Technical Report, Internet Engineering Task Force (IETF) (June 1994).

[6] G. Brasche and B. Walke, Concepts, services and protocols of the new gsm phase 2+ general packet radio service, IEEE Communications Magazine 35 (August 1997) 94–104.

[7] J. Cai and D. Goodman, General packet radio service in GSM, IEEE Communications Magazine 10 (October 1997) 122–131.

[8] I. Costa and L. Dell'Uomo, Quality of service in future 4G mobile wireless systems, in: *Proc. of the IEEE Internat. Conf. on 3rd Generation Wireless and beyond (3Gwireless '01)*, June 2001, pp. 502–507.
[9] A. Dutta-Roy, The cost of quality in Internet-style networks, IEEE Spectrum (September 2000) 57–62.
[10] ETSI 3GPP, Digital cellular telecommunications system (Phase 2+) (GSM); Universal mobile telecommunications system (UMTS); General packet radio service (GPRS); Service description; Stage 1 (3G TS 22.060 version 3.2.0, Release 1999), Technical specification ETSI TS 122 060, European Telecommunications Standards Institute, Sophia Antipolis, France (January 2000).
[11] ETSI 3GPP, Digital cellular telecommunications system (Phase 2+) (GSM); Universal mobile telecommunications system (UMTS); General packet radio service (GPRS); Service description; Stage 2 (3G TS 23.060 version 3.2.1, Release 1999), Technical specification ETSI TS 123 060, European Telecommunications Standards Institute, Sophia Antipolis, France (January 2000).
[12] ETSI 3GPP, Selection procedures for the choice of radio transmission technologies of the universal mobile telecommunication system UMTS (UMTS 30.03, 3G TR 101 112), Technical Report, European Telecommunications Standards Institute, Sophia Antipolis, France (April 1998).
[13] ETSI 3GPP, Universal mobile telecommunications system (UMTS); QoS concept and architecture (3G TS 23.107 version 3.1.0, Release 1999), Technical specification ETSI TS 123 107, European Telecommunications Standards Institute, Sophia Antipolis, France (January 2000).
[14] ETSI 3GPP, Universal mobile telecommunications system (UMTS); Service aspects; Services and service capabilities (3G TS 22.105 version 3.8.0, Release 1999), Technical specification ETSI TS 122 105, European Telecommunications Standards Institute, Sophia Antipolis, France (March 2000).
[15] ETSI, Digital cellular telecommunications system (Phase 2+) (GSM); Radio transmission and reception (GSM 05.05), Technical specification 5.2.0, European Telecommunications Standards Institute, Sophia Antipolis, France (January 1996).
[16] ETSI TC-SMG, Digital cellular telecommunications system (Phase 2+); General packet radio service (GPRS); Service description; Stage 2 (GSM 03.60 version 7.4.0, Release 1998), Draft European Standard ETSI EN 301 344, European Telecommunications Standards Institute, Sophia Antipolis, France (April 2000).
[17] ETSI TC-SMG, Digital cellular telecommunications system (Phase 2+); Mobile radio interface layer 3 specification (GSM 04.08 version 7.7.1, Release 1998), European standard ETSI EN 300 940, European Telecommunications Standards Institute, Sophia Antipolis, France (October 2000).
[18] A. Furuskär, S. Mazur, F. Müller and H. Olofsson, EDGE: Enhanced data rates for GSM and TDMA/136 evolution, IEEE Personal Communication Magazine (June 1999) 56–65.
[19] H. Gudding, Capacity analysis of GPRS, Master thesis, NTNU Norwegian University of Science and Technology, Trondheim, Norway (March 2000).
[20] ITU-T SG 10, Functional specification and description language (SDL), ITU-T recommendation Z.100, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland (1993).
[21] 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).
[22] M. Junius et al., 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).
[23] R. Kalden, I. Meirick and M. Meyer, Wireless internet access based on GPRS, IEEE Personal Communication Magazine 4 (April 2000) 8–18.
[24] L. Kleinrock, *Queueing Systems*, Vol. 1: *Theory* (Wiley, New York, 1975).
[25] R. Koodli and M. Puuskari, Supporting packet-data QoS in next-generation cellular networks, IEEE Communication Magazine 39 (February 2001) 180–188.

[26] K. Nichols, S. Blake, F. Baker and D. Black, Definition of the differentiated services field in the IPv4 and IPv6 headers, Request for comments 2474, Technical Report, Internet Engineering Task Force (IETF) (December 1998).

[27] V. Paxson, Empirically-derived analytic models of wide-area TCP connections, IEEE/ACM Transactions on Networking 2(4) (August 1994) 316–336; see also `ftp://ftp.ee.lbl.gov/papers/WAN-TCP-models.ps.Z`.

[28] M. Steppler, Performance analysis of communication systems formally specified in SDL, in: *Proc. of the 1st Internat. Workshop on Simulation and Performance '98 (WOSP '98)*, 1998, pp. 49–62.

[29] R. Stevens, *TCP/IP Illustrated*, Vol. 1 (Addison-Wesley, Reading, MA, 1996).

[30] P. Stuckmann, H. Finck and T. Bahls, A WAP traffic model and its appliance for the performance analysis of WAP over GPRS, in: *Proc. of the IEEE Internat. Conf. on 3rd Generation Wireless and beyond (3Gwireless '01)*, San Francisco, USA, June 2001, pp. 338–343.

[31] P. Stuckmann and J. Franke, The capacity and performance gain reachable with link quality control in EGPRS networks, in: *Proc. of the IEEE Internat. Conf. on 3rd Generation Wireless and beyond (3Gwireless '01)*, June 2001, pp. 781–786.

[32] P. Stuckmann and F. Müller, GPRS radio network capacity considering coexisting circuit-switched traffic sources, in: *Proc. of European Conf. on Wireless Technology (ECWT 2000)*, Paris, France, October 2000 (Miller Freeman, New York, 2000) pp. 66–69.

[33] P. Stuckmann and F. Müller, Quality of service management in GPRS networks, in: *Proc. of the IEEE Internat. Conf. on Networking (ICN '01)*, July 2001, Lecture Notes in Computer Science, Vol. 2093(1) (Springer, Berlin) pp. 276–285.

[34] P. Stuckmann, Simulation environment GPRSim: Tool for performance analysis, capacity planning and QoS enhancement in GPRS/EDGE networks, Technical Report, `http://www.comnets.rwth-aachen.de/~pst`.

[35] U. Vornefeld, Packet scheduling in SDMA based wireless networks, in: *Proc. of the Vehicular Technology Conference*, September 2000.

[36] B. Walke, *Mobile Radio Networks – Networking, Protocols and Traffic Performance*, 2nd ed. (Wiley, Chichester, 2001).

[37] J. Wigard and P. Mogensen, A simple mapping from C/I to FER and BER for a GSM type of air-interface, in: *Internat. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 1996, pp. 78–82.

[38] 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: *Vehicular Technology Conference (VTC)*, 1998, pp. 1588–1592.