Packet Delay Analysis in GPRS Systems

Marco Ajmone Marsan, Paola Laface, Michela Meo
Dipartimento di Elettronica, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
{ajmone,paola.laface,michela}@polito.it

***Abstract*— In this paper we describe an analytical model to compute the packet delay distribution in a cell of a wireless network operating according to the GSM/GPRS standard. GSM (Global System for Mobile communications) is the most widely deployed wireless telephony standard, and GPRS (Generalized Packet Radio Service) is the technology that is now available to integrate packet data services into GSM networks. By comparing the performance estimates produced by the analytical model against those generated by detailed simulation experiments, we show that the proposed modeling technique is quite accurate. In addition, we show that the results produced by the analytical model are extremely useful in the design and planning of a wireless voice and data network.**

***Index Terms*— GSM, GPRS, Packet delay distribution, Performance analysis, Matrix analytic techniques, Markovian models.**

I. Introduction and Motivations

After many years of incredible (and somewhat unexpected) success, wireless telecommunications system manufacturers and network operators are now facing a period of slowdown. If we consider the events in recent years, we see that the enormous success of GSM, in Europe and worldwide, motivated operators to invest huge resources for the acquisition of the 3G UMTS licenses; this created the need for even more resources to build 3G networks; however, competition has been driving down tariffs, and the general economic situation is not favoring an increased access to services. This implies that operators are short of cash, and are thus delaying the investments for the development of new networks, with an impact on the manufacturers that heavily invested in research and development of the 3G technology.

In order to survive in this condition, rather than overprovisioning their networks, as they did in the recent past, operators and manufacturers are trying to identify ways to effectively use the deployed resources, possibly even sharing them among several operators, a scenario unheard of until recently.

Achieving efficiency in the use of resources, calls for very effective design and planning approaches, in order to avoid degradations in the Quality of Service (QoS) offered to end users, which could imply the loss of customers and of the revenues they generate.

In the particular case of 2G and 3G wireless systems, GSM and UMTS in particular, an efficient use of resources implies a careful design of the network cells, and an appropriate partitioning of the resources between voice and data services. Indeed, while voice is still generating the largest (by far) share of revenues, operators have great expectations from wireless Internet access based on data services. It is thus necessary to dimension networks so as to keep the voice customers happy, while attracting data service users.

While techniques for the dimensioning of wireless telephony networks have been extensively investigated in the past, not as much work has been done for the combined planning of voice and data wireless networks, and the proposed design approaches were mostly based on metrics such as the average packet delay. However, we all know that in the case of Internet access the average packet delay is not the most important metric: much more interesting are the delay distribution and the delay quantiles, or the fraction of packets that experience a delay higher than a specified threshold.

In this paper we develop an analytical model to compute the packet delay distribution in a cell of a wireless network operating according to the GSM/GPRS standard[1], we validate our model by comparison against detailed simulation experiments, and we discuss the applicability of the results to the design and planning of a wireless voice and data network.

While our model refers to only one cell, it represents the key element for the development of a planning technique for multicell networks, possibly with hierarchical structure; indeed, all of the planning approaches in the literature that consider multicell systems, study one cell at a time, and then combine the results of the analysis of individual cells in order to obtain metrics at the network level.

The paper is organized as follows. The system under analysis is presented in Section II together with the assumptions introduced in the model development. The model is then explained in Section III. Numerical results are presented in Section IV; finally, Section V concludes the paper.

II. System and Modeling Assumptions

Within a cell of a GSM system, one or more carrier frequencies are activated, and over each carrier a TDMA *frame* of $T_f = 60/13$ ms is defined, comprising 8 *slots* of $15/26$ ms each. A circuit (or channel) is defined by a slot position in the TDMA frame, and by a carrier frequency. Since some channels must be allocated for signaling, each carrier frequency can devote to the transmission of end user information from 6

This work was supported in part by the Italian Ministry for University and Scientific Research under the project PlanetIP and by the Center for Multimedia Radio Communications (CERCOM).

[1]The technology that is now available to integrate packet data services into GSM networks is GPRS (Generalized Packet Radio Service).

to 8 channels, depending on the cell configuration; we will assume that the TDMA frame allocates 7 slots to end users and 1 slot to signaling. In the model development, we assume that each cell is equipped with a generic number N of traffic channels.

We consider two services: telephony and data transfer. Telephony provisioning relies on the usual circuit-based GSM service; data packets are instead transferred according to the GPRS standard, using the same resources deployed for telephony. Based on the provider strategic decisions, different channel allocation policies can be adopted for the simultaneous delivery of telephony and data transfer services. The typical allocation policy is called *voice priority* and results from strategic decisions that acknowledge the primary role of the telephony service (telecommunications network operators today still generate most of their revenues through voice services). Telephone calls are set up as long as at least one channel is available in the cell of interest. As a consequence, data packets can be transmitted only over the channels which are not used by voice connections.

A different channel allocation policy is necessary when the telecommunications services provider desires to guarantee at least a minimum QoS level to the data service. In this case, a fixed number R of channels can be reserved to data transfers, while all remaining channels are shared by voice and data connections, with priority to voice. The improvement in the QoS provided to data is obtained at the cost of reducing the resources available for telephony, thus a performance degradation for voice is expected. This policy will be called *R-reservation*.

Hybrid approaches may be applied when the telecommunications services provider expects that the introduction of GPRS may involve a small number of users only, so that a static channel reservation may result in an inefficient use of radio resources, but still some QoS must be provided to data services users. In this scenario, it may happen that during long time intervals no data transfers are required. It is then convenient to introduce some mechanisms that detect the presence of active GPRS users, and only in this case reserve channels to data traffic. While we proved that such dynamic reservation schemes can be quite effective [1], we do not study them in this paper, but our models can be rather easily extended to also cope with dynamic schemes.

In our performance analysis we focus on traditional performance metrics: the telephone call blocking probability (where call blocking may result from the lack of channels to allocate either a new call request or a handover request), the data packet loss probability, and the probability that a data packet perceives a delay longer than a given maximum allowable value. The latter performance metric will be express in terms of a pair of values (D_M, P_M), where D_M is the maximum allowable delay and P_M is the probability that the delay constraint D_M is not met, i.e., P_M expresses the fraction of packets which perceive a delay longer than D_M.

In this paper we focus on the interaction between voice and data services in a one-level cellular system. Extensions to hierarchical cellular structures are easily derived from the proposed model, similarly to [1], [2].

The telecommunications system we consider supports user mobility. Users can roam from a cell to a neighboring cell during active voice calls: an active user (i.e., a user that has established a voice or data call) that roams from a cell to another, must execute a handover procedure transferring the call from the channel in the old cell to a channel in the new cell without interrupting the communication. If no channel is available in the new cell entered by the user, the call is lost or *blocked*. In case a handover fails, the call must be terminated.

Since the duration of a data transfer is typically much smaller than the time spent by a user in a cell, we neglect the possibility that a user requests a handover procedure while transferring data. We instead account for handovers of voice connections.

As is normally done when modeling cellular telephony systems, we consider one cell at a time [3], and we neglect the impact of signaling. Moreover, in order to model the system, we introduce the assumptions discussed below.

As customary in models of telephone systems, we assume that the sequence of new call requests follows a Poisson process with rate λ, and that the duration of calls is an exponentially distributed random variable with mean $1/\mu$. We also assume that incoming handover requests follow a Poisson process, whose rate is equal to λ_h (λ_h is derived by balancing the incoming and outgoing handover flows, as explained below). Thus, the voice call arrival process is Poisson with rate

$$\lambda_v = \lambda + \lambda_h \, .$$

The time spent by a user within a cell (which is normally called *dwell time*) is assumed to be exponentially distributed with mean $1/\mu_h$. The call activity time within a cell (the channel holding time) is thus a random variable with negative exponential distribution with rate $\mu_v = \mu_h + \mu$.

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. More recently, these assumptions were used in modeling wireless telephony systems, [2], [4], [5], [6], [7].

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. In order to cross the wireless link, IP packets are fragmented into *radio blocks*, that are transmitted in 4 slots in identical positions within consecutive GSM frames over the same carrier frequency. 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 signaling, or (more usually) the same signaling resources that are available for telephony.

In order to describe the GPRS traffic, we adopt the model of Internet traffic defined by the 3GPP (3rd Generation Partnership Project) in [8]; a sketch of the GPRS traffic model that we use is shown in Fig. 1. Active users within a cell execute a *packet session*, which is an alternating sequence of *packet calls* and *reading times*. According to [8], the number of packet calls within a packet session can be described by

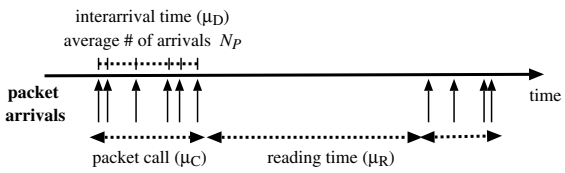

Fig. 1. Model of GPRS traffic: a packet session.

a geometrically distributed random variable; however, since we will study the system behavior for a fixed number D 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_R. Each packet call comprises a geometrically distributed number of packets with mean value N_P; the interarrival time between packets in a packet call is an exponentially distributed random variable with rate μ_D. According to [8], we shall assume $1/\mu_R = 41.2\ s$ and $N_P = 25$; we will let μ_D vary in order to change data traffic. The average packet call duration, $1/\mu_C$, is equal to the average packet interarrival time multiplied by the average number of packets generated during a packet call, so that $\mu_C = \frac{\mu_D}{N_P}$.

According to [8], the packet size in radio blocks can have a number of different distributions, some with heavy tail. In general we will denote by p_i the probability that the size of a packet is equal to i radio blocks, and we will let i vary between 1 and a maximum value, P.

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. In case of failure, the radio block transmission must be repeated. These events are modeled by a random choice: with probability c a radio block transfer is successful, and with probability $1 - c$ it fails.

III. Analytical Model

In describing the analytical model, we first focus on the voice priority channel allocation policy. We then present the extensions to be introduced in the model in order to deal with the R-reservation channel allocation policy.

A. Model of a cell

Each cell can be modeled by a queue with N servers, which represent the N available channels. Two classes of customers enter the queue. Customers in the first class represent voice connections. They arrive at the system according to a Poisson process with parameter λ_v and require a negative exponential service time with rate μ_v; these users do not queue waiting for service: if no channel is available to set up the connection, i.e., if no server is free when the customer joins the queue, the request fails, and the customer is lost.

The second class of customers represents GPRS radio blocks. From the GPRS traffic representation in Fig. 1, we observe that a GPRS user can be modeled as an *On-Off* traffic source. The time spent in state *Off* represents the reading time, while state *On* describes packet calls. In the latter state the GPRS user generates packets according to a Poisson process with rate μ_D. Depending on their size, IP packets are segmented into different numbers of radio blocks, so that the radio block arrivals at the buffer occur in batches. According to the packet size distribution, the size of a batch is equal to i radio blocks with probability p_i, and i varies between 1 and P. Radio blocks are queued waiting to be served in a transmission buffer whose capacity is equal to B radio blocks.

A radio block is transmitted over the wireless link if a channel is available (i.e., it is not used by voice connections), hence if a server is idle. The radio block is removed from the buffer if the transmission is successful, with probability c. In the analytical model we assume that the transmission time of a radio block is a random variable with negative exponential distribution with mean value equal to 4 GSM frames, $1/\gamma = 4{\cdot}T_f$. Of course, the radio block transmission time is constant and equal to $4{\cdot}T_f$, rather than exponential. However, the impact of this assumption on the system performance was shown to be very limited, due to the small value of radio block transmission times compared to voice dynamics. This phenomenon was studied in [9] by comparing the results obtained from a model which includes the exponential assumption for radio block transmission times against the results obtained from a discrete-time model with constant radio block transmission times. The exponential assumption was observed to produce accurate results.

Given the assumptions introduced above, we develop a continuous-time Markov chain (CTMC) model of the system, whose state is defined by the vector $\mathbf{s} = (b, d, v)$: where

- b is the number of radio blocks in the buffer, b varies between 0 and the buffer capacity B;
- d is the number of active packet calls, d varies between 0 and the number of data sessions, D;
- v is the number of active voice calls, v varies between 0 and the number of channels in the cell, N.

Let $\mathcal{S}$ be the state space. The number of states in $\mathcal{S}$ is equal to $(B+1)(D+1)(N+1)$.

According to the ordering of the variables presented above, we can acknowledge a block banded structure in the infinitesimal generator matrix $\mathbf{Q}$ of the CTMC, as can be seen in Fig. 2.

The blocks $\mathbf{B}_{i,j}$ are $(D+1)(N+1) \times (D+1)(N+1)$ and correspond to the transitions which make the radio block buffer occupancy change from i to j. Since an IP packet can be segmented into at most P radio blocks, all blocks $\mathbf{B}_{i,j}$ with $j > i + P$ are null.

The structure of $\mathbf{B}_{i,i}$ is the following,

$$\mathbf{B}_{i,i} = \begin{bmatrix} \mathbf{D}_{0,0} & \mathbf{D}_{0,1} & \mathbf{0} & \mathbf{0} & \cdots & & \\ \mathbf{D}_{1,0} & \mathbf{D}_{1,1} & \mathbf{D}_{1,2} & \mathbf{0} & \cdots & & \\ \mathbf{0} & \mathbf{D}_{2,1} & \mathbf{D}_{2,2} & \mathbf{D}_{2,3} & \cdots & & \\ \vdots & \ddots & & & & \ddots & \\ \mathbf{0} & \cdots & & \mathbf{0} & \mathbf{D}_{D-1,D} & \mathbf{D}_{D,D} \end{bmatrix}$$

where blocks $\mathbf{D}_{k,l}$ are $(N+1) \times (N+1)$ and correspond to changes in the number of active data sources from k to l. The diagonal blocks $\mathbf{D}_{k,k+1}$ collect transitions corresponding to

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B}_{0,0} & \mathbf{B}_{0,1} & \cdots & \mathbf{B}_{0,P} & \mathbf{0} & \mathbf{0} & & \\ \mathbf{B}_{1,0} & \mathbf{B}_{1,1} & \cdots & \mathbf{B}_{1,P} & \mathbf{B}_{1,P+1} & \mathbf{0} & & \\ \mathbf{0} & \mathbf{B}_{2,1} & \cdots & \mathbf{B}_{2,P} & \mathbf{B}_{2,P+1} & \mathbf{B}_{2,P+2} & \mathbf{0} & \cdots \\ \vdots & \ddots & & & & & & \ddots \\ \mathbf{0} & \cdots & & & & \mathbf{B}_{B-1,B-2} & \mathbf{B}_{B-1,B-1} & \mathbf{B}_{B,B+1} \\ \mathbf{0} & \cdots & & & & \mathbf{0} & \mathbf{B}_{B-1,B} & \mathbf{B}_{B,B} \end{bmatrix}$$

Fig. 2. Structure of the infinitesimal generator matrix $\mathbf{Q}$.

the activation of new data sources. When d sources are active, the activation rate is equal to $(D-d)\mu_R$, hence we have:

$$\mathbf{D}_{k,k+1} = \mu_R \mathbf{D}^{(-)}$$

where the diagonal block $\mathbf{D}^{(-)}$ contains integers from D to 0 along the main diagonal:

$$\mathbf{D}^{(-)} = \begin{bmatrix} D & 0 & 0 & \cdots & \\ 0 & (D-1) & 0 & \cdots & \\ 0 & 0 & (D-2) & \cdots & \\ \vdots & & & & \\ 0 & \cdots & & 1 & 0 \\ 0 & \cdots & & 0 & 0 \end{bmatrix}$$

Similarly, since the rate by which sources switch off is proportional to μ_C and to the number of active sources, $\mathbf{D}_{k,k-1}$ is

$$\mathbf{D}_{k,k-1} = \mu_C \mathbf{D}^{(+)}$$

with

$$\mathbf{D}^{(+)} = \begin{bmatrix} 0 & 0 & 0 & \cdots & \\ 0 & 1 & 0 & \cdots & \\ 0 & 0 & 2 & \cdots & \\ \vdots & & & & \\ 0 & \cdots & & D-1 & 0 \\ 0 & \cdots & & 0 & D \end{bmatrix}$$

The tridiagonal blocks $\mathbf{D}_{k,k}$ describe the dynamics of voice calls:

$$\mathbf{D}_{k,k} = \begin{bmatrix} X & \lambda_v & 0 & \cdots & \\ \mu_v & X & \lambda_v & \cdots & \\ 0 & 2\mu_v & X & \lambda_v & \cdots \\ \vdots & & & & \\ 0 & \cdots & (N-1)\mu_v & X & \lambda_v \\ 0 & \cdots & & N\mu_v & X \end{bmatrix}$$

The terms X along the main diagonal of $\mathbf{D}_{k,k}$ correspond also to terms along the main diagonal of $\mathbf{Q}$, and are adjusted so that the rows of $\mathbf{Q}$ sum to 0.

Going back to the structure of $\mathbf{Q}$, the blocks $\mathbf{B}_{i,i+k}$ are related to the arrival at the buffer of batches of radio blocks. The batch size is equal to k with probability p_k, and the arrival rate is proportional to the number of active data sources and to the IP packet generation rate μ_D. Thus, we have:

$$\mathbf{B}_{i,i+k} = \mu_D p_k \mathbf{D}^{(+)} \otimes \mathbf{I}_{N+1} \quad \text{for } k = 1, 2, \cdots P \quad \text{and } i+k \leq B \tag{1}$$

where $\mathbf{I}_{N+1}$ is the $(N+1)\times(N+1)$ identity matrix, and $\otimes$ is the Kronecker product. We assume that when not all the radio blocks composing an IP packet can be accommodated in the buffer, the whole packet is lost. Thus, the structure reported in (1) holds even in the lower right corner of $\mathbf{Q}$. Finally, blocks $\mathbf{B}_{i,i-1}$ collect the transitions which describe the successful transmissions of radio blocks,

$$\mathbf{B}_{i,i-1} = \mathbf{I}_{D+1} \otimes \left(c\gamma \mathbf{N}_i^{(-)} \right)$$

where $\mathbf{N}_i^{(-)}$ accounts for the number of radio blocks which can be transmitted during the same set of four frames. Radio blocks are transmitted employing all the resources not used by voice connections: when v voice calls are active, $N-v$ channels are used if at least $N-v$ radio blocks are in the buffer.

$$\mathbf{N}_i^{(-)} = \begin{bmatrix} \min(N,i) & 0 & \cdots & \\ 0 & \min(N-1,i) & \cdots & \\ \vdots & & & \\ 0 & \cdots & \min(1,i) & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix}$$

Let $\pi(\mathbf{s})$ be the steady state probability of state $\mathbf{s}$ (we will also write $\pi(b,d,v)$) and let the vector $\boldsymbol{\pi}$ collect the steady state probabilities of all states in $\mathcal{S}$. From the flow balance equations

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$$

together with the normalization condition, we compute $\boldsymbol{\pi}$, by standard techniques for the solution of CTMC. In particular, we employ the block reduction method.

The value of the arrival rate of incoming handover requests, λ_h, is derived by balancing incoming and outgoing handover flows for voice users in a fixed point procedure. The outgoing handover flow is computed as:

$$\lambda_h^{(\text{out})} = \sum_{v=1}^{N} \sum_{d=0}^{D} \sum_{b=0}^{B} v\mu_h \pi(b,d,v)\,.$$

The approach of using a fixed point procedure to compute incoming handover flows was widely used in the literature for the analysis of cellular systems, see for example, [4], [5], [2].

B. Performance Metrics

From $\boldsymbol{\pi}$ some interesting performance metrics can be computed. Let O and O_{IP} be the offered traffic in radio blocks and in IP packets.

$$O = \frac{D\mu_R}{\mu_C + \mu_R} \mu_D \sum_{i=1}^{P} i p_i \tag{2}$$

where the first term is the average number of active packet calls, and the sum is the average size of batches. Similarly,

$O_{\rm IP}$ is given by:

$$O_{\rm IP} = \frac{D\mu_R}{\mu_C + \mu_R}\mu_D \sum_{i=1}^{P} p_i \,. \tag{3}$$

By accounting for all the cases in which the buffer overflows, the probability that a radio block is lost can be computed as:

$$L = \frac{1}{O}\sum_{v=0}^{N}\sum_{d=1}^{D}\sum_{b=B-P+1}^{B}\sum_{i=B-b+1}^{P} d\mu_D i p_i \pi(b,d,v) \,. \tag{4}$$

The probability that an IP packets is lost is, instead, given by:

$$L_{\rm IP} = \frac{1}{O_{\rm IP}} \cdot \sum_{v=0}^{N}\sum_{d=1}^{D}\sum_{b=B-P+1}^{B}\sum_{i=B-b+1}^{P} d\mu_D p_i \pi(b,d,v) \,. \tag{5}$$

The throughput in radio blocks and in IP packets is, respectively,

$$X = O\,(1-L) \qquad \text{and} \qquad X_{\rm IP} = O_{\rm IP}\,(1-L_{\rm IP}) \,. \tag{6}$$

The average buffer occupancy is computed from:

$$E[b] = \sum_{v=0}^{N}\sum_{d=0}^{D}\sum_{b=1}^{B} b\pi(b,d,v) \,. \tag{7}$$

We also evaluate the average buffer occupancy given that v voice calls are active:

$$E[b|v] = \frac{\sum_{d=0}^{D}\sum_{b=1}^{B} b\pi(b,d,v)}{\sum_{d=0}^{D}\sum_{b=0}^{B}\pi(b,d,v)} \,. \tag{8}$$

The voice call blocking probability is given by the probability that all channels are busy with voice connections:

$$L_v = \sum_{d=0}^{D}\sum_{b=0}^{B}\pi(b,d,N) \,. \tag{9}$$

Observe that, since voice has priority over data, the presence of data traffic is transparent to voice users; thus, L_v can also be computed by simply applying the Erlang-B formula.

Some of the main QoS metrics for data services are related to delay. By applying Little's formula, the average delay perceived by radio blocks can be easily computed as:

$$E[T] = \frac{E[b]}{X} \,. \tag{10}$$

However, for data networks design and planning, some of the most interesting QoS metrics are related to the delay distribution. A typical example is the probability P_M that packets perceive a delay longer than a maximum allowable value D_M. In fact, many protocols interpret excessive delays as indications of packet losses (TCP is a relevant example). In the case of protocols developed to carry real-time traffic, these losses are not recovered, and translate into a deterioration of the quality of the communication. In the case of protocols aimed at the reliable transfer of user information, a packet loss is recovered by retransmission, and the QoS deterioration is perceived by the user in terms of large delays in accessing the requested information.

Unfortunately, delay distributions are quite hard to compute. We therefore exploit some characteristics of our system in order to derive a simple approximate formula based on the steady-state probabilities only. In particular, we exploit the fact that voice dynamics are on a much slower time scale than data traffic. We study data delays given the number of active voice calls, v, as if v were constant. More formally, letting the random variable T be the delay perceived by a radio block, we write the cumulative distribution function (CDF) $F(t)$ as,

$$F(t) = P\{T \le t\} = \sum_{v=0}^{N} P\{T \le t|v\}P\{v\} \tag{11}$$

where $P\{v\}$ is the probability that v voice calls are active. Consider the cases with $v < N$. Since v is assumed to be constant, the rate at which radio blocks are removed from the buffer is also constant and equal to $(N-v)c\gamma$. When a radio block arrives and finds b radio blocks already in the buffer, it perceives a delay given by the sum of $(b+1)$ services times (its own service time included), where each service time is negative exponentially distributed with mean $1/[(N-v)c\gamma]$. The sum of $(b+1)$ exponential random variables is distributed according to an Erlang-$(b+1)$, whose variance decreases with increasing values of b. Therefore, we introduce only a small error by assuming that the delay is deterministically equal to the mean delay[2]. A radio block which finds buffer occupancy b is thus assumed to experience a constant delay equal to $(b+1)/[(N-v)c\gamma]$. Now, in order to derive $P\{T \le t|v\}$ in (11), we compute the maximum buffer occupancy $K_v(t)$ which makes the radio block perceive a delay smaller than t,

$$\frac{K_v(t)+1}{c\gamma(N-v)} = t \tag{12}$$

from which we derive

$$K_v(t) = \lfloor tc\gamma(N-v)\rfloor - 1 \qquad \text{for } v < N \,. \tag{13}$$

Consider now the case $v = N$ (this case never occurs when the R-reservation policy is adopted). Since all the channels are busy with voice connections, the delay perceived by a radio block which enters the buffer and finds b radio blocks, is given by two contributions: the time till a channel is released by a voice connection and, the $(b+1)$ service times with just one channel available. Thus, we have,

$$\frac{K_N(t)+1}{c\gamma} + \frac{1}{N\mu_v} = t \tag{14}$$

from which we derive,

$$K_N(t) = \max\left(0, \lfloor c\gamma\left(t - \frac{1}{N\mu_v}\right)\rfloor - 1\right) \,. \tag{15}$$

Note that the case $v = N$ contributes only for values of t larger than $\frac{1}{N\mu_v}$.

A radio block perceives a delay smaller than t if, at its arrival at the buffer, the number of radio blocks before it in the buffer is smaller than $K_v(t)$. Suppose that an IP packet finds, at its arrival at the buffer, b radio blocks already in the

[2]A further motivation for this assumption is that the radio block transmission time is constant in reality.

buffer. The buffer occupancy observed by the first radio block of the burst coincides with the occupancy observed by the IP packet. The second radio block in the burst observes $b+1$ radio blocks in the buffer with the same probability, the third one observes $b+2$ radio blocks and so on. Then, since IP packets arrive at the buffer according to a Markov modulated Poisson process, we can write,

$$F(t) = P\{T \le t\} = \frac{1}{X} \sum_{v=0}^{N} \sum_{b=0}^{K_v(t)} \sum_{i=1}^{P} \sum_{d=1}^{D} p_i f_i(b) d \mu_D \pi(b,d,v) \quad (16)$$

$$\text{with } f_i(b) = \begin{cases} i & \text{if } b+i \le K_v(t) \\ K_v(t) - b & \text{if } b+i > K_v(t) \end{cases}$$

and

$$K_v(t) = \begin{cases} \lfloor tc\gamma(N-v) \rfloor - 1 & \text{for } v < N \\ \max\left(0, \lfloor c\gamma \left(t - \frac{1}{N\mu_v}\right) \rfloor - 1\right) & \text{for } v = N \end{cases} \quad (17)$$

where the term $f_i(b)$ accounts for the number of radio blocks in an IP packet (possibly all of them) which perceive a delay smaller than t.

Similarly, we can derive the cumulative distribution function of the delay perceived by an IP packet,

$$F_{\text{IP}}(t) = \frac{1}{X_{\text{IP}}} \sum_{v=0}^{N} \sum_{b=0}^{K_v(t)} \sum_{i=1}^{P} \sum_{d=1}^{D} p_i g_i(b) d \mu_D \pi(b,d,v) \quad (18)$$

$$\text{with } g_i(b) = \begin{cases} 1 & \text{if } b+i \le K_v(t) \\ 0 & \text{if } b+i > K_v(t) \end{cases}$$

and $K_v(t)$ as in (17). Notice that, as accounted for by the term $g_i(b)$, the delay perceived by an IP packets is equal to the delay of the last radio block of the packet.

Summarizing, in order to derive (16) and (18) we introduced two approximations. First, we decomposed the system and found the delay distribution given the number of active voice calls as if it were constant. Second, we substituted Erlang-n distributions by deterministic ones. As will be shown in the next section, these approximations have a marginal impact, and the estimates of delay distribution provided by (16) and (18) are extremely accurate.

C. R-reservation policy

We explain in this section how the proposed model can be extended in order to describe the R-reservation channel allocation policy.

As already mentioned, in a cell which adopts the R-reservation policy, R channels are reserved to data traffic, while the remaining $N-R$ channels are shared between voice and data traffic, with priority to voice.

The Markovian model introduced in Section III-A can be used also to describe the behavior of a cell adopting the R-reservation policy. The only difference is that under the R-reservation channel allocation policy, no more than $N-R$ voice calls can be accepted. Therefore, under the R-reservation channel allocation policy, the state variable v ranges from 0 to $N-R$. The state space cardinality reduces to $(B+1)(D+1)(N-R+1)$ and the block sizes in matrix $\mathbf{Q}$ change accordingly.

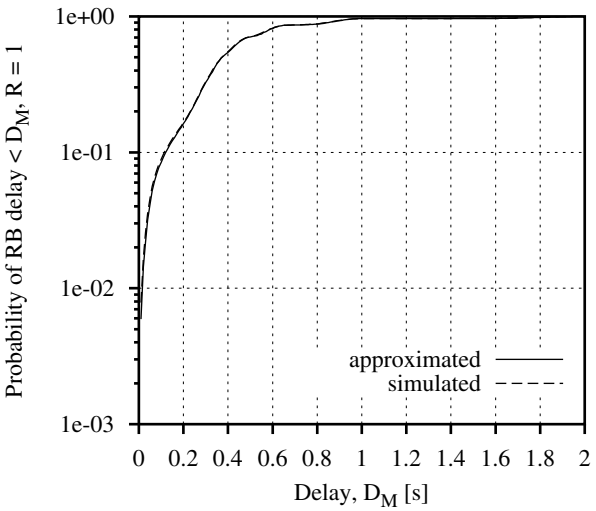

Fig. 3. CDF of radio block delay. 10 data sources and R-reservation channel allocation policy, $R=1$, $\lambda_v = 1/30$ s^{-1}, $\mu_D = 20$ s^{-1}.

As we will observe by means of numerical results in Section IV, the adoption of the R-reservation channel allocation policy implies an improvement of the QoS of the data transfer service. The cost of this improvement is paid in terms of a QoS deterioration of the telephony service. The voice call blocking probability is in this case:

$$P_v = \sum_{d=0}^{D} \sum_{b=0}^{B} \pi(b,d,N-R) \quad (19)$$

which is larger than for the voice priority policy.

IV. Numerical Results

In this section, we validate the accuracy of the proposed model by comparison against simulation results obtained by a discrete-event simulator. The main difference in the assumptions lying below the simulation and analytical models concerns the radio block transmission time. While in the simulator the transmission time of a radio block is constant and equal to four GSM frame times, in the analytical model an exponentially distributed transmission time is assumed, with mean value equal to four GSM frame times. The parameters of the considered scenarios are summarized in Table I. For the sake of simplicity, we assume that the packet size is equal to either 1 or P 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.

Fig. 3 shows the cumulative distribution function (CDF) of the radio block delay in the case of R-reservation policy with $R=1$, 10 data sources ($D=10$), $\lambda_v = 1/30$ s^{-1}, $\mu_D = 20$ s^{-1}. The solid line refers to analytical results, the dashed line to simulations. In order to observe the tail of the CDF we also show in Fig. 4 the complement of the CDF. The smooth step behavior that can be observed in Fig. 4 indicates

TABLE I

PARAMETERS OF THE CONSIDERED SCENARIOS

Parameter	Value
No. of traffic channels in the cells, N	7
No. of channels reserved to data, R	0,1
$1/\mu$	180 s
μ_h	$\mu/2$
Buffer size, B	100
No. of packet sessions, D	10, ..., 50
Max. no. of radio blocks per packet	$P = 16$
Distribution of the no. of radio blocks per packet	$p_1 = 0.5$ $p_{16} = 0.5$
N_P	25
$1/\mu_R$	41.2 s
c	0.95

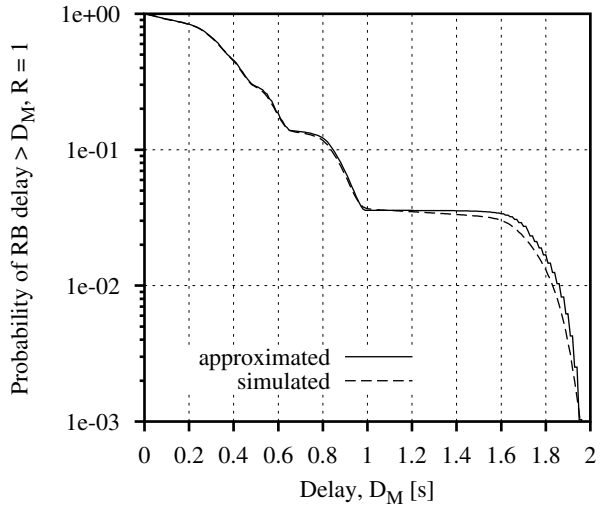

Fig. 4. Complement of CDF of radio block delay. 10 data sources and R-reservation channel allocation policy, $R = 1$, $\lambda_v = 1/30\ \mathrm{s}^{-1}$, $\mu_D = 20\ \mathrm{s}^{-1}$.

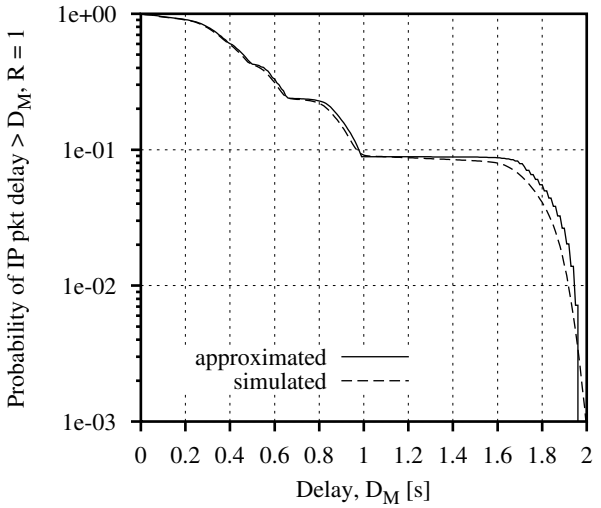

Fig. 5. Complement of CDF of IP packet delay. 10 data sources and R-reservation channel allocation policy, $R = 1$, $\lambda_v = 1/30\ \mathrm{s}^{-1}$, $\mu_D = 20\ \mathrm{s}^{-1}$.

that the probability density function of the radio block delay exhibits a kind of multi-modal behavior. Depending on the number of active voice calls, the radio block service rate changes remarkably, so that the delay perceived by radio blocks tends to concentrate around values which depend on the number of active voice calls. This phenomenon is also emphasized by the different time scales of voice and data traffic dynamics. Despite being approximate, the analytical model provides extremely accurate estimations of the CDF, even for small values of probability. Moreover, due to its simplicity, the computational cost of the analytical approach is much smaller than that of simulation. It took us almost one hour CPU time to obtain the simulation results shown in Figs. 3 and 4, while 4 minutes only were required for deriving the analytical results.

For the same scenario, the complement of the CDF for the delay of IP packets is shown in Fig. 5. Notice, again, the accuracy of the analytical predictions even for the performance at the IP packet level.

Fig. 6 shows the impact of the number of data sources on the probability that the radio block delay is larger than a delay constraint D_M in the case of the voice priority channel allocation policy. The value of μ_D is set so that the total data traffic load is kept constant. The voice traffic load in the x-axis is given by λ_v/μ_v. When the number of sources is small, the traffic is burstier; on the contrary, the superposition of On/Off sources reduces the burstiness of the total traffic. However, observe that the impact of the number of sources is marginal. The impact of the burstiness of On/Off sources is slightly more evident when the radio block loss probability is considered, as can be seen in Fig. 7 in the case of the R-reservation channel allocation policy. We conclude that for the QoS assessment we should use the On/Off model when a small number of data sources is considered, i.e., with less than 10 or 20 sources. When a larger number of sources is considered, we can as well adopt a simple Poisson arrival process of data packets (see the solid line referring to Poisson traffic in Fig. 7). The reason for this conclusion is partially due to the fact that the multiplexing of a number of independent On/Off sources with exponential On and Off times tends to a Poisson process, and partially due to the effect of the very different time scales of voice and data dynamics: the fact that voice is so much slower than data makes the performance of data essentially depend on the steady-state voice behavior, and on the *average* data arrival rate, thus canceling the effect of the short-term burstiness of data sources.

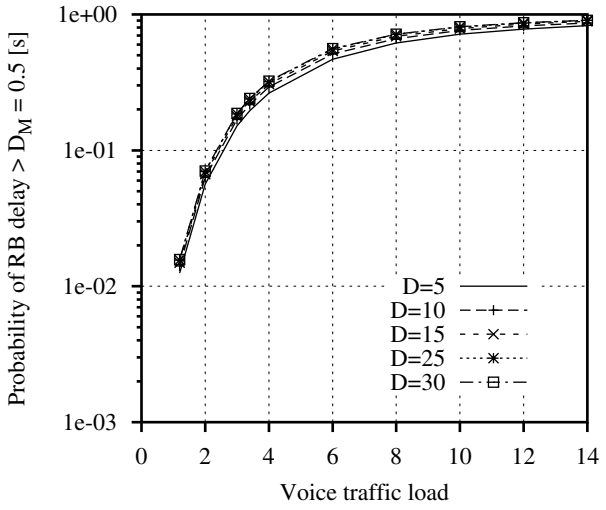

Fig. 6. Probability that radio block delay is larger than the delay constraint $D_M = 0.5$ s, versus the voice traffic load and for different number of data sources. Voice priority channel allocation policy.

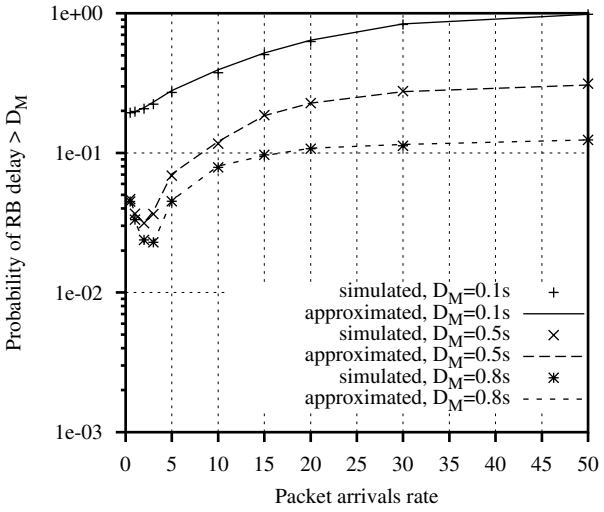

Fig. 8. Probability that radio block delay is larger than the delay constraint D_M versus data packet arrival rate. Poisson data traffic, voice priority channel allocation policy, $\lambda_v = 1/30$ s^{-1}.

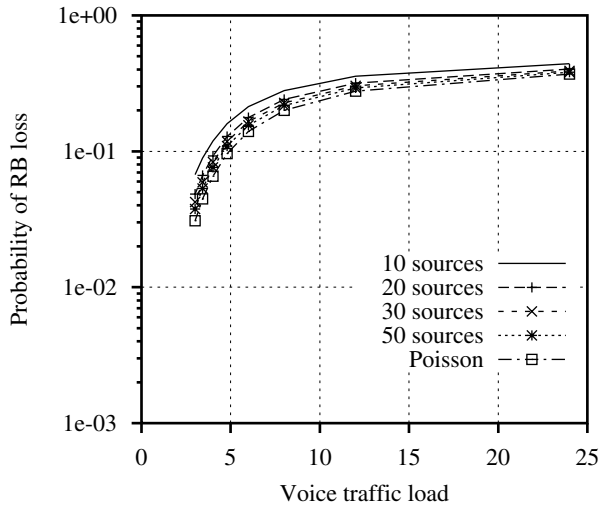

Fig. 7. Radio block loss probability versus the voice traffic load for different number of data sources and for Poisson data traffic. R-reservation channel allocation policy, with R =1.

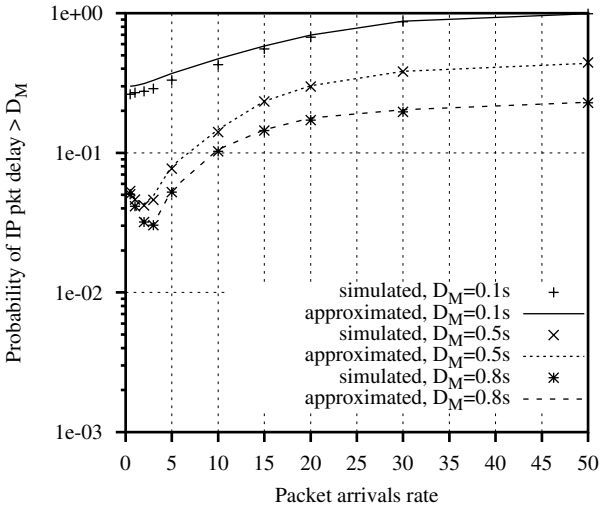

Fig. 9. Probability that IP packet delay is larger than the delay constraint D_M versus data packet arrival rate. Poisson data traffic, voice priority channel allocation policy, $\lambda_v = 1/30$ s^{-1}.

We now evaluate the impact of the data traffic load on the probability that a constraint on the maximum radio block delay can be met. We plot in Fig. 8 the probability that the radio block delay is larger than a constraint D_M. The voice priority channel allocation scheme is adopted, the arrival rate of voice calls is equal to $1/30$ s^{-1}, and Poisson data traffic is considered. Of course, as the constraint becomes less tight (i.e., D_M increases) the probability of not being able to meet the constraint decreases. The decreasing behavior of the curve that can be observed for small values of the load is due to the fact that the delay is computed only for those packets which enter the buffer. By increasing the load we cause higher losses which occur mainly for large values of v, so that the fraction of packets which perceive short delays (i.e., enter when v is small) increases. When the load becomes high, however, the increased delay for all values of v dominates, and the curve monotonically increases with the data load.

Fig. 9 reports similar results for IP packets. Of course, the probabilities referring to whole IP packets are higher (recall that the IP packet delay corresponds to the delay of the last radio block in the burst).

The impact of voice traffic can be observed in Figs. 10 and 11 for radio block and IP packet delay, respectively. The voice priority scheme is adopted, the delay constraint is D_M =0.8 s. Clearly, the increase of voice traffic causes the deterioration of the QoS perceived by radio blocks, which is expressed in our case by the increased probability that the QoS constraint on maximum delay cannot be met.

V. Conclusions

In this paper we described an approximate Markovian model for the estimation of the packet delay distribution in a cell of a GSM/GPRS network simultaneously supporting voice and data services. In addition, we validated the analytical model by comparison against discrete-event simulation of the system, and we showed how the model results can be instrumental for

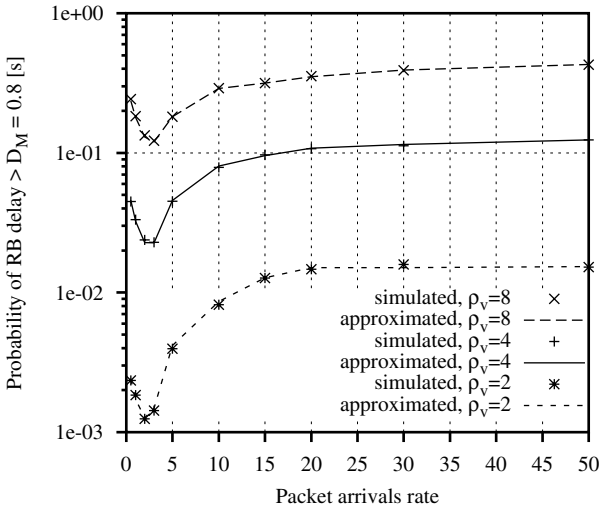

Fig. 10. Probability that radio block delay is larger than the delay constraint D_M =0.8 s versus data packet arrival rate for different values of the voice traffic load. Poisson data traffic, voice priority channel allocation policy.

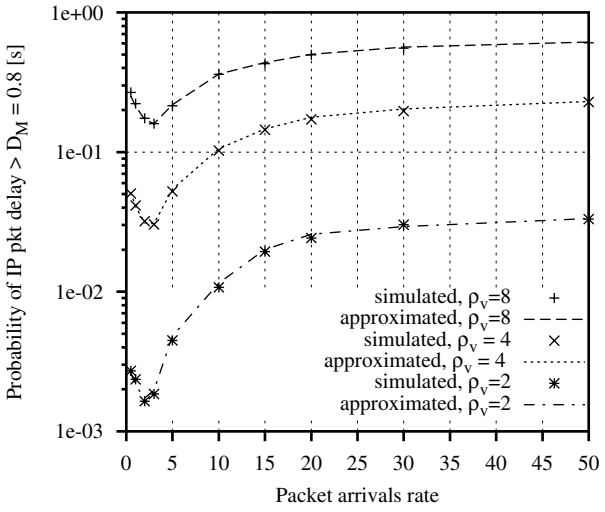

Fig. 11. Probability that IP packet delay is larger than the delay constraint D_M =0.8 s versus data packet arrival rate for different values of the voice traffic load. Poisson data traffic, voice priority channel allocation policy.

the dimensioning of the cell resources, and for the assessment of the effectiveness of the channel allocation policies to voice and data.

The presented model is applied to only one cell, but it can serve as the basic building block for the complete design and planning of multi-cell wireless voice and data networks, possibly adopting a layered cell architecture, like in 900-1800 MHz GSM/GPRS systems.

REFERENCES

[1] M. Meo, M. Ajmone Marsan, C. Batetta, "Resource Management Policies in GPRS Wireless Internet Access Systems," *IEEE International Performance and Dependability Symposium*, Washington, DC, June 23rd - 26th, 2002.

[2] M. Meo, M. Ajmone Marsan, "Approximate Analytical Models for Dual-Band GSM Networks Design and Planning," *Infocom 2000*, Tel Aviv, Israel, March 2000.

[3] 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.

[4] D. Hong, S. Rappaport, "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Non-Prioritized Handoff Procedures," *IEEE Transactions on Vehicular Technology*, Vol. VT-35, N. 3, pp. 77-92, August 1986.

[5] K. K. Leung, W. A. Massey, W. Whitt, "Traffic Models for Wireless Communication Networks," *IEEE JSAC*, Vol. 12, N. 8, pp. 1353-1364, October 1994.

[6] Y. Lin, "Modeling Techniques for Large-Scale PCS Networks," *IEEE Communication Magazine*, Vol. 35, N. 2, pp. 102-107, February 1997.

[7] L.R. Hu, S.S. Rappaport, "Personal Communication Systems Using Multiple Hierarchical Cellular Overlays," *IEEE ICUPC'94*, San Diego, CA, September 1994.

[8] "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).

[9] M. Ajmone Marsan, M. Gribaudo, M. Meo, M. Sereno, "On Petri Net-Based Modeling Paradigms for the Performance Analysis of Wireless Internet Accesses," *9th International Workshop on Petri Nets and Performance Models,* September 11-14, 2001, RWTH Aachen, Germany.