Performance analysis of the general packet radio service

Christoph Lindemann [*], Axel Thümmler

Department of Computer Science, University of Dortmund, August-Schmidt-Strasse 12, 44227 Dortmund, Germany

Abstract

This paper presents an efficient and accurate analytical model for the radio interface of the general packet radio service (GPRS) in a GSM network. The model is utilized for investigating how many packet data channels should be allocated for GPRS under a given amount of traffic in order to guarantee appropriate quality of service. The presented model constitutes a continuous-time Markov chain. The Markov model represents the sharing of radio channels by circuit switched GSM connections and packet switched GPRS sessions under a dynamic channel allocation scheme. In contrast to previous work, the Markov model explicitly represents the mobility of users by taking into account arrivals of new GSM and GPRS users as well as handovers from neighboring cells. Furthermore, we take into account TCP flow control for the GPRS data packets. To validate the simplifications necessary for making the Markov model amenable to numerical solution, we provide a comparison of the results of the Markov model with a detailed simulator on the network level.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Mobile Internet services; Packet switched data transmission for mobile communication; Continuous-time Markov chains; Discrete-event simulation

1. Introduction

The general packet radio service (GPRS) is a standard from the European Telecommunications Standards Institute (ETSI) on packet data in GSM systems [10]. By adding GPRS functionality to the existing GSM network, operators can give their subscribers resource-efficient wireless access to external Internet protocol-based networks, such as the Internet and corporate Intranets. The basic idea of GPRS is to provide a packet-switched bearer service in a GSM network. As impressively demonstrated by the Internet, packet-switched networks make more efficient use of the resources for bursty data applications and provide more flexibility in general.

To evaluate the performance of GPRS, several simulation studies were conducted. Early simulation studies for GPRS have been reported in [6,7]. Meyer evaluated the performance of TCP over GPRS under several carrier to interference conditions and data coding schemes [13,17]. Malomsoky et al. developed a simulator for dimensioning GSM networks with GPRS [15]. Stuckmann and Müller developed a system simulator for GPRS

---

[*] Corresponding author.
E-mail address: cl@cs.uni-dortmund.de (C. Lindemann).
URL: http://www4.cs.uni-dortmund.de/~Lindemann/.

and studied the correlation of GSM and GPRS users for fixed and on-demand channel allocation techniques [18].

In previous work, several analytical models based on continuous-time Markov chains have been introduced for studying performance issues in GSM networks. Marsan et al. evaluated the impact of reserving channels for data and multimedia services on the performance in a circuit switched GSM network [1]. Marsan et al. developed an approximate analytical model for evaluating the performance of dual-band GSM networks [3]. Boucherie and Litjens developed a Markov model for analyzing the performance of GPRS under a given GSM call characteristic [5]. Markoulidakis et al. developed a Markov model for third generation mobile telecommunication systems [16]. They employed the Markov model for estimating the cell border crossing rate and the time it takes a busy mobile user to leave a cell area. Recently, Ermel et al. developed a Markov model for deriving blocking probabilities and average data rates for GPRS in GSM networks [9]. In none of these previous work, the question how many packet data channels (PDCH) should be allocated for GPRS for a given amount of traffic in order to guarantee appropriate quality of service (QoS) has been investigated.

This paper presents an efficient and accurate analytical performance model for the radio interface of the GPRS in a GSM network. The presented model constitutes a continuous-time Markov chain. The Markov model introduced in this paper represents the sharing of radio channels by circuit switched GSM connections and packet switched GPRS sessions under a dynamic channel allocation scheme. We assume a fixed number of physical channels permanently reserved for GPRS sessions and the remaining channels to be shared by GSM and GPRS connections. The model is utilized for investigating how many PDCH should be allocated for GPRS for a given amount of traffic in order to guarantee appropriate QoS. We present performance curves for average carried data traffic, packet loss probability, throughput per user, and queueing delay for different network configurations and traffic parameters.

In contrast to previous work, the Markov model explicitly represents the mobility of users by taking into account arrivals of new GSM and GPRS users as well as handovers from neighboring cells. Furthermore, we employ the traffic model defined by the 3rd Generation Partnership Project (3GPP) in [11] that can be effectively represented by an interrupted Poisson process (IPP), i.e., an on–off source. We consider a cluster comprising of seven hexagonal cells in an integrated GSM/GPRS network, serving circuit-switched voice and packet-switched data sessions. To allow the effective employment of numerical solution methods, the Markov model represents just one cell (i.e., the mid-cell) and employs the procedure for balancing incoming and outgoing handover rates introduced in [2]. To validate this simplification, we provide a comparison of the results of the Markov model with a detailed simulator implemented using the simulation library CSIM [8]. The simulator represents the entire cell cluster on the network level. Furthermore, an accurate implementation of the TCP flow control mechanism is included in the simulator. This validation shows that almost all performance curves derived from the Markov model lie in the confidence intervals of the corresponding curve of the simulator. Because of the employment of a numerical method for steady-state analysis, we can efficiently and accurately compute sensitive performance measures such as loss probabilities. In fact, using the presented Markov model sensitive performance measures can be computed on a modern PC within few minutes of CPU solution time. Note, that even with simulation runs in the order of hours proper estimates for such measures cannot be derived using discrete-event simulation because the large width of confidence intervals makes the results meaningless.

The remainder of the paper is organized as follows. Section 2 describes the basic GPRS network architecture and the radio interface which provide the technical background of the simulator and the analytical model. In Section 3, we describe the model and introduce its parameters. Section 4 derives the state space and driving processes for the analysis of the Markov model. Comprehensive performance studies for GPRS are presented in

Section 5. A detailed comparison of the performance between different network configurations and percentages of GPRS users is provided. Finally, concluding remarks are given.

2. General packet radio service

The introduction of GPRS as an additional service to GSM networks requires several modifications to the network architecture. The nodes corresponding to the access network and the circuit switched part of the core network already implemented in current GSM systems can be shared between GPRS and GSM. Two new node types, serving GPRS support node (SGSN) and gateway GPRS support node (GGSN), have to be introduced in the core network to handle packet switched data. The GGSN is the gateway node between an external packet-switched data network (e.g. IP, X.25) and the GPRS core network. In case of an external IP network, the GGSN is seen as an ordinary router serving all addresses that were static or temporarily assigned to the mobile stations (MS). Its task is to assign the correct SGSN for a MS depending on the location of the MS. The SGSN connects the GPRS core network and the radio access network, and switches the packets to the correct base station controller (BSC) via the Gb interface. The base transceiver station (BTS) is only a relay station without protocol functions. It performs the modulation of the carrier frequencies and demodulation of the signals. Fig. 1 shows the network architecture of GPRS in GSM networks.

On the physical layer, GSM uses a combination of frequency division multiple access (FDMA) and time division multiple access (TDMA) for multiple access. Two frequency bands are reserved for GSM operation, one for transmission from the mobile station to the BTS (uplink) and one for transmission from the BTS to the mobile station (downlink). Each of these bands is divided into 124 single carrier channels of 200 kHz width. A certain number of these frequency channels is allocated to a BTS, i.e., to a cell. Each of the 200 kHz frequency channels is divided into eight time slots that form a TDMA frame. A time slot lasts for a duration of 0.577 ms and carries 114 bits of information. The recurrence of one particular time slot defines a physical channel. GSM channels are called traffic channels (TCH) and channels allocated for GPRS are called PDCH.

The channel allocation in GPRS is different from the original allocation scheme of GSM. GPRS allows a single mobile station to transmit on multiple time slots of the same TDMA frame. This results in a very flexible channel allocation: one to eight time slots per TDMA frame can be allocated to one mobile station. On the other hand a time slot can be assigned temporarily to a mobile station, so that one to eight MS can use one time slot. Moreover, uplink and downlink channels are allocated separately, which efficiently supports asymmetric data traffic flows, i.e., non real-time traffic like WWW browsing or FTP.

In conventional GSM, a channel is permanently allocated for a particular user during the entire call period (whether data is transmitted or not). In contrast to this, in GPRS the channels are only allocated when data packets are sent or received, and they are released after the transmission. For bursty traffic this results in a much more efficient usage of the scarce radio resource. With this principle, multiple users can share one physical channel. GPRS includes the functionality to increase or decrease the amount of radio resources allocated to GPRS on a dynamic basis. The PDCHs are taken from the common pool of all channels available in the cell. The mapping of physical channels to either packet switched (GPRS) or circuit switched (conventional GSM) services can be performed statically or dynamically (capacity on demand), depending on the current traffic load. A load supervision procedure monitors the load of the
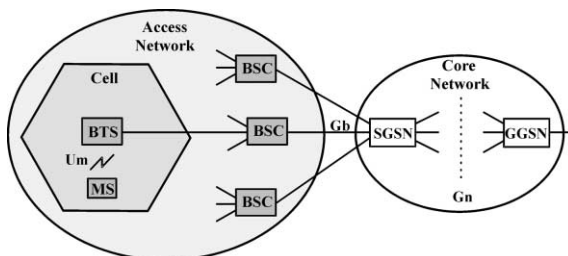

Fig. 1. Basic GPRS network architecture.

PDCHs in the cell. According to the current demand, the number of channels allocated for GPRS can be changed. Physical channels not currently in use by conventional GSM can be allocated as PDCHs to increase the QoS for GPRS. When there is a resource demand for services with higher priority, e.g. GSM voice calls, PDCHs can be deallocated.

3. Description of the Markov model of GPRS

Following [2], the performance model considers a single cell in an integrated GSM/GPRS network, serving circuit-switched voice and packet-switched data calls. We assume that GSM calls and GPRS calls arrive according to two mutually independent Poisson processes, with arrival rates λ_{GSM} and λ_{GPRS}, respectively. GSM calls are handled circuit-switched, so that one physical channel is exclusively dedicated to the corresponding mobile station. After the arrival of a GPRS call, a GPRS *session* begins. During this time, the BSC schedules the radio interface (i.e., the physical channels) among different GPRS users. GPRS users receive packets according to a specified traffic model explained below. The amount of time that a mobile station with an ongoing call remains within the cell is called *dwell time*. If the call is still active after the dwell time, a handover toward an adjacent cell takes place. The *call duration* is defined as the amount of time that the call will be active, assuming it completes without being forced to terminate due to handover failure. We assume the dwell time to be an exponentially distributed random variable with mean $1/\mu_{h,GSM}$ for GSM calls and $1/\mu_{h,GPRS}$ for GPRS sessions. The call durations are also exponentially distributed with mean values $1/\mu_{GSM}$ and $1/\mu_{GPRS}$ for GSM calls and GPRS sessions, respectively. In order to limit the amount of packet traffic in the cell we restricted the maximal number of active GPRS sessions by a value M. This provides a form of first comes first served admission control in order to guarantee certain QoS for the GPRS users.

To exactly model the user behavior in the cell, we consider the handover flow of active GSM calls and GPRS sessions from adjacent cells. It is impossible to specify in advance the intensity of the incoming handover flow. This is due to the fact that the handover rate out of the cell depends on the number of active customers within the cell. On the other hand, the handover rate into the cell depends on the number of customers in the neighboring cells. Thus, the iterative procedure introduced in [2] is employed for balancing the incoming and outgoing handover rates. The iteration is based on the assumption that the incoming handover rate $\lambda_{h,GSM}^{(i+1)}$ of GSM calls and $\lambda_{h,GPRS}^{(i+1)}$ of GPRS sessions at step $i+1$ is equal to the corresponding outgoing handover rate computed at step i.

Since in the end-to-end data path, the wireless link is typically the performance bottleneck, the model represents the radio interface of an integrated GSM/GPRS network. The functionality of the GPRS core network is not included. Because of the anticipated traffic asymmetry (most of the GPRS traffic will be WWW browsing), the model focuses on resource contention in the downlink (i.e., the path BSC → BTS → MS) of the radio interface. The amount of uplink traffic, e.g., induced by acknowledgments, is assumed to be negligible. The arrival stream of data packets is modeled at the network layer, assuming a data packet size of 480 byte [11]. Data packets arriving at the BSC are stored in a FIFO buffer with limited size of K data packets until they are transmitted on a free physical channel. Let N be the overall number of physical channels available in the cell. We assume that N_{GPRS} channels are permanently reserved as PDCHs for GPRS and the remaining $N_{GSM} = N - N_{GPRS}$ channels can be used either as GSM TCH or "on-demand" as PDCHs. Among the on-demand channels, GSM calls have priority. That is on-demand channels allocated as PDCH are immediately released, when requested by a GSM call. Fig. 2 illustrates the partitioning scheme of physical channels in GSM TCH and GPRS PDCH.

In order to describe the GPRS traffic, we adopt the model defined by the 3GPP in [11]. Active users within a cell execute a *packet service session*, which is an alternating sequence of *packet calls* and *reading times* (see Fig. 3). During a packet call several packets may be generated. Therefore, a
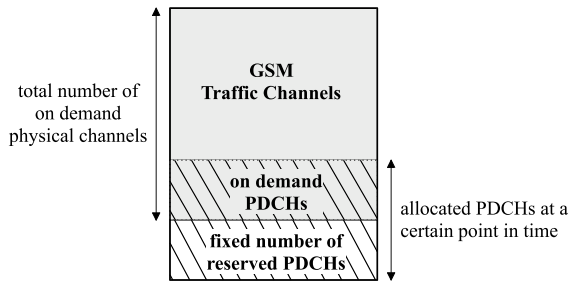
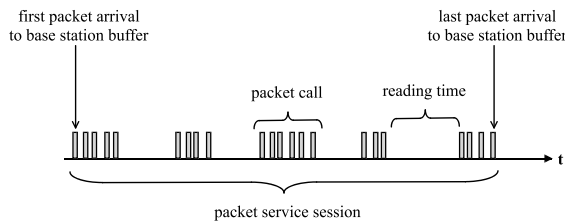Fig. 2. Partitioning of physical channels among GSM and GPRS.

Fig. 3. Typical characteristic of a packet service session [11].

Fig. 4. IPP traffic model for one GPRS session.

packet call constitutes a bursty sequence of packets. Note, that the burstiness during a packet call is a characteristic feature of packet transmissions that must be taken into account in an accurate traffic model [4]. For example in a WWW browsing session a packet call corresponds the downloading of a WWW document. After the document is entirely arrived to the terminal, the user is consuming certain amount of time, i.e., the reading time, studying the information. It is also possible that the packet service session contains only one packet call. In fact this is the case for a file transfer via FTP.

According to [11], the number of packet calls within a packet session should be a geometrically distributed random variable with mean N_{pc}. The reading time between packet calls is an exponentially distributed random variable with parameter $1/D_{pc}$. Each packet call comprises a geometrically distributed number of data packets with mean N_d and the interarrival time between packets in a packet call is an exponentially distributed random variable with parameter $1/D_d$.
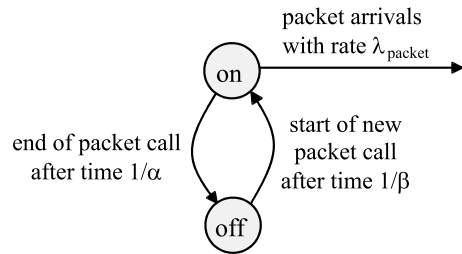
The traffic model described in [11] can be represented by a Markov modulated Poisson process (MMPP) [12]. In particular, we consider an MMPP with two alternating states named ''on'' and ''off''. The on-state corresponds to an active packet call of one GPRS user and the off-state represents the reading time of the GPRS user (see Fig. 4). During the on-state packets are generated by an exponentially distributed random variable with parameter $\lambda_{packet} = 1/D_d$. In the off-state no packets are generated. The average on and off times are exponentially distributed random variables with parameter $\alpha = 1/(N_d D_d)$ and $\beta = 1/D_{pc}$. Thus, we consider a special case of a MMPP, i.e. an IPP. The average packet service session time corresponds to the GPRS session duration $1/\mu_{GPRS} = N_{pc}(D_{pc} + N_d D_d)$.

Because most of today's Internet traffic (around 90%) is transported via the TCP/IP protocol, we take into account a TCP flow control mechanism in the Markov model. In case of network congestion, the buffer of the router at the beginning of the bottleneck link, i.e., the BSC, overflows and packets get lost. TCP detects lost packets due to timer expiration or the reception of three duplicate acknowledgements from the receiver and reacts on these losses by reducing the packet sending rate of the source. As illustrated by the validation with simulation results in Section 5.2, the following approximate representation of a TCP flow control in the Markov model can effectively represent the reaction of TCP sources to network congestion, i.e., buffer overflow at the BSC. Therefore, in the Markov model the sending rate of the TCP sources is reduced when the buffer occupancy exceeds a certain percentage η of the buffer size K.

To provide a reliable wireless link for data transfer a forward error correction (FEC) mechanism on the physical layer as well as an automatic repeat request (ARQ) mechanism in the radio link control (RLC) protocol on the link layer are specified for GPRS [7]. For FEC four different coding schemes CS-1 to CS-4 based on convolutional coding are currently defined. Coding scheme CS-1 corresponds to the coding required for a channel with high block error rate, i.e., code rate 1/2, and coding scheme CS-4 corresponds to no coding, i.e., code rate 1. In order to take into account the influence of block errors on performance measures we consider the fixed coding scheme CS-2 in the Markov model. It allows a data transfer rate of 13.4 Kbit/s on each PDCH [7].

Note, that we do not consider packet losses due to interference on the wireless link. Because TCP is unaware of these losses, it could be possible that TCP will react wrong on wireless losses with congestion control, i.e., slowing down its sending rate. As shown in [17], GPRS provides a sufficiently fast working ARQ mechanism, which allows typically several retransmissions before TCP recognizes a loss due to timer expiration. Therefore, TCP observes just packet delays rather than losses. Nevertheless, in the Markov model we assume that almost all packet losses can be recovered by the FEC mechanism of the coding scheme and therefore no retransmissions of lost packets are necessary. Taking into account packet retransmissions that would lead to a decrease in overall throughput could be considered in future work.

4. Analysis of the Markov model

4.1. State definition and derivation of transition rates

The analysis of the GPRS model introduced in Section 3 is performed by means of a continuous time Markov chain. From the steady state distribution of the Markov chain performance measures of interest can be computed. A *state* of the model representing the considered cell is determined by the number of GSM connections currently active, denoted by n $(0 \leqslant n \leqslant N_{\text{GSM}})$, the number of active

GPRS sessions, denoted by m $(0 \leqslant m \leqslant M)$, the number of packets in the BSC buffer denoted by k $(0 \leqslant k \leqslant K)$, and the states r_i of the two-state MMPPs for active GPRS sessions with $0 \leqslant i \leqslant m$. As a consequence, the state can be specified as the vector $s = (n, k, m, r_1, \ldots, r_M)$ with $r_i = 1$ or $r_i = 2$ for $1 \leqslant i \leqslant m$ and $r_i = 0$ for $m + 1 \leqslant i \leqslant M$. This leads to $(2^{M+1} - 1)(N_{\text{GSM}} + 1)(K + 1)$ feasible states. Due to its large state space, such a Markov model can be analyzed by discrete-event simulation only. Making the common assumption that all GPRS users behave statistically identical, allows us to derive an aggregated Markov model whose state space is tractable for numerical solution. The rationale behind the aggregation lies in the fact that m identical two-state MMPPs corresponding to m active GPRS sessions can be represented by one MMPP with $m + 1$ states [12]. Employing this aggregation, the state of the Markov model for the cell can be expressed by a vector $s = (n, k, m, r)$. In this tuple, r represents the state of the MMPP for m concurrently active GPRS sessions. The state r of the aggregated MMPP models that r MMPPs of individual GPRS sessions are in off-state and the remaining $m - r$ MMPPs are in on-state. This reduces the state space significantly to an overall number of $\frac{1}{2}(M + 1)(M + 2)(N_{\text{GSM}} + 1)(K + 1)$ states.

The behavior of GSM users in the considered cell can be represented by an $M/M/c/c$ queue with $c = N_{\text{GSM}}$ servers. This is because GSM users are not effected by data traffic of GPRS sessions due to their higher priority. The arrival process of GSM voice calls is the superposition of two Poisson processes corresponding to newly arriving voice calls and incoming handover requests. Therefore, the arrival rate of the $M/M/c/c$ queue is given by $\lambda_{\text{GSM}} + \lambda_{\text{h,GSM}}$. In the same way, the service rate of the $M/M/c/c$ queue is derived as $\mu_{\text{GSM}} + \mu_{\text{h,GSM}}$. Moreover, the behavior of GPRS users can be represented in the same way by an $M/M/c/c$ queue with $c = M$ servers and arrival and service rates $\lambda_{\text{GPRS}} + \lambda_{\text{h,GPRS}}$ and $\mu_{\text{GPRS}} + \mu_{\text{h,GPRS}}$, respectively. The Markov model constitutes a compound queueing system whose arrival process is governed by the number of active GPRS users (i.e., the customers of the latter $M/M/c/c$ queue) and whose service process is governed by the number

of active GSM connections (i.e., the customers of the former $M/M/c/c$ queue).

Let S be the state space of the Markov model we just described. For notational convenience, we enumerate its states from 0 to $S_{\max}$. The model dynamics are determined by the underlying continuous-time Markov chain which cause state transitions at random instants. State transitions correspond to different kinds of events that must be processed in the cell. The following kinds of events may occur:

(i) incoming GSM calls and handovers,
(ii) incoming GPRS sessions and handovers,
(iii) leaving GSM calls due to completion or handover,
(iv) leaving GPRS sessions due to completion or handover,
(v) arrivals of data packets,
(vi) service of data packets,
(vii) state changes of the MMPP to a more bursty or less bursty arrival of data packets.

One can easily show that the continuous-time Markov chain underlying the Markov model has finite state space and is homogeneous and irreducible. Thus, the steady state distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_{S_{\max}})$ can be computed through the matrix equation $\boldsymbol{\pi} \cdot \mathbf{Q} = 0$ together with the normalization condition $\sum_{i=0}^{S_{\max}} \pi_i = 1$. Here, $\mathbf{Q}$ denotes the infinitesimal generator matrix. The transition rates, i.e. the entries of matrix $\mathbf{Q}$, are obtained from the analysis of the system events (i)–(vii). For each event, it is possible to determine what state transitions can happen, i.e. what are the possible successor states of a generic state $s = (k, n, m, r)$. This is what we discuss next, referring to Table 1 which shows the conditions on the model state for a transition to be possible, the rate associated with the transition, and the successor state, for each type of events.

Incoming GSM calls and handovers are accepted in the cell if the number of free channels, excluding those reserved as PDCHs, is such that the call can be accommodated. Incoming GPRS sessions and handovers are accepted in the cell if the maximal number of GPRS users M is not reached. A new GPRS session in the cell starts sending packets according to an MMPP described above. We assume the MMPP to start in steady state, that is in on-state with probability $\beta/(\alpha + \beta)$ and in off-state with probability $\alpha/(\alpha + \beta)$. This assumption guaranties that the MMPP is still in steady state when the GPRS session is terminated. Both the completion of calls and the outgoing handover requests have the effect of freeing a channel in the cell. Thus with n active GSM calls the rate of freeing a channel is $n(\mu_{\text{GSM}} + \mu_{\text{h,GSM}})$.

Table 1
Transitions from a state (k, n, m, r) in the Markov chain

Event type	Condition	Successor state	Rate
GSM call arrival	$n < N_{\text{GSM}}$	$(k, n+1, m, r)$	$\lambda_{\text{GSM}} + \lambda_{\text{h,GSM}}$
GPRS session arrival	$m < M$	$(k, n, m+1, r)$	$\frac{\beta}{\alpha+\beta}(\lambda_{\text{GPRS}} + \lambda_{\text{h,GPRS}})$
		$(k, n, m+1, r+1)$	$\frac{\alpha}{\alpha+\beta}(\lambda_{\text{GPRS}} + \lambda_{\text{h,GPRS}})$
GSM call leaving cell	$n > 0$	$(k, n-1, m, r)$	$n(\mu_{\text{GSM}} + \mu_{\text{h,GSM}})$
GPRS session leaving cell	$(m > 0) \wedge (r = 0)$	$(k, n, m-1, r)$	$m(\mu_{\text{GPRS}} + \mu_{\text{h,GPRS}})$
	$(m > 0) \wedge (r = m)$	$(k, n, m-1, r-1)$	$m(\mu_{\text{GPRS}} + \mu_{\text{h,GPRS}})$
	$(m > 0) \wedge (0 < r < m)$	$(k, n, m-1, r-1)$	$\frac{r}{m}m(\mu_{\text{GPRS}} + \mu_{\text{h,GPRS}})$
		$(k, n, m-1, r)$	$\frac{m-r}{m}m(\mu_{\text{GPRS}} + \mu_{\text{h,GPRS}})$
Packet arrival	$(k \leqslant \eta K) \wedge (m > 0)$	$(k+1, n, m, r)$	$(m-r)\lambda_{\text{packet}}$
	$(\eta K < k < K) \wedge (m > 0)$	$(k+1, n, m, r)$	$\min\{(m-r)\lambda_{\text{packet}}, \min(N-n, 8k)\mu_{\text{service}}\}$
Packet service	$\min(N - n, 8k) > 0$	$(k-1, n, m, r)$	$\min(N-n, 8k)\mu_{\text{service}}$
MMPP less bursty	$r < m$	$(k, n, m, r+1)$	$(m-r)\alpha$
MMPP more bursty	$r > 0$	$(k, n, m, r-1)$	$r\beta$

For GPRS sessions leaving the cell we have to distinguish if the packet arrival process of the terminated session is in on-state or off-state. In state (k,n,m,r) are r GPRS sessions (out of m) in off-state and the remaining $m-r$ sessions are in on-state. Therefore, the probability that the leaving session is in off-state is r/m and that it is in on-state is $(m-r)/m$.

If the number of data packets queued in the BSC buffer is less or equal ηK the arrival rate of data packets is determined by the number of active GPRS sessions in the cell and by the state of the aggregated MMPP. In this case the average arrival rate of data packets corresponds to $m-r$ sessions in on-state. The same argument holds for the time spent in a particular state of the $(m+1)$-state Markov chain controlling the arrival process of data packets. With rate $(m-r)\alpha$ the aggregated MMPP changes to a less bursty state, i.e. one GPRS session changes from on-state to off-state, and with rate $r\beta$ it changes to a more bursty state, respectively. For a queue length of more than ηK data packets the arrival rate is simply bounded by the service rate. In the service process for data packets, the PDCHs are fairly shared by all packets in transfer up to a maximum of 8 PDCHs per data packet (multislot mode) and a maximum of 8 packets per PDCH [10]. With k packets residing in the BSC buffer a maximum of $8k$ PDCHs could be used for data transfer. We assume that at each time all free on-demand channels are allocated as PDCHs. Furthermore, N_{GPRS} fixed PDCHs are utilized for data transfer. This results in an overall number of $N-n$ physical channels that are available for the transfer of GPRS packet data. Putting it altogether, we get a utilization of $\min(N-n, 8k)$ PDCHs in state (k,n,m,r).

4.2. Derivation of performance measures

Recall that the arrival and service behavior for GSM calls and GPRS sessions constitute a $M/M/c/c$ queueing systems. Since the steady state solution for such a queue is known in closed-form, we can immediately derive performance measures.

With

$$
\rho_{\text{GSM}} = \frac{\lambda_{\text{GSM}} + \lambda_{\text{h,GSM}}}{\mu_{\text{GSM}} + \mu_{\text{h,GSM}}},
$$
$$
\rho_{\text{GPRS}} = \frac{\lambda_{\text{GPRS}} + \lambda_{\text{h,GPRS}}}{\mu_{\text{GPRS}} + \mu_{\text{h,GPRS}}},
\tag{1}
$$

the steady state solutions $\pi_{\text{GSM},n}$ for n active GSM calls in the cell and $\pi_{\text{GPRS},m}$ for m active GPRS sessions in the cell is given by

$$
\pi_{\text{GSM},0} = \left(\sum_{n=0}^{N_{\text{GSM}}} \frac{\rho_{\text{GSM}}^n}{n!} \right)^{-1},
$$
$$
\pi_{\text{GSM},n} = \pi_{\text{GSM},0} \frac{\rho_{\text{GSM}}^n}{n!} \quad \text{for } n = 1, 2, \ldots, N_{\text{GSM}},
\tag{2}
$$

$$
\pi_{\text{GPRS},0} = \left(\sum_{m=0}^{M} \frac{\rho_{\text{GPRS}}^m}{m!} \right)^{-1},
$$
$$
\pi_{\text{GPRS},m} = \pi_{\text{GPRS},0} \frac{\rho_{\text{GPRS}}^m}{m!} \quad \text{for } m = 1, 2, \ldots, M.
\tag{3}
$$

We apply the steady-state solutions (2) and (3) to iteratively balance the handover flows of GSM calls and GPRS sessions in advance. Assuming that in steady state the average handover flow entering the cell equals the average handover flow leaving the cell and the initialization $\lambda_{\text{h,GSM}}^{(0)} = \lambda_{\text{GSM}}$ and $\lambda_{\text{h,GPRS}}^{(0)} = \lambda_{\text{GPRS}}$, the handover flows can be balanced as follows ($i \geqslant 0$):

$$
\lambda_{\text{h,GSM}}^{(i+1)} = \mu_{\text{h,GSM}} \sum_{n=1}^{N_{\text{GSM}}} n\pi_{\text{GSM},n}^{(i)} \quad \text{for GSM calls,} \tag{4}
$$

$$
\lambda_{\text{h,GPRS}}^{(i+1)} = \mu_{\text{h,GPRS}} \sum_{m=1}^{M} m\pi_{\text{GPRS},m}^{(i)}
$$

for GPRS sessions. $\tag{5}$

Furthermore, from the solution (2) and (3) we can calculate performance measures such as carried voice traffic (CVT) and average number of GPRS sessions (AGS):

$$
\text{CVT} = \sum_{n=1}^{N_{\text{GSM}}} n\pi_{\text{GSM},n},
\tag{6}
$$

$$AGS = \sum_{m=1}^{M} m\pi_{GPRS,m}. \tag{7}$$

The GSM call blocking probability and GPRS session blocking probability is simply given by the steady-state probabilities $\pi_{GSM,N_{GSM}}$ and $\pi_{GPRS,M}$, respectively.

In the following, we show how to compute additional performance indices that are plotted in the curves presented in Section 5. These performance measures are obtained from the steady state solution π of the Markov model that can easily be computed numerically as explained in Section 4.1. The carried data traffic (CDT) is the average number of channels in use for data transfer, i.e., PDCHs, and is given by

$$CDT = \sum_{i=0}^{S_{max}} n(i)\pi_i, \tag{8}$$

where $n(i)$ is the number of PDCH utilized in state i and π_i is the steady-state probability of state i. Furthermore, we can derive the average packet arrival rate λ_{avg} from the steady-state distribution by summing up the arrival rates in states i weighted by the probabilities π_i. The packet loss probability (PLP) is the probability that an arriving data packet finds a BSC buffer with already K packets queued and, thus, cannot be stored. It can be computed from the average packet arrival rate and the overall throughput of data packets $CDT \mu_{service}$:

$$PLP = 1 - \frac{CDT \mu_{service}}{\lambda_{avg}}. \tag{9}$$

The queueing delay (QD) is the time packets are waiting in the BSC queue until a free PDCH is available for transfer. It can be computed by the quotient of the mean queue length (MQL), which can be directly derived from the steady state distribution, and the overall throughput of data packets:

$$QD = \frac{MQL}{CDT \mu_{service}}. \tag{10}$$

A last performance measure of interest is the average throughput per user (ATU) that can be derived by the overall throughput of data packets

and the average number of GPRS sessions in the cell:

$$ATU = \frac{CDT \mu_{service}}{AGS}. \tag{11}$$

5. Performance results

5.1. The base parameter setting

The base parameter setting underlying the performance experiments are summarized in Table 2. These values are used for the derivation of all numerical results unless specified otherwise. The overall number of physical channels in a cell is set to $N = 20$ among which at least one channel is reserved for GPRS. Our study is mainly focussed on the introduction of GPRS into the GSM network. Therefore, we assume as base value that only 5% of the arriving calls corresponding to GPRS session requests and the remaining 95% are GSM calls. GSM call duration is set to 120 s and call dwell time to 60 s, so that users make 1–2 handovers on average. These values are quite often used in design and planning of mobile telephony systems. For GPRS sessions the average session duration is obtained from the different traffic model parameters described below. The session dwell time is assumed to be 120 s. We assume slower movement of GPRS users than for GSM users because higher visual attention is required for GPRS services like WWW browsing that do not allow fast movement in many cases. In all experiments, we fix the modulation and coding

Table 2
Base parameter setting of the Markov model of GPRS

Parameter	Base value
Number of physical channels, N	20
Number of fixed PDCHs, N_{GPRS}	1
BSC buffer size, K	100 data packets
Transfer rate for one PDCH (CS-2), $\mu_{service}$	13.4 Kbit/s
Average GSM voice call duration, $1/\mu_{GSM}$	120 s
Average GSM voice call dwell time, $1/\mu_{h,GSM}$	60 s
Average GPRS session dwell time, $1/\mu_{h,GPRS}$	120 s
Percentage of GSM users	95%
Percentage of GPRS users	5%

scheme to CS-2 [7]. It allows a data transfer rate of 13.4 Kbit/s on each PDCH.

Traffic models are derived from the traffic parameter characterization defined by the 3GPP in [11]. In particular, we consider two traffic models, (1) for 8 Kbit/s and (2) for 32 Kbit/s WWW browsing. In both models, the average number of packet calls per session is $N_{pc} = 5$, the average reading time between packet calls is $D_{pc} = 412$ s. The average number of data packets within a packet call is $N_d = 25$. The models only differ in the burstiness of the packet arrival process. That is the interarrival time between data packets during a packet call. For the 8 Kbit/s model $D_d = 0.5$ and for 32 Kbit/s $D_d = 0.125$, respectively. The corresponding parameters of the Markov model are obtained as described in Section 3.

Table 3 specifies the parameters of the traffic models. Traffic model 1 corresponds to 8 Kbit/s bandwidth and traffic model 2 corresponds to 32 Kbit/s bandwidth for WWW browsing, respectively. As we will observe from the curves presented in Section 5.3, these two traffic models produce a low traffic load that can be managed by one or two PDCHs. In order to study the cell under heavier traffic load and therefore the usage of on-demand PDCHs, we introduce a third traffic model. This model is obtained from traffic model 2 by setting the off-duration of the traffic process equal to the on-duration and assuming a GPRS session duration for 50 packet calls. The refined traffic model corresponds to traffic model 3 in Table 3.

Table 3
Parameter setting of different traffic models

Parameter	Traffic model 1	Traffic model 2	Traffic model 3
Maximum number of active GPRS sessions, M	50	50	20
Average GPRS session duration, $1/\mu_{GPRS}$ (s)	2122.5	2075.6	312.5
Average arrival rate of data packets, λ (Kbit/s)	8	32	32
Average duration of a packet call, $1/\alpha$ (s)	12.5	3.1	3.1
Average reading time between packet calls, $1/\beta$ (s)	412	412	3.1

5.2. Validation of the Markov model of GPRS

Recall that the major simplification in the Markov model of GPRS stems from the choice of studying just one cell in isolation, instead of considering the entire cell cluster and the interactions among adjacent cells. This simplification relies on the assumption that under operating conditions of the cellular network (i.e., in steady-state) the average incoming handover flow is equal to the average outgoing handover flow. Furthermore, the model consists of a simplified TCP flow control mechanism. To validate these simplifications of the Markov model, we additionally implemented a detailed simulator using the simulation library CSIM [8]. This simulator represents a cellular network comprising seven hexagonal cells and takes explicitly into account the handover procedures for GSM and GPRS users. Moreover, the transmission of data packets over the wireless link is modeled in more detail than in the Markov model. That is, we explicitly consider the segmentation of data packets into TDMA frames. Furthermore, all relevant TCP mechanisms, such as slow start, congestion avoidance, and retransmission based on both timeouts and duplicate acknowledgements, have been implemented. The simulation results of the mid-cell of the cell cluster are compared with corresponding results obtained from the Markov model. Confidence intervals with confidence level of 95% for simulation results are computed using batch means. For the validation we considered traffic model 3 because in this configuration most valuable statements can be derived from the presented experiments.

In the first experiment, we determine the optimal value for the threshold η in order to closely approximate the flow control of TCP. Recall that in the Markov model the arrival rate of data packets slows down, when the queue at the BSC reaches a length of more than ηK packets. Fig. 5 shows the PLP for different values of η in comparison to the simulation result. The borders of the confidence intervals are drawn as dashed lines. Numerical results are drawn in solid lines. From the curves, we conclude that a setting η equal to 0.7 is optimal for modeling a TCP flow control in the Markov model. A value of η below 0.7 slows
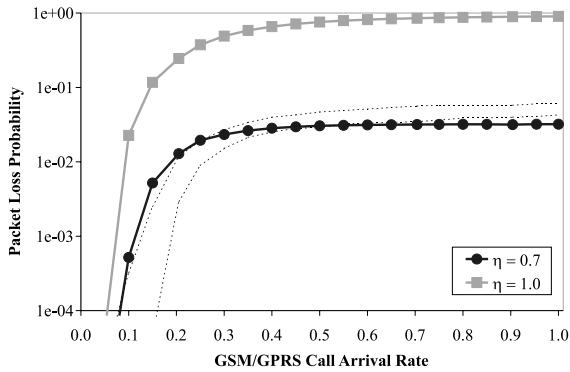
Fig. 5. Calibrating the threshold η to closely represent the flow control of TCP.

down the traffic, even if the network is not really congested. Subsequently, we consider $\eta = 0.7$ in the following experiments. A threshold of $\eta = 1.0$ corresponds to the case without flow control. In this case the PLP approaches the value 1.0 for increasing call arrival rate.

In order to validate relevant performance measures, Fig. 6 plots curves for CDT and ATU for different percentages of GPRS users in comparison to the numerical results. These curves clearly indicate that the simplifications introduced in the Markov model do not alter significantly the performance measures of interest. Thus, the Markov model is highly accurate and can effectively be utilized for studying the performance of GPRS.

The shape of the CDT curve of Fig. 6 can be explained as follows: For low traffic the fraction of the channel utilization corresponding to GPRS users increases up to 4.8 in case of 10% GPRS

users. However, with increasing traffic the fraction of the channel utilization of GPRS users decreases because more and more GSM users occupy the radio resources. This is due to the assumption that GSM users have priority over GPRS users. Therefore, for very high traffic the fraction of the channel utilization corresponding to GPRS users decreases to its minimum which corresponds to the one reserved PDCH. The reduction of CDT for increasing traffic load clearly decreases the throughput for every GPRS user as depicted in the right curve of Fig. 6.

5.3. A comparative performance study of GPRS

This section presents numerous performance curves of the cellular mobile communication network derived from steady-state solutions of the Markov model. In particular, we investigate the impact of the number of PDCHs reserved for GPRS users on the performance of the cellular network. This results give valuable hints for network designers on how many PDCHs should be allocated for GPRS for a given amount of traffic in order to guarantee appropriate QoS. In the curves presented in this section, we assume the base parameter setting of Table 2 if not mentioned otherwise. In all curves the arrival rate of GSM and GPRS users is varied to study the cell under increasing traffic intensity due to more user requests.

Figs. 7–9 present a comparative study of the mobile network considering traffic models 1 and 2. As performance measures, we consider CDT, PLP, and QD as defined in Section 4. In each figure we
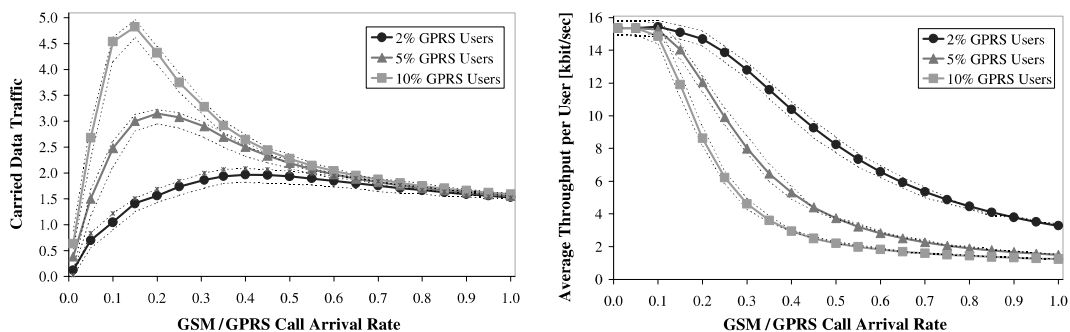

Fig. 6. Validation of numerical results with detailed simulator, 1 reserved PDCH.
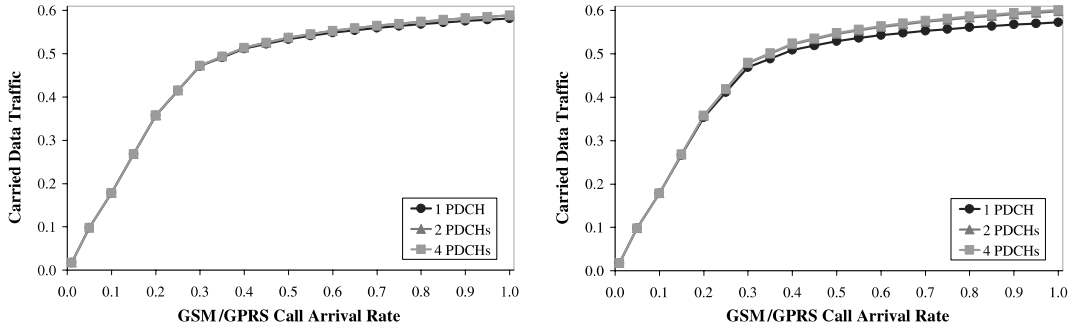
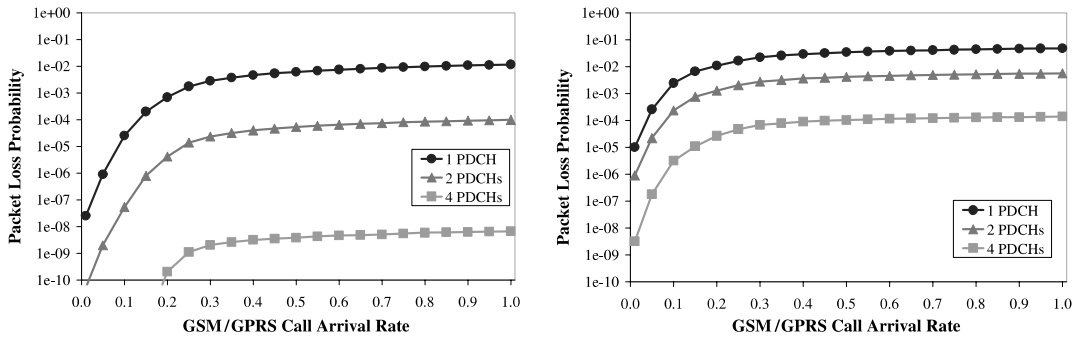Fig. 7. CDT for traffic model 1 (left) and 2 (right).

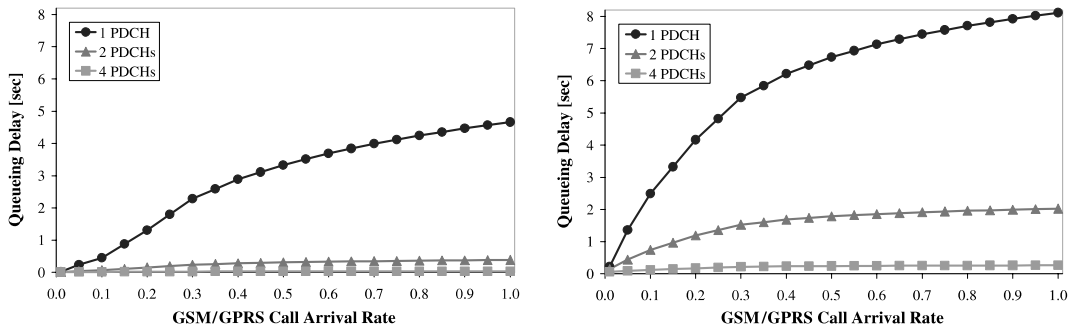Fig. 8. PLP for traffic model 1 (left) and 2 (right).

Fig. 9. QD for traffic model 1 (left) and 2 (right).

vary the number of reserved PDCHs (1, 2, and 4). The maximum number of GPRS sessions that can be concurrently active in the cell is restricted to $M = 50$. From the curves presented in Fig. 6 we see that for both traffic models the CDT remains nearly the same even if we reserve 1, 2 or 4 PDCHs for GPRS. For a GSM/GPRS call arrival rate of 1

call per second only 0.6 PDCHs are used on average. Note, that a GSM/GPRS call arrival rate of 1 call per second corresponds to 0.05 new GPRS session requests per second in case of 5% GPRS users. To derive the overall GPRS session request rate, we have to add the handover request rate that is obtained by the balancing procedure
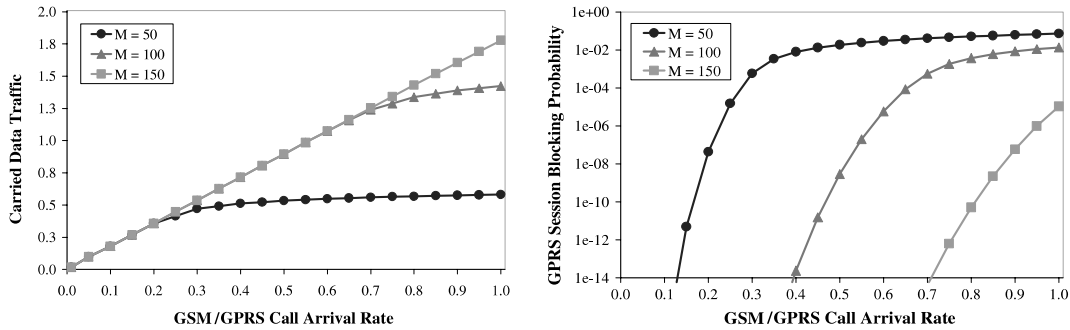
Fig. 10. CDT and GPRS session blocking probability.

(see Sections 3 and 4). In case of traffic model 1 and 2 this handover rate is very high for GPRS users because their dwell time in the cell is very low compared to their average session duration. Therefore, we conclude that for both traffic models one PDCH is sufficient to carry the data traffic of the GPRS users.

From Figs. 8 and 9, we observe that reserving more PDCHs decreases QD and the probability of packet loss due to buffer overflow. This is surely important to provide certain QoS guaranties to GPRS users. Consequently, we conclude that on the one hand reserving more PDCHs will decrease QD and packet loss (in bursty packet arrival phases) but on the other hand these extra physical channels will be idle most of the time. It is noteworthy that due to the scarce radio resources the reservation of PDCHs has to be decided carefully. Reserving two or even more PDCHs would only be desirable for providing certain QoS guaranties to GPRS users. Comparing the curves in Figs. 8 and 9, we find that traffic model 2 which produces more bursty traffic (arrival rate of 32 Kbit/s during a packet call) results in longer delay and higher PLP.

Due to the long GPRS session duration of approximately 2100 s (equals to 35 min) in traffic models 1 and 2, the handover arrival rates of active GPRS sessions is very high (about 0.3 GPRS handover requests per second at an GSM/GPRS arrival rate of 1 call per second). Therefore, the blocking probability of arriving GPRS sessions is also high because the maximal number of active GPRS sessions in the cell is restricted to $M = 50$

and this limit is reached very quickly (about 10% of GPRS users are not admitted in the cell at an arrival rate of 1 GSM/GPRS call per second). This effect justifies the following experiment where we studied how many PDCHs are needed to satisfy *almost all* GPRS session requests up to a GSM/GPRS call arrival rate of 1 call per second (see Fig. 10). Therefore, we increase the maximum number of active GPRS sessions allowed in the cell to values $M = 50$, 100, and 150, respectively. The curves of Fig. 10 plot CDT and GPRS session blocking probability versus GSM/GPRS call arrival rate. They are computed using traffic model 1. For $M = 150$ we find a maximal GPRS session blocking probability that is below 10^{-5} with an utilization of 1.8 PDCHs on average: In fact no more PDCHs are needed! We conclude from Fig. 10 that the reservation of 2 PDCHs for GPRS is sufficient to satisfy *almost all* GPRS session requests up to a new call arrival rate of 1 call per second.

In the next experiments, we investigate the system under higher GPRS traffic load, i.e. traffic model 3. Figs. 11–13 present a comparison of the mobile network for different system configurations. The comparison is made in two dimensions: the amount of GPRS users and the number of reserved PDCHs. In each curve, we vary the number of reserved PDCHs (0, 1, 2, and 4) and the fraction of GPRS users among newly arriving calls (2%, 5%, and 10%). As performance measure, we consider the CDT and ATU.

For low traffic the utilization of physical channels for packet transfer is independent from the
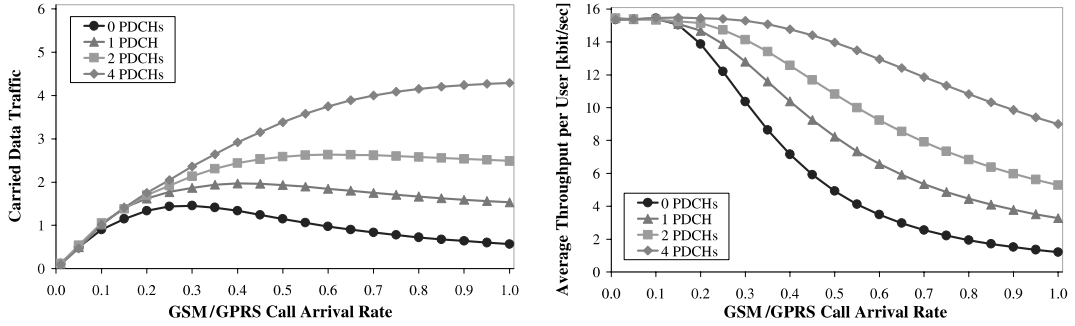
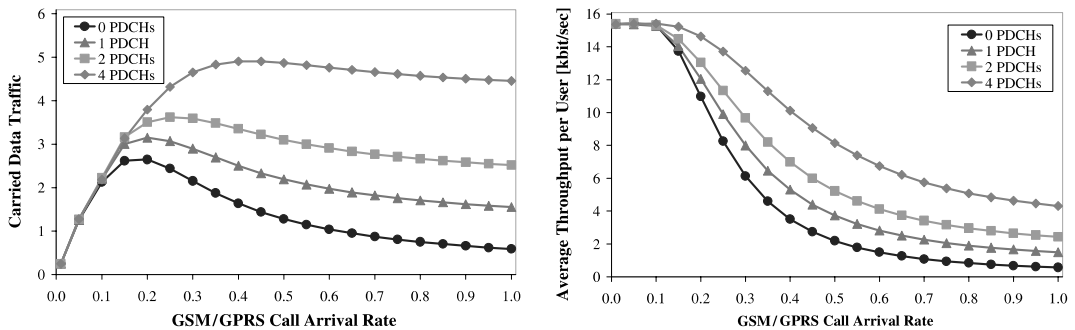Fig. 11. CDT and throughput per user for 2% GPRS users.

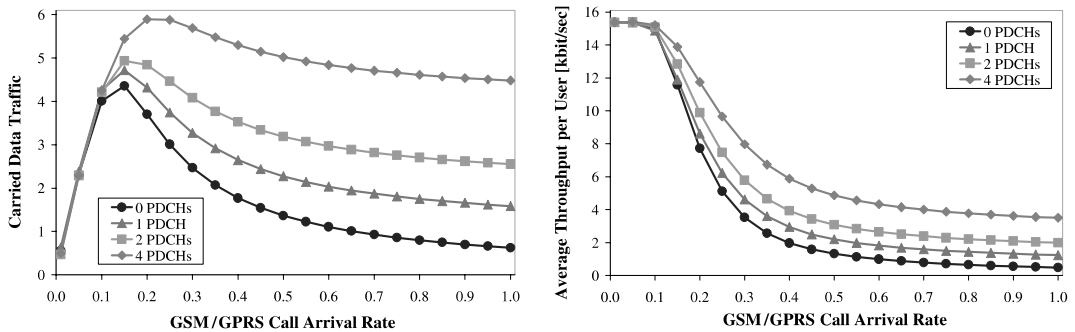Fig. 12. CDT and throughput per user for 5% GPRS users.

Fig. 13. CDT and throughput per user for 10% GPRS users.

numbers of reserved PDCHs. This is because the low amount of traffic can be completely managed by the cell even with no reserved PDCH. However, for increasing traffic intensity the channel utilization for data transfer decreases. This can be ex-

plained by the same argument as for Fig. 6. Furthermore, we observe that the decrease of allocated PDCHs due to high traffic intensity becomes less significant when more PDCHs are reserved. This observation can also be concluded
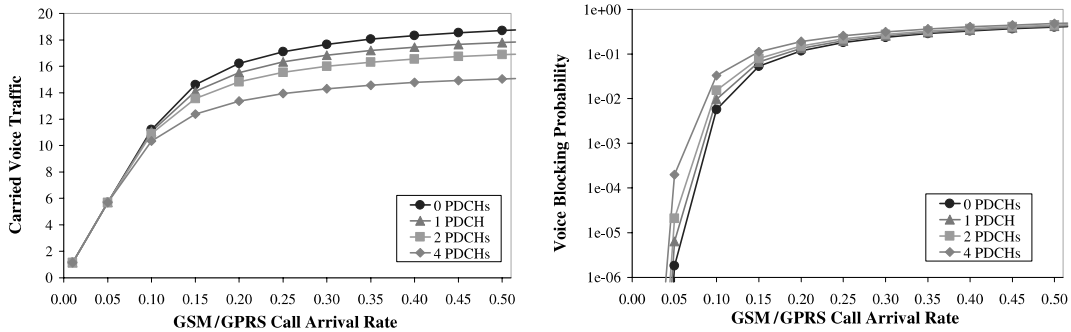
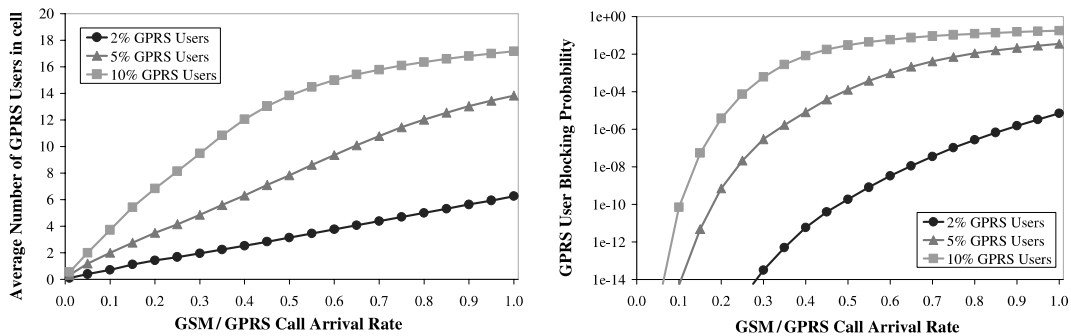Fig. 14. Influence of GPRS on GSM voice service (95% GSM calls).

Fig. 15. Average number of GPRS users in cell and GPRS user blocking probability.

from the curves plotting ATU in Figs. 11–13. Surely, under very low traffic intensity for GSM and GPRS users, each GPRS user reaches the maximum throughput. With increasing load, the throughput per user decreases much more slightly in case of four reserved PDCHs. This is opposed to the case of no reserved PDCHs where the throughput approaches nearly zero.

Comparing the different GPRS user populations, we discuss an example of high interest for network designers: Consider GPRS users with a QoS profile that allows a throughput degradation of at most 50%. Then, we can conclude that for 2% GPRS users the reservation of 4 PDCHs is sufficient up to an GSM/GPRS call arrival rate of 1 call per second. However, for the case of 5% and 10% GPRS users, the QoS profile can only be guaranteed up to a call arrival rate of 0.5 and 0.3 calls per second, respectively. In this case network designers should think about more re-

strictive call admission conditions to meet the requirements.

In an additional experiment, we study the performance loss in the GSM voice service due to the introduction of GPRS. Fig. 14 plots the CVT and voice blocking probability for different numbers of reserved PDCHs. The presented curves indicate that the decrease in channel capacity and, thus, the increase of the blocking probability of the GSM voice service is negligible compared to the benefit of reserving additional PDCHs for GPRS.

Fig. 15 presents curves for average number of GPRS users in the cell and blocking probabilities of GPRS session requests due to reaching the limit of M active GPRS sessions. We observe that for 2% GPRS users the maximum number of 20 active GPRS sessions is not reached. Therefore, the blocking probability remains below 10^{-5}. For 10% GPRS users and increasing call arrival rate, the average number of sessions approaches its

maximum. Thus, some GPRS users will be rejected. It is important to note that the curves of Fig. 15 can be utilized for determining the average number of GPRS users in the cell for a given call arrival rate. In fact, together with the curves of Figs. 11–13, we can provide estimates for the maximum number of GPRS users that can be managed by the cell without degradation of QoS. For example, for 5% GPRS users and 4 PDCHs reserved, the QoS profile of maximal 50% throughput degradation is achieved until the call arrival rate exceeds 0.5 calls per second, i.e., until there are on average eight active GPRS users in the cell.

6. Conclusions

This paper presented a comprehensive performance study of the radio resource sharing by circuit switched GSM connections and packet switched GPRS sessions under a dynamic channel allocation scheme. We assumed a fixed number of physical channels permanently allocated to GPRS and the remaining channels to be on-demand channels that can be used by GSM voice service and GPRS packets. Performance results are derived from the steady-state analysis of a Markov model. A validation of the Markov model with a detailed simulator on the network level shows that almost all performance curves derived from the Markov model lie in the confidence intervals of the corresponding curve of the simulator.

We investigated the impact of the number of PDCH reserved for GPRS users on the performance of the cellular network. That is for example, for GPRS users with a QoS profile allowing a throughput degradation of at most 50%, we concluded that for 2% GPRS users among all incoming calls, the reservation of four PDCHs is sufficient up to an GSM/GPRS call arrival rate of 1 call per second. However, for the case of 5% and 10% GPRS users, the QoS profile can only be guaranteed up to a call arrival rate of 0.5 and 0.3 calls per second, respectively. Such results give valuable hints for network designers on how many PDCHs should be allocated for GPRS for a given

amount of traffic in order to guarantee appropriate QoS.

Note, that determining the number of PDCHs for GPRS is a tradeoff between GSM and GPRS performance. Therefore, an optimal value of PDCHs can be only determined with respect to the desired performance requirements for GSM and GPRS that must be selected by the network operator. Applying adaptive performance management [14], future work considers the dynamic adjustment of the number of PDCHs with respect to the current GSM and GPRS traffic load and the desired performance requirements.

References

[1] M.A. Marsan, S. Marano, C. Mastroianni, M. Meo, Performance analysis of cellular communication networks supporting multimedia services, in: Proceedings of the Sixth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Montreal, Canada, 1998, pp. 274–281.

[2] M.A. Marsan, G. De Carolis, E. Leonardi, R. Lo Cigno, M. Meo, How many cells should be considered to accurately predict the performance of cellular networks? Proceedings of European Wireless, Munich, Germany, 1999.

[3] M.A. Marsan, M. Meo, Approximate analytical models for dual-band GSM networks design and planning, in: Proceedings of the 19th Conference on Computer Communications (Infocom), Tel-Aviv, Israel, 2000.

[4] E. Anderlind, J. Zander, A traffic model for non real-time data users in a wireless radio network, IEEE Commun. Lett. 1 (1997) 37–39.

[5] R.J. Boucherie, R. Litjens, Radio Resource Sharing in a GSM/GPRS Network, in: Proceedings of the 12th ITC Specialist Seminar on Mobile Systems and Mobility, Lillehammer, Norway, 2000, pp. 261–274.

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

[7] J. Cai, D.J. Goodman, General packet radio service in GSM, IEEE Commun. Mag. 35 (1997) 122–131.

[8] CSIM18—The Simulation Engine. http://www.mesquite. com.

[9] M. Ermel, K. Begain, T. Müller, J. Schüler, M. Schweigel, Analytical comparison of different GPRS introduction strategies, in: Proceedings of the Third ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Boston, MA, 2000, pp. 3–10.

[10] ETSI, Digital cellular telecommunications system (Phase 2+); General Packet Radio Service (GPRS); Service description; Stage 2, GSM recommendation 03.60, 1999.

[11] ETSI, Universal Mobile Telecommunication System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS, Technical Report TR 101 112 v3.2.0, 1998.

[12] W. Fischer, K. Meier-Hellstern, The Markov-modulated poisson (MMPP) process cookbook, Perform. Evaluation 18 (1993) 149–171.

[13] R. Kalden, I. Meirick, M. Meyer, Wireless internet access based on GPRS, IEEE Personal Commun. 7 (2000) 8–18.

[14] C. Lindemann, M. Lohmann, A. Thümmler, Adaptive performance management for UMTS networks, Comp. Networks 38 (2002) 477–496.

[15] Sz. Malomsoky, Sz. Nádas, G. Tóth, P. Zarándy, Simulation based GPRS network dimensioning, in: Proceedings of the 12th ITC Specialist Seminar on Mobile Systems and Mobility, Lillehammer, Norway, 2000.

[16] J.G. Markoulidakis, G.L. Lyberopoulos, M.E. Anagnostou, Traffic model for third generation cellular mobile telecommunication systems, Wireless Networks 4 (1998) 389–400.

[17] M. Meyer, TCP performance over GPRS, in: Proceedings of the First Wireless Communications and Networking Conference (IEEE WCNC), New Orleans, MS, 1999, pp. 1248–1252.

[18] P. Stuckmann, F. Müller, GPRS radio network capacity and quality of service using fixed and on-demand channel allocation techniques, in: Proceedings of the 51th Vehicular Technology Conference, Tokyo, Japan, 2000.

Christoph Lindemann is an Associate Professor in the Department of Computer Science at the University of Dortmund and leading the Computer Systems and Performance Evaluation group. From 1994 to 1997, he was a Senior Research Scientist at the GMD Institute for Computer Architecture and Software Technology (GMD FIRST) in Berlin. In the summer 1993 and during the academic year 1994/1995, he was a Visiting Scientist at the IBM Almaden Research Center, San Jose, CA. Christoph Lindemann is a Senior Member of the IEEE. He is author of the monograph Performance Modelling with Deterministic and Stochastic Petri Nets published by John Wiley in 1998. Moreover, he co-authored the survey text Performance Evaluation—Origins and Directions, Springer-Verlag, 2000. He served on the program committees of various well-known international conferences. His current research interests include mobile computing, communication networks, Internet search technology, and performance evaluation.

Axel Thümmler received the degree Diplom-Informatiker (M.S. in Computer Science) from the University of Dortmund in April 1998. Presently, he is a Ph.D. student in the Computer Systems and Performance Evaluation group at the University of Dortmund. His research interests include mobile computing, communication networks, and performance evaluation.