Burkhard Monien
Viktor K. Prasanna
Sriram Vajapeyam (Eds.)

LNCS 2228

High Performance Computing – HiPC 2001

8th International Conference
Hyderabad, India, December 2001
Proceedings

Springer

Performance Analysis of Data Services over GPRS[★]

Marco Ajmone Marsan¹, Marco Gribaudo², Michela Meo¹, and Matteo Sereno²

¹ Dipartimento di Elettronica, Politecnico di Torino, 10129 Torino, Italy
`{ajmone,michela}@polito.it`

² Dipartimento di Informatica, Università di Torino, 10149 Torino, Italy
`{marcog,matteo}@di.unito.it`

Abstract. In this paper we describe a model for the performance analysis of mobile packet data services sharing a common GSM/GPRS (Global System for Mobile communications/General Packet Radio Service) cellular infrastructure with mobile telephony services. The performance model is developed using the DSPN (Deterministic and Stochastic Petri Net) paradigm. With limited computational cost, the DSPN solution allows the derivation of a number of interesting performance metrics, which can be instrumental for the development of accurate design and planning algorithms for cellular networks offering integrated services.

1 Introduction

Seven hundred millions mobile telephony service users are being offered (or will soon be offered) mobile packet data services for a variety of different applications, the most relevant of which is the mobile, ubiquitous access to the Internet.

In GSM (Global System for Mobile communications) systems [1,2], the technology that has been standardized for the packet transfer of data from/to mobile terminals is GPRS (General Packet Radio Service) [3,4,5,6,7], which is considered the forerunner of the Universal Mobile Telecommunication System (UMTS) [8,9], for whose licenses mobile telephony companies in Europe have made huge investments.

The fact that GPRS exploits the same resources used by mobile telephony raises a number of questions concerning the dimensioning and the management of the GSM/GPRS radio interface. In this paper we tackle this issue, and we develop a Petri net based model of the behavior of the radio interface in a GSM/GPRS cell, providing numerical performance results that give insight into the possible design tradeoffs.

The particular class of Petri nets used for the model construction is called DSPN (Deterministic and Stochastic Petri Net) [10,11]. These are Petri nets with timed transitions, where the transition firing delays can be either deterministic or exponentially distributed. While no restriction exists in the use of

[★] This work was supported in part by the Italian National Research Council, and by CERCOM.

exponentially distributed transition firing delays, constant firing delays must be used for transitions that are enabled one at a time. In our case this is not a problem, since just one constant delay will be used to model the elementary time slot in our model.

The paper is organized as follows. In Section 2 we describe the characteristics of the GSM/GPRS cellular mobile communication network that we consider, together with the probabilistic assumptions that are needed to describe the system dynamics with a DSPN model. In Section 3 we describe the characteristics of the DSPN model and its solution. Results are shown and discussed in Section 4. Section 5 briefly comments on the complexity of the DSPN model solution. Section 6 concludes the paper.

2 System and Modeling Assumptions

GSM adopts a *cellular architecture*: the area served by a GSM network is covered by a number of non-overlapping *cells*. Each cell is controlled by a *base station* (called BTS – Base Transmitter Station, in the GSM jargon), which handles the transmissions from and to mobile terminals over the wireless link. Base stations are then connected through a (wired or wireless) point-to-point link to the fixed part of a GSM network.

A small number of frequency bands (resulting from a partition of the portion of spectrum allocated by international organizations to GSM services, around 900 MHz and 1.8 GHz), each one of 200 kHz, is activated in any cell (typical numbers are a few units). Each frequency band defines a FDMA *channel*, on which a TDMA *frame* of 60/13 ms is defined, comprising 8 *slots* of 15/26 ms. A circuit (or channel) is defined by a slot position in the frame, and by a frequency band; thus, the maximum number of telephone calls that can be simultaneously activated in a cell corresponds to 8 times the number of frequency bands. At least one channel must be allocated for signalling in each cell, so that a frequency band can use for the transmission of end user information (either voice or data) from 6 to 8 channels, depending on the cell configuration; we will assume that in cells with one active frequency band, the TDMA frame allocates 7 slots to end users and 1 slot to signalling. The pair (frequency,slot) defines a *control channel* if the channel is used for signalling, and defines a *traffic channel* if the channel is used for the transmission of end user information (either voice or data).

When an end user requests the activation of a new GSM telephone call, the system must transfer some signalling information to inform all the involved entities about the request and to distribute the necessary commands, and allocate a traffic channel on one of the frequency bands of the cell. If no slot is free among those that can be allocated to end-users, the new call cannot be established, and is *blocked*.

Similarly, when users move from one cell to another (we call the former origin cell, and the latter destination cell) during a telephone call, the system must transfer the call from the origin cell to the destination cell through a *handover* procedure. If no traffic channel can be allocated in the destination cell, the call

handover fails (a *handover blocking* occurs) and the telephone call is forced to terminate.

As customary in models of telephone systems, we shall assume that the sequence of new call requests follows a Poisson process, and that the duration of calls is an exponentially distributed random variable. Moreover, as is normally done when modeling cellular telephony systems, we consider one cell at a time [12], we neglect the impact of signalling, and we include incoming handover requests into the Poisson arrival process, whose rate will be set to λ_v (thus, λ_v is the total rate of requests incoming to a cell, and each request should be satisfied with one of the slots on the frequency bands that are active within the cell, i.e., with a traffic channel). The total active time of a call within a cell (which is normally called the call *dwelt time*) differs from the call duration, because a call may traverse several cells during its active time; we extend the exponential assumption to dwell times also, defining a rate μ_v for the time from the beginning of the life of a call in a cell (resulting from either the successful activation of a new call or a successful incoming handover) until the end of the call activity in the same cell (due to either the termination of the call by one of the end users or an outgoing handover). Note that exponential assumptions are generally considered not to be critical in telephony models: telephone systems have been dimensioned using exponential assumptions for almost a century. In the derivation of numerical results we shall assume $1/\mu_v = 120$ s. This value is quite often used in design and planning of mobile telephony systems; it is somewhat shorter than the traditional average call holding time of fixed telephony (180 s), in order to account for handovers.

GPRS was conceived for the transfer of packets over a GSM infrastructure, with a simplified allocation of resources over the wireless link, and an IP transport among additional elements of the wired GSM network, called SGSNs (Serving GPRS Support Nodes). In order to cross the wireless link, IP packets are fragmented in *radio blocks*, that are transmitted in 4 slots in identical positions within consecutive GSM frames of the same frequency band. Depending on the length of the IP packet, the number of radio blocks necessary for the transfer may vary. The allocation of the radio link to radio block transmissions can either use dedicated resources for signalling, or (more usually) the same signalling resources that are available for telephony.

In order to describe the GPRS traffic, we adopted the model of Internet traffic defined by the 3GPP (3rd Generation Partnership Project) in [13]; a sketch of the GPRS traffic model that we use is shown in Figure 1. Active users within a cell execute a *packet session*, which is an alternating sequence of *packet calls* and *reading times*. According to [13], the number of packet calls within a packet session should be a geometrically distributed random variable with average N_{pc} ; however, since we will study the system behavior for a fixed number of concurrently active packet sessions, we will assume that packet sessions remain active for an indefinite amount of time. The reading time between packet calls is an exponentially distributed random variable with rate μ_{Dpc} . Each packet call comprises a geometrically distributed number of packets with average Nd ; the

Fig. 1. Model of GPRS traffic: a packet session.

interarrival time between packets in a packet call is an exponentially distributed random variable with rate μ_{Dd} . According to [13], we shall assume $1/\mu_{Dpc} = 41.2$, $Nd = 25$ and $1/\mu_{Dd} = 0.5$, all times being expressed in seconds. The packet size in radio blocks can have a number of different distributions, some with heavy tail. In our model the packet size is equal to either 1 or 16 radio blocks, with the same probability. Other discrete distributions could be easily introduced in our models in order to approximate heavy tailed distributions of the packet size.

The transfer of radio blocks over the radio channel can either be successful, thus allowing the removal of the radio block from the buffer, or result in a failure due to noise, fading, or shadowing. These probabilities are modeled with a random choice: with probability c a radio block transfer is successful and with probability $1 - c$ it fails.

3 The DSPN Model

In this section we describe the DSPN model of one cell of a GSM system where just one frequency band is available to simultaneously offer telephony and GPRS services. We focus on transmissions from the base station to end user terminals (this is the so-called *downlink* direction).

The DSPN model is shown in Fig. 2. The only deterministic transition in the model is T_{4F} , whose delay corresponds to four consecutive GSM frames, i.e., $4 \cdot 60/13$ ms. The beginning and end of each quadruple of frames is modeled by the firing of immediate transitions t_b and t_e , respectively. Thus, the marking of place BEG-4F indicates that a quadruple of frames is starting, the marking of place END-4F indicates that a quadruple of frames is ending, and the marking of place 4F indicates that a quadruple of frames is in progress. One token is in place 4F in the initial marking.

We assume that a traffic channel can be allocated to new telephone call requests and incoming handover requests only at boundaries of quadruples of frames. Similarly, the scheduling of transmissions of radio blocks on traffic channels not used by voice can take place only at boundaries of quadruples of frames. Both assumptions appear to be quite realistic, at least for the first implementations of GPRS services.

At each firing of transition T_{4F} , one radio block can be transferred on each traffic channel that is not used by voice connections.

Fig. 2. The DSPN model.

The number of tokens in place *CIRCUITS* in the initial marking of the DSPN equals the number of traffic channels available in the cell (i.e., 7, according to our assumptions). Voice calls that await the allocation of a traffic channel are modeled by tokens in place *REQ-V*. Voice calls that are in progress are modeled by tokens in place *ACTIVE-V*. The total number of voice calls (that either await the allocation of a traffic channel, or are using a traffic channel) in the cell is modeled by tokens in place *TOT-V*.

Arrivals of new voice call requests and incoming voice call handover requests into the cell are modeled by the timed transition T_{in-v} , which must thus have rate λ_v . The arrival of a voice call increases the marking of places *REQ-V* and *TOT-V*; then, since no more than 7 calls can be simultaneously established in a cell, if the marking of place *TOT-V* exceeds 7, the immediate transition t_{loss-v} fires, modeling a voice call blocking (either a new call blocking or a handover blocking), and one token is removed from both *REQ-V* and *TOT-V*. At the beginning of a quadruple of GSM frames, voice calls that await the allocation of a traffic channel become active by firing the immediate transition $t_{allocate-v}$. This removes one token from *REQ-V*, and increases by one the marking of *ACTIVE-V*, but does not alter the marking of *TOT-V*.

The completion of voice calls during a quadruple of frames is modeled by the firing of the timed transition T_{end-v} , which must thus have rate μ_v and infinite server semantics. The firing of T_{end-v} decreases the marking of both *ACTIVE-V* and *TOT-V*.

The number N of ongoing packet sessions in the cell is assumed to be constant, and is modeled by the number of tokens in place *OFF-D* in the initial marking. The beginning and end of packet calls within a packet session is modeled by the firing of the timed transitions T_{new-d} and T_{end-d} , respectively, which must thus have infinite server semantics with rates μ_{Dpc} and μ_{Dd}/Nd . Tokens in place *ON-D* model active packet calls. The generation of a packet within an active packet call is modeled by the firing of the timed transition T_{pkt} , and the generation of one token in *PKT*; the firing rate is μ_{Dd} with infinite server semantics. The firing of one of the two immediate transitions t_{1rb} , t_{16rb} , models the generation of either 1 or 16 radio blocks from the packet segmentation, with

equal probability. Note that these two immediate transitions describe the pdf of the number of radio blocks generated by segmenting a packet. We chose to describe a sort of bimodal packet length distribution with just two transitions, but the inclusion in the model of a larger number of transitions does not increase the model complexity (the model state space remains the same), so that more complex or realistic distributions can be very simply incorporated in the model. Tokens in place RB model radio blocks requesting access to the buffer where they can wait their turn for transmission. Radio blocks enter the buffer (modeled by place BUFFER) if free positions are available (some token marks place FREE-BUF), through the immediate transition t_{in-buf} . If the buffer is full (place FREE-BUF is empty), radio blocks are lost through transition t_{loss-d} . The initial marking of place FREE-BUF corresponds to the buffer capacity in radio blocks, which is denoted by B .

The allocation of traffic channels not used by voice calls to radio blocks is modeled by the immediate transitions $t_{allocate-d}$, which is enabled at the beginning of a quadruple of GSM frames, with lower priority than $t_{allocate-v}$. Transmitted radio blocks are modeled by tokens in place ACTIVE-D. If the radio block transmission is not successful, a copy of the radio block must remain in the buffer for retransmission. The immediate transitions t_{rb-OK} and t_{rb-KO} , with probabilities c and $1 - c$, model the successful and unsuccessful transmission of a radio block.

Table 1 reports weights and priorities for all immediate transitions in the DSPN. Table 2 reports the firing rates for all exponential transitions; Table 3 reports the initial marking for all places.

Table 1. Weights and priorities of immediate transitions in the DSPN model.

Transition	Weight	Priority
t_{loss-v}	1	4
t_{loss-d}	1	4
t_{in-buf}	1	4
$t_{allocate-v}$	1	3
$t_{allocate-d}$	1	2
t_{1rb}	$1/2$	4
t_{16rb}	$1/2$	4
t_b	1	1
t_e	1	1
t_{rb-OK}	c	2
t_{rb-KO}	$1 - c$	2

Table 2. Firing rates of exponential transitions in the DSPN model.

Transition	Firing rate	Semantics
T_{in-v}	λ_v	single server
T_{end-v}	μ_v	infinite server
T_{new-d}	μ_{Dpc}	infinite server
T_{end-d}	μ_{Dd}/Nd	infinite server
T_{pck}	μ_{Dd}	infinite server

Table 3. Initial marking of the DSPN model.

Place	Initial marking
CIRCUITS	7
4F	1
FREE-BUF	B
OFF-D	N

3.1 Performance Indices

The DSPN model allows the computation of many interesting performance indices. In this paper we shall mainly consider the following performance measures:

- voice call blocking probability (for both new call and handover requests); this is the ratio between the throughput of transition t_{loss-v} and the throughput of T_{in-v}
- average number of radio blocks in the base station buffer; this is the average number of tokens in place BUFFER
- average number of radio blocks in the base station buffer conditional on the number of active voice calls being equal to k ; this is the average number of tokens in place BUFFER when the number of tokens in place ACTIVE-V equals k
- radio block loss probability; this is the ratio between the throughput of transition t_{loss-d} and the sum of the throughput of transitions t_{loss-d} and t_{in-buf} .

4 Numerical Results

We consider a cell with just one active frequency band, which allows the definition of one control channel and seven traffic channels. We assume that the voice and data traffic in this cell are as specified in Section 2. The traffic parameters values are summarized in Table 4, and are used for the derivation of all numerical results, unless otherwise specified.

The left plot of Figure 3 shows results for the voice call blocking probability and the radio block loss probability versus increasing values of the voice call

Table 4. Parameters of the basic scenario.

Parameter	Value
Number of traffic channels	7
Buffer size, B	100
Number of packet sessions, N	1, 3, 4, 5
$1/\mu_v$	120 s
Nd	25
$1/\mu_{Dd}$	0.5 s
$1/\mu_{Dpc}$	41.2 s
c	0.95

Fig. 3. Voice call blocking probability and radio block loss probability (left plot) and average queueing delay of radio blocks (right), with 1, 3 or 5 packet sessions.

request rate, λ_v . The radio block loss probability is plotted for different values of the number of active data sessions, N . The voice call blocking probability is insensitive to data traffic; instead, the radio block loss probability slightly increases with the number of data sources. This effect is mostly due to the increase in data traffic resulting from the higher number of data sources.

For the same scenario, the average queueing delay experienced by radio blocks is presented in the right plot of the same Figure. It can be observed that the delay suffered by radio blocks is quite large, due to the low bit rate of the radio channel and to the priority of voice traffic in accessing the radio resources.

The average number of radio blocks in the base station buffer is shown in the left plot of Figure 4, again as a function of the voice call request rate in the case of 1, 3 or 5 packet sessions. It is interesting to observe that the average number of radio blocks in the base station buffer remains quite small, even for rather large values of the radio block loss probability. The responsibility for the buffer overflow resides in the fact that voice has priority over data and in the different time scales of data and voice traffic. Indeed, the radio block service rate depends on the voice calls dynamics and, in particular, when all channels in the cell are busy with voice connections, no radio block can be transmitted until at least one traffic channel becomes free. During these intervals of service

Fig. 4. Average buffer occupancy with 1, 3 or 5 packet sessions (left), and conditional average buffer occupancy, given that 5, 6 or 7 voice calls are active, with 3 packet sessions (right).

Fig. 5. Radio block loss probability (left) and average buffer occupancy (right) for different values of buffer size, with 3 packet sessions.

interruption, which are quite long compared to the data traffic dynamics, the radio block buffer fills up, and a large number of radio blocks is lost.

This is confirmed by the right plot of Figure 4 which shows the average number of radio blocks in the base station buffer conditional on the number of active voice calls, for 5, 6, or 7 active voice calls, and 3 data sessions. The buffer fills when all traffic channels are busy with voice calls; this is the main source of the base station buffer overflow, hence of the radio block loss. Moreover, the fact that the voice call dynamics are much slower than the radio block generation dynamics makes the buffer overflow periods quite long in relative terms. As a result, although the average number of radio blocks in the base station buffer is quite low, the radio block loss probability is rather high.

A further proof of this phenomenon comes from the left plot of Figure 5, which reports the radio block loss probability for different values of the buffer size. Observe that the buffer size has a minor impact on loss probability. This is again due to the different time scales of voice call and radio block generation dynamics. The time interval during which radio blocks are lost is much larger than the time needed to fill the buffer so that radio block loss probability results to be almost independent from the buffer capacity. We notice a significant increase of the radio block loss probability only when the buffer size reduces to 25 radio blocks.

Fig. 6. Voice call blocking probability and radio block loss probability (left) and average queueing delay of radio blocks (right) when either 0 or 1 traffic channels are reserved to GPRS traffic, with 4 packet sessions.

This is due to the data traffic burstiness: the probability that two packets of 16 radio blocks each are generated close to each other is quite large. The right plot of Figure 5 reports the average radio block buffer occupancy for different values of buffer capacity. Observe again that the average buffer occupancy is always quite small.

Finally, the left plot of Figure 6 compares the radio block loss probability and the voice call blocking probability estimates when either zero or one traffic channels are reserved to GPRS traffic, assuming that 4 packet sessions are in progress. Note that the presence of one reserved channel produces a drastic reduction of the radio block loss probability (over one order of magnitude), with a modest increase of the voice call blocking probability (about a factor 3). A similar gain can be observed for the average radio block queueing delay in the right plot of the same Figure.

5 Complexity of the Model Solutions

The CPU times required for the solution of the DSPN models are reported in Table 5 together with the state space size of the underlying stochastic model. All models were solved on a Pentium III running at 600 Mhz with 512 Mbytes of main memory. The models were solved by using the software package DSPNexpress [11].

6 Conclusions

In this paper we have described a DSPN model for the performance analysis of mobile packet data services based on GPRS, and exploiting a common cellular infrastructure with GSM mobile telephony services.

The DSPN model is quite flexible and computationally parsimonious, allowing the accurate estimation of a variety of interesting performance metrics for the considered system.

Table 5. CPU times for the DSPN solution.

N	State space size	DSPN sol.[s]
1	58176	172
2	87264	522
3	116352	1136
4	145440	2046
5	174528	4281

Simple and accurate models of GSM/GPRS systems, like the one that we presented here, are a must for the effective design and planning of multiservice cellular systems.

References

1. M. Mouly, M.-B. Pautet, "The GSM System for Mobile Communications", Cell&Sys, 1992.
2. B.H. Walke, "Mobile Radio Networks: Networking and Protocols", John Wiley, 1999.
3. "Digital Cellular Telecommunications System (Phase 2+), General Packet Radio Service, Service Description, Stage 1", GSM 02.60 v.6.2.1, August 1999.
4. "Digital Cellular Telecommunications System (Phase 2+), General Packet Radio Service, Service Description, Stage 2", GSM 03.60 v.6.3.0, April 1999.
5. G. Brasche, B.H. Walke, "Concepts, Services, and Protocols of the New GSM Phase 2+ General Packet Radio Service", *IEEE Communications Magazine*, Vol. 35, N. 8, pp. 94-104, August 1977.
6. J. Cai, D. Goodman, "General Packet Radio Service in GSM", *IEEE Communications Magazine*, Vol. 35 N. 10, pp. 122-131, October 1977.
7. R. Kalden, I. Meirick, M. Meyer, "Wireless Internet Access Based on GPRS", *IEEE Personal Communications*, Vol. 7, N. 2, pp. 8-18, April 2000.
8. J.P. Castro, "Air Interface Techniques for Future Mobile Networks: The UMTS Radio Access Technology", John Wiley, 2001.
9. A. Samukic, "UMTS Universal Mobile Telecommunications System: Development of Standards for the Third Generation", *IEEE Transactions on Vehicular Technology*, Vol. 47, N. 4, pp. 1099-1104, November 1998.
10. M. Ajmone Marsan, G. Chiola, "On Petri Nets with Deterministic and Exponentially Distributed Firing Times," in: G.Rozenberg (editor), *Advances in Petri Nets 1987*, Lecture Notes in Computer Science n.266, Springer Verlag, 1987.
11. C. Lindemann, "Performance Modelling with Deterministic and Stochastic Petri Nets", John Wiley, 1998.
12. M. Ajmone 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?", *European Wireless'99 and 4th ITG Mobile Communications*, Munich, Germany, October 1999.
13. "Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0)", ETSI TR 101 112 V3.2.0 (1998-04).