An improved model for GSM/GPRS/EDGE performance evaluation

COPCA*
Facultad de Ingeniería, Universidad de la República
ANTEL: Administración Nacional de Telecomunicaciones
Montevideo, Uruguay
copca@fing.edu.uy

ABSTRACT

Nowadays massive access to the new cellular technologies is no longer an idea but a tangible reality. Among them, the GSM/GPRS/EDGE architecture represents without any question the most worldwide spread of them. Costs decreases in both infrastructure and personal equipment has largely increased the number of users. This abrupt growth poses a difficult challenge for telecom operators when it comes to dimensioning and evaluating the performance of their networks.

Previous works have studied the problem of modelling GSM/GPRS/EDGE networks, paying little attention to many technical implementation features of determinant impact. This work addresses the performance evaluation problem of these networks, focusing on the influence of different operational details. Different data and voice models are studied, including a traffic prioritization model of great relevance facing the increasing popularity of new multimedia services offered through these networks. The proposed models are finally applied to the performance evaluation of the local operator's network.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: [design studies, measurement techniques, modeling techniques, performance attributes]; G.3 [**Probability and Statistics**]: [queueing theory, markov processes]

General Terms

Performance, Measurement, Algorithms, Design.

*COPCA is a joint research group of the Electrical Engineering Faculty (Universidad de la República) and ANTEL. Members: Pablo Belzarena, Paola Bermolen, Pedro Casas, Andrés Ferragut, Federico Larroca, Valeria Meilan, Javier Pereira, Natalia Pignataro, Sergio Nesmachnow, Franco Robledo, Bruno Bazzano, Carlos Piana and Alejandro Reyna.

Keywords

Network performance evaluation, cellular networks modelling, GSM-GPRS/EDGE, traffic priorization.

1. INTRODUCTION

Initially, GSM (Global System for Mobile Application, standard developed by 3GPP, 3rd Generation Partnership Project [1]) cellular networks were developed to offer, mainly, telephonic services. Only a basic circuit oriented service at low transfer rate was provided to data access. This service, named CSD (Circuit Switched Data), can achieve transfer rates up to 9.6 kbps per time slot. However, with the arise of GPRS/EDGE (General Packet Radio Service, Enhanced Data rates for GSM Evolution), data services through GSM cellular networks have been increasing in popularity. Analysis and study of data services in a cellular network require new models, since traditional ones (like Erlang's formulas) are not applicable to this kind of traffic. At the same time, particularities of each GSM/GPRS/EDGE network, mainly on the resources assignment, require an adaptation of the general model depending on the needs of each provider.

This work analyzes the dimensioning and performance evaluation of a GSM/GPRS/EDGE cell, taking into account some technical characteristics of this architecture operational equipments. The focus is on modeling data services but considering carefully the voice/data interaction.

To study the performance of data applications on the GSM/GPRS/EDGE network, two Quality of Service (QoS) parameters were considered, *blocking probability* and *throughput* for a typical user of a certain *connection type*.[1] This is due to the fact GPRS/EDGE networks are designed and used by most of providers as a best effort network. The majority of carried traffic is elastic (Web, Wap, MMS, mail) and its most important QoS parameter is throughput (how long it takes to finish the transfer). Real time multimedia applications are sensitive to other parameters besides throughput, like delays, jitter and packet loss, but GPRS/EDGE networks have not been thought to provide any guarantees over these parameters. A multimedia service offered by many providers over GPRS/EDGE is the Push To Talk (PTT) service. However, because the way this application operates (less interactive than a normal conversation), if a given amount of throughput is assured, it is possible to decide wether or not the minimum requirements are granted.

[1]By connection type we understand the different data services connections that are carried through the network, as Web, Wap, MMS, etc.

All models presented are based on the analysis of resources in a given cell (basically time slots) and on the way that these are shared by the users. It was not pretended to study the problems that may arise after the BSC (Base Station Controller) and the rest of the network. This is because the air interface is normally the bottleneck of the system. Moreover, the air interface introduces inherent problems like interference, retransmissions, delays and packets loss. In order to alleviate this problems, GPRS/EDGE implements an adaptive codification rate depending on the carrier to interference relation ($\frac{c}{i}$). So, if this relation is under certain levels, the information is encoded with more redundancy bits, which yields to transfer effective information at lower rates.

The aspects concerning mobility, like handover, were not considered because it was assumed that such issues were not as relevant as the resource sharing.

The rest of the paper is organized as follows: in section 2 there is a brief description of related models for GSM and GPRS/EDGE systems. In section 3 the model for a single cell with GSM/GPRS/EDGE service interaction is described. In section 4 the model is extended to include traffic differentiation between data services. In section 5 a method to analyze real data from a GSM/GPRS/EDGE cell and adjust the model parameters is presented. Finally, in section 6 the influence of different modifications on model paramters in the performance of the system is analyzed, and the conclusions are given in section 7.

2. RELATED WORK

There has been much research concerning modelling and dimensioning cellular networks.

In [5], the GPRS/EDGE network is modeled assuming that packets that wait to be served are in a *SSQ* (single server queue). Besides, it is supposed that the arrival process is a Markov Modulated Poisson Process (MMPP). Finally, matrix based analytical models are used in order to obtain numerical results over the stationary distribution of the queue size. With this result, the mean packet delay is obtained. The main drawback of this model is that all users are modeled with the same MMPP (lacking flexibility) and minor errors in the modeling process yields to considerable errors on the results. Besides, the model is focused on the delays of the queue and not in the throughput.

In [9], an analytical model for the GPRS/EDGE air interface is presented. The interaction between GSM and GPRS/EDGE connections are analyzed over a dynamic scheme of channel assignment. The resulting model is used to find how many data channels have to be assigned to GPRS/EDGE under a certain condition of GSM traffic, in order to guarantee a given level of QoS. However, the focus of the model is on the analysis of the handover impact over data service which is is not considered in this work.

In [8], different ways of assigning bandwidth for a multiservice cellular network are considered, aiming at guaranteeing certain QoS level for different applications. Two different bandwidth assignment strategies are presented through simulations and theoretical analysis. The authors present a theoretical analysis of the performance for both schemes through Markovian models. An homogeneous network is assumed, where cells have the same number of channels and the same arrival rate of calls and handoff petitions (both for data and voice), considered as Poisson process.

In [7], general QoS concepts like delay, throughput and service precedence are taken into account. Moreover, ETSI recommended values for this parameters are presented. The authors propose a certain combination of different techniques to achieve this levels of QoS, like admission control of calls, resource reservation and the implementation of scheduling mechanisms.

In [3] and [2], the authors present analytical flow models of the GPRS/EDGE network, giving explicit formulas for cell dimensioning. Their proposals are based on the Engset model, assuming a finite quantity of users that generates ON/OFF sessions, and a bandwith share between active users. Besides, an analytical GSM/GPRS/EDGE network model is presented, using different resource assignment schemes (complete sharing and partial sharing). This model considers the interaction between voice and data. Dahmouni et al. in [4] extend the model developed in [2, 3] to the case of multiservice networks.

The models developed in [2–4] provide an excellent description of a cell's behaviour at the *flow level* timescale, which is the apropriate timescale for QoS dimensioning. However, certain important capabilities and configuration parameters present in the providers equipment have a deep impact on the network performance, and they are not considered in their work. In this paper, the Dahmouni et. al. model is improved to take account for many specific details of GSM/GPRS/EDGE networks, and the impact of these modifications is thorougly analyzed through comparison with a real network data.

3. MODEL DESCRIPTION

The purpose of this work is to model the behavior of a cell in a network using a GSM/GPRS/EDGE architecture, where voice and data services share the same resources. The model provides information about the blocking probability and throughput of the different services, considering the interaction between them through a detailed resource sharing policy.

For voice traffic (GSM), the guideline will be the classical Erlang's model (c.f. [6]), i.e. the voice calls arrive into the system as a Poisson process with intensity λ_v and each call have a random exponential duration of mean $1/\mu_v$.

For data traffic (GPRS/EDGE), the model will be made at a *flow* level timescale. A *session* (e.g. one user browsing through Web pages) can be modeled as a series of *flows* (each page download) separated by inactivity periods (*think times*) with no data transfer (user's analysis of the information). At this timescale, the traffic offered by the users is modeled as an ON/OFF process. The mean duration of the ON period is T_{ON} and during this time a random amount of traffic with mean B_{ON} (measured in bytes) is offered. These values depend in general of the application being considered (Web traffic, Wap traffic, etc.). This flow level model avoids the problem of characterizing the packet level dynamics which are difficult to describe in a detailed manner. The flow level model was introduced by [10] for wired networks and by [2] in the context of GSM/GPRS/EDGE networks.

In the flow level timescale, each flow acts as a unique job or "telephone call", so we can use the classical techniques of queueing theory (which cannot be applied in the packet timescale, c.f. [10]) to obtain performance results in terms of mean job size and arrival intensity.

The base station of a GSM/GPRS/EDGE cell provides

resource sharing through a TDMA scheme, where two types of timeslots (voice and data) can coexist. Let T be the total number of slots in the cell. The voice traffic is given priority, and C_v slots will be reserved for voice service. For data traffic, a typical strategy is to reserve C_d slots for minimum service, and leave the remaining $T - C_v - C_d$ slots on an "on demand" status. This "on demand" slots will be assigned to, voice traffic in priority, and to data traffic subject to availability. Therefore, the total rate reserved for data traffic in a base station will be variable, and it will depend on the number of timeslots not used for voice. To describe this phenomenon, in the next subsections a model for each traffic type in isolation is presented, to finally describe how they interact following this timeslot sharing policy.

3.1 GPRS/EDGE model

The data traffic model described here is based in the works by Dahmouni et. al. [2–4]. However, the models presented so far do not consider some frequent strategies used by operators in their cells to share the timeslots and enhance performance. Specifically, it is important to consider the possibility to use different Code Schemes in each timeslot and a more complex strategy to share the timeslots between users than the one described in [2].

It is supposed that there are M different traffic types (i.e. corresponding to different data applications). The users of type i will generate independent flows of mean size $E(\sigma_i)$, separated by thinking times of mean $E(\tau_i)$, which are also independent. In this first model, it is assumed that all traffic types are treated equally by the system. Let N_i be the number of users of type i in the cell. In general the number of data users in a cell is small, so we cannot assume that flows arrive as a Poisson process. Instead, the model will be based in the finite population Engest model [6].

Let C_d be the number of timeslots available for GPRS/EDGE traffic at a given time. Throughout this subsection this number is assumed constant. In the original model of [2], each incoming flow will receive a number of d timeslots (if there are enough available), and each timeslot allows a service rate of μ_{gprs}. When all the timeslots are assigned, a simple strategy is to share all of them between all the users present in the system.

If the flow size and think time distributions are exponential, the number of users of each class in the system $j = (j_1, \ldots, j_M)$ will be a Markov birth-death process in M dimensions. Let n_{max} be the maximum number of GPRS/EDGE users that can be active simultaneously. This quantity is bounded by the maximum number of users sharing the same timeslot m, the number of data users in the cell, and the maximum number of active flows in the cell, which is 32 by technological limitations, so n_{max} is given by:

$$n_{max}(C_d) = \min\{N, 32, mC_d\}$$

and the space state of the process is:

$$E = \{(j_1, \ldots, j_M) \,/ j_i \leq N_i \text{ y } j_1 + \ldots + j_M \leq n_{max}\}$$

With the above assumptions, Dahmouni et. al. show that the stationary probability of the system is given by:

$$p(j) = p(0) \frac{\prod_{i=1}^{M} C_{j_i}^{N_i} \rho_i^{j_i}}{\prod_{i=1}^{j_1+\ldots+j_M} \min\left\{d, \frac{C_d}{i}\right\}} \tag{3.1}$$

where $\rho_i = \frac{E(\sigma_i)}{E(\tau_i)} \frac{1}{\mu_{GPRS}}$, C_j^N are the binomial coefficients and $p(0)$ is given by normalization. From (3.1) we can derive the user throughput and the blocking probabilities.

As mentioned before, two important modifications will be introduced in the above model in order to generalize it:

Timeslot assignment. The timeslot assignment policy considered before is somewhat simplistic. Different equipment providers implement different strategies, which affects the throughput of the system. A common assignment policy consists in the following behavior: each incoming flow is assigned (when there are enough timeslots available) to the *same* set of d slots, sharing them until the number of active flows in the set is greater than a threshold T_{TBF}[2]. When the threshold is attained, a new set of d slots is reserved for data, and the new flows will be allocated to this block. The procedure is repeated until all data slots are reserved and after that, all slots are shared equally by the active flows in the system.

This changes in the assignment policy also changes the death rates in the process. For ease of exposition we describe the case $M = 1$, (only one type of data traffic). The transmission rates depend now on T_{TBF} and the number j of active flows in the following way:

$$\mu_j = \min\left\{C_d, d\left(\left[\frac{j-1}{dT_{TBF}}\right] + 1\right)\right\}\mu \tag{3.2}$$

where $\mu = \frac{\mu_{GPRS}}{E(\sigma)}$ and $j = 1, \ldots, n_{max}$ If the number of active flows is less than dT_{TBF}, a rate $d\mu$ is obtained, if it is between dT_{TBF} and $2dT_{TBF}$, the rate is $2d\mu$ and so on until reaching the limit of timeslots where the rate becomes $\frac{C_d}{d}$, i.e. all timeslots are equally shared. In figure 1 we show the transition diagram of the new process. The flow arrival rate is the same as in the original model, that is $\lambda_j = (N - j)\lambda$ where $\lambda = E(\tau)^{-1}$ is the flow arrival rate per user.

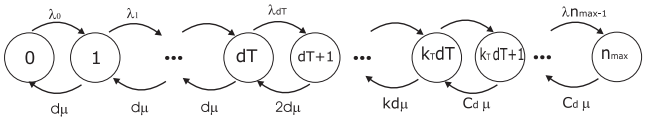

Figure 1: Transition diagram for the timeslot assignment policy

Timeslot transmission rate. In many GPRS/EDGE networks, the transmission rate assigned to a timeslot is not constant. It depends on the Code Scheme used by the endpoints, which allows different transmission rates. For instance, many operators reserve some timeslots for mcs-9 coding which corresponds to EDGE connections, and GPRS/EDGE slots with code schemes CS-1 or CS-2. The number of slots of each kind is a design parameter that the operator will choose considering, for instance, equipment costs.

To incorporate this feature in our model, suppose that there are n_1 timeslots with transmission rate μ_1 and n_2 with rate μ_2 (where $n_1 + n_2 = C_d$). The timeslots with rate μ_1

[2]TBF stands for Temporary Block Flow, the identifier of a flow assigned by the network

are assigned first, and then those with rate μ_2. Again, for ease of presentation we describe the case $M = 1$.

If the number of active flows is j, the death rate becomes:

$$\mu_{GPRS}(j) = \frac{x_j^1\mu_1 + x_j^2\mu_2}{x_j^1 + x_j^2}$$

where x_j^i is the number of slots of type i being used when there are j concurrent flows in the system.

Therefore, we have that:

$$\begin{aligned} x_j^1 &= \min\{k_j d, n_1\} \\ x_j^2 &= \min\left\{(k_j d - n_1)^+, n_2\right\} \end{aligned}$$

where $k_j d$ is the number of slots in use when there are j flows present. As before, this value dependes on the T_{TBF}.

With the two modifications described above, it is possible to generalize equation (3.1). This is done by solving the balance equations of the Markov chain pictured in figure 1, and including the rates $\mu_{GPRS}(j)$ as defined above.

When M classes of data traffic are considered, if $j = (j_1, \ldots, j_M)$ denotes the number of active users of each type, the stationary probability of the process is given by:

$$p(j) = p(0)\frac{(j_1 + \ldots + j_M)! \prod_{i=1}^{M} C_{j_i}^{N_i}}{\prod_{i=0}^{j_1+\ldots+j_M} \min\left\{C_d, d\left(\left[\frac{i-1}{dT_{TBF}}\right] + 1\right)\right\} \mu_{GPRS}(i)} \tag{3.3}$$

From $p(j)$ it can be derived that the mean throughput obtained by one user of type i is:

$$Th_{C_d,N_i} = \frac{\sum_{(j_1,\ldots,j_M)\in E^*} j_i p(j) r(j_1 + \ldots + j_M)}{\sum_{(j_1,\ldots,j_M)\in E^*} j_i p(j)}$$

being E^* the set of j such that $j_i > 0$.

The blocking probability for a user of type i is given by:

$$B_{C_d,N_i} = 1 - \frac{E(\sigma_i)}{E(\tau_i)} \frac{\sum_{(j_1,\ldots,j_M)\in E^*} j_i p(j) r(j_1 + \ldots + j_M)}{\sum_{(j_1,\ldots,j_M)\in E^*} (N_i - j_i) p(j)}$$

where:

$$r(i) = \min\left\{C_d, d\left(\left[\frac{i-1}{dT_{TBF}}\right] + 1\right)\right\} \frac{\mu_{GPRS}(i)}{i}$$

is the throughput obtained by one user when there are i active users in the.

3.2 GSM model

In order to model GSM traffic, as mentioned above, it can be assumed that the number of users present in a cell is big enough to ensure that the call arrivals are a Poisson process of intensity λ_v. Each call will have a random exponential duration of mean $1/\mu_v$. If we assume that the system has $T - C_d$ timeslots available for voice traffic, a simple model for the number of busy timeslots is the Erlang $M/M/T - C_d/T - C_d$ queue, where it is assumes that for each call a full timeslot is given.

However, also in the GSM solutions offered by providers, different strategies are applied in the timeslot assignment in order to improve the number of calls that the network can manage. A common strategy in today solutions is to share the same timeslot between two users, which is called

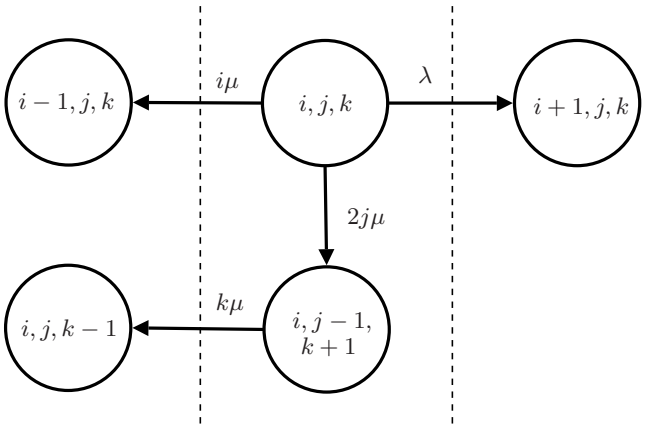

Figure 2: Case 1, *full-rate* arrivals

half-rate assignment. This policy allows to admit more calls into the system with less quality, usually in a congestion situation when there are few timeslots available. In general, a threshold HRTh is defined and when the fraction of free timeslots is less than HRTh, the calls are assigned at *half-rate*.

A problem derived from the above policy is that it can drive the system into a state in which, at a given moment of time, many slots are assigned at *half-rate* but each of them to only one user. This happens for instance when a *half-rate* call ends, freeing half of a slot. These slots will be allocated to voice traffic, and so they are not available for data, leading to inefficiencies in the system. The strategy to avoid this problem is to define a second threshold PACKTh. When the number of free slots is less than PACKTh, the system will reallocate the *half-rate* calls "packing" isolated calls in pairs sharing the same slot, thereby reducing the total number of slots occupied by voice traffic. In what follows, we describe a queueing model that takes account for these two strategies.

Let be the process $N(t)$ which will be a Markov chain where $N(t) = (n_1(t), n_2(t), n_3(t))$, being $n_1(t)$ the number of slots assigned in *full-rate* mode, $n_2(t)$ the number of slots occupied by two users in *half-rate* mode and $n_3(t)$ the number of slots with only one user in *half-rate* mode. In what follows, the transition of $N(t)$ are described.

Case 1. Let $N(t) = (i, j, k)$, if the total number of occupied slots $z = i + j + k$ is such that the fraction of free slots is greater than HRTh, the incoming calls are assigned in *full-rate* mode. The possible transitions are shown in figure 2.

The dotted lines in figure 2 indicates "macro" states where $z = i + j + k$ is constant. There are three death transitions, corresponding to the three possible call endings. A call arrival is assigned *full-rate* so the birth transition increases the first component of N.

Case 2. If the number of occupied slots z is such that the fraction of free slots is between PACKTh and HRTh, the incoming calls are assigned *half-rate* as shown in figure 3 but there is no packing of *half-rate* calls. The difference here is that birth transitions increase the second component of N when there are slots assigned in *half-rate* mode ($k > 0$) and

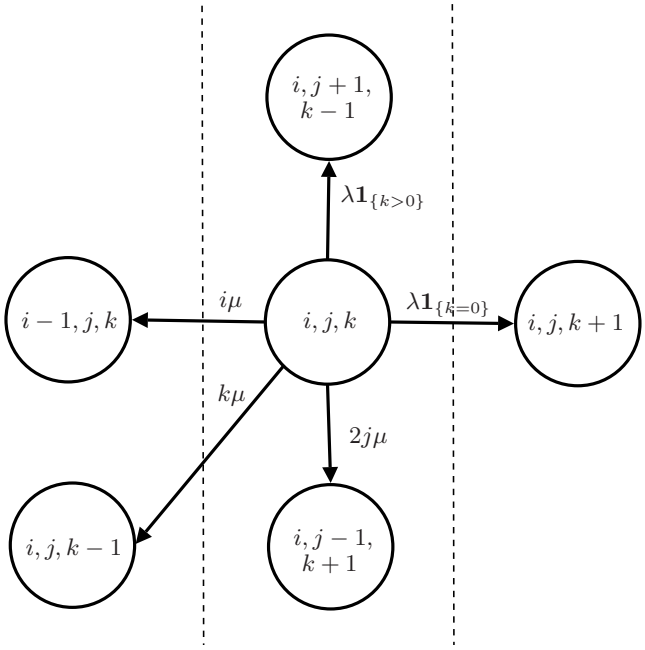

Figure 3: Case 2, *half-rate* arrivals

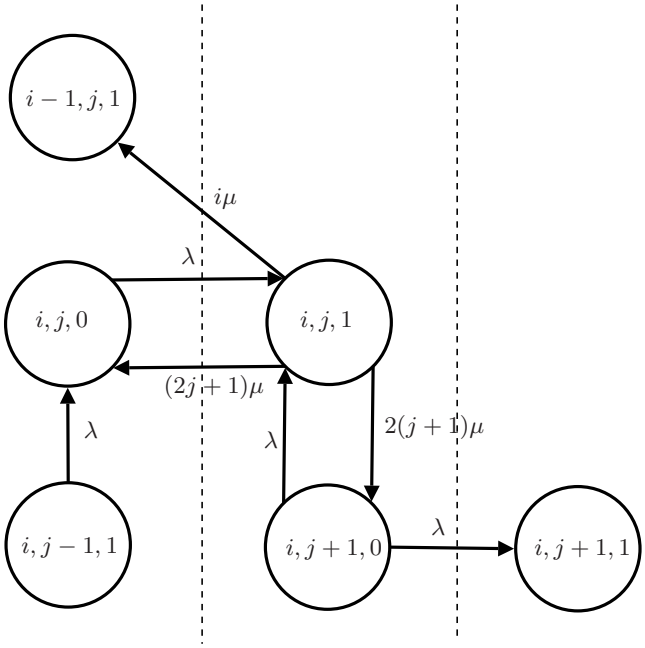

Figure 4: Case 3, *half-rate* arrivals with packing

so the incoming call is merged into a slot with an existing one, or it increases the third component, if there are no isolated *half-rate* calls ($k = 0$).

Case 3. If the number of occupied slots z is such that the fraction of free slots is less than $PACKTh$, the system starts packing *half-rate* calls whenever there are two or more slots assigned at *half-rate*. Therefore, in this case k can only be equal to 0 or 1 and the transitions are shown in figure 4.

As described in the figure, when the system is in state $(i, j, 1)$ and a *half-rate* call ends (either isolated or in a shared slot), the system goes to state $(i, j + 1, 0)$, thus packing the isolated calls. When a new call arrives, it either puts the system in the macro-state $z+1$ if $k = 0$ or it stays in macro state z if $k = 1$, by combining the new call with the isolated *half-rate* one. In all these cases, the first coordinate can only vary between 0 and the maximum number of slots occupied before calls start to be assigned in *half-rate*.

For this model, it is not possible to find an explicit formula for the stationary distribution of the process. Nevertheless, once identified all the state transitions and their rates, it is possible to construct the Q-matrix of the process and solve numerically the equation $\pi Q = 0$ with the normalization condition $\sum \pi(k) = 1$, to find the stationary distribution.

From π, it is possible to calculate $R(z)$, the distribution of the number of slots occupied by voice in the system, by summing $\pi(n)$ over the macro states $\{n = (i, j, k) : i + j + k = z\}$. This distribution will be of use in the modelling of GSM/GPRS/EDGE interaction, as we describe in the next subsection.

3.3 Modelling GSM/GPRS/EDGE interaction

Until now, both traffic types, voice an data are treated in isolation, with a fixed number of slots assigned to each one of them. In general, as it was described at the beginning of the section, there are C_d slots reserved for data C_v slots reserved only for voice and the remaining $T - C_d - C_v$ are shared, with priority to voice traffic. Therefore, the number of available slots for data traffic at any given time is determined by the number of slots occupied by voice traffic. Actually, the probability that there are ξ slots available for GPRS/EDGE is equal to the probability that there are $T - C_d - \xi$ slots occupied by GSM. In the stationary situation, this probability is the result of the analysis done in subsection 3.2. For GSM,the system performance is unchanged by data traffic, due to its priority over data traffic.

To analyze the GPRS/EDGE performance in the integrated environment, a *time scale separation* is assumed, as it is done in [2–4]. In general, data transfers in GPRS/EDGE are shorter interactions with the network, than GSM calls. A flow, like a web page download stays in the system a time which is much shorter than that of a phone call. Therefore, the Markov chain controlling the number of GPRS/EDGE flows in the system operates in a much faster time scale than that controlling GSM calls. Then, it is possible to assume that while there are z slots occupied by GSM, the GPRS/EDGE Markov chain reaches its stationary regime. With this assumption, if $R(z)$ is the stationary probability of finding z slots occupied by GSM , the mean throughput of a GPRS/EDGE connection can be calculated as:

$$Th_d = \sum_{z=0}^{T-C_d} R(z) X_{\min\{T-C_v, T-z\}}$$

and the blocking probability is:

$$B_d = \sum_{z=0}^{T-C_d} R(z) B_{\min\{T-C_v, T-z\}}$$

where $X_C = X_{C,N}$ y $B_C = B_{C,N}$ are obtained by the equations derived in subsection 3.1.

4. MODEL FOR GPRS/EDGE WITH TRAFFIC DIFFERENTIATION.

The GPRS/EDGE model described in the preceding section assumes that the base station makes an equal treatment of each application in the system. However, there are data services which may be more demanding in terms of Quality of Service (QoS) and may be treated in a differential way by the system. One example is the Push To Talk (PTT) service, which is a packetized voice service and requires a minimum throughput, as well as low delay, jitter and loss rate in order to provide an usable service. In this case, it may be of interest for the network operator to introduce differential treatment between the applications.

In what follows, a generalization of the model presented in section 3.1 is developed incorporating priorities between the traffic classes. In section 6 the impact of priorities in the overall performance of the system is discusses.

Two traffic classes labeled 1 and 2 are considered in such a way that whenever there are flows of type 1 in the system, the full capacity of the base station is used to serve these flows, that is, traffic 1 has an absolute priority over traffic 2. This model assumes that flows of type 2 are waiting the end of service of type 1. This is an approximation of a packet level policy where packets from type 1 are handled in priority. The flow level approximation is valid when the load offered by priority traffic is low in comparison with the system load, and that priority flows are composed by bursts of short duration. These assumptions are valid for instance for PTT traffic.

Let N_i be the number of users of type i in the cell. Let $j(t) = (j_1(t), j_2(t))$ be the random process counting the number of active flows of each traffic type at time t. The state space for this process is:

$$E = \{(j_1, j_2)/j_1 \leq n_{1,max}, j_2 \leq n_{2,max}\}$$

where, as before, $n_{i,max} = \min\{32, m \cdot C_d, N_i\}$.

The process $j(t)$ is a Markov birth and death process. The birth rates of this new process are the same that in the non-priority case studied in section 3.1. The death rates are also similar with the following remark: when the system is in a state with $j_1 > 0$, the system only serves type 1 flows at full capacity. When the state of the system is of the form $j_1 = 0, j_2 > 0$, only type 2 traffic is served at full capacity. Summarizing:

$$\lambda_i = \frac{1}{E(\tau_i)} \quad \text{for } i = 1, 2 \tag{4.1}$$

$$\mu_i(j_i) = \min\left\{C_d, d\left(\left[\frac{j_i - 1}{dT}\right] + 1\right)\right\} \frac{\mu_{GPRS}(j_i)}{E(\sigma_i)} \tag{4.2}$$

The transitions are represented in figures 5 and 6.

As before, knowing the transition rates, the Q-matrix of the process can be constructed. Although in this case it is not possible to find an analytical expression of $p(j_1, j_2)$, the stationary distribution, but it is possible to obtain it numerically by solving the equation $pQ = 0$.

Once having the stationary distribution, we can calculate the throughput attained by each traffic class, and the corresponding blocking probabilities by adapting the formulae used in the model without priorities. The results are summarized in the following equations.

Throughput of priority traffic:

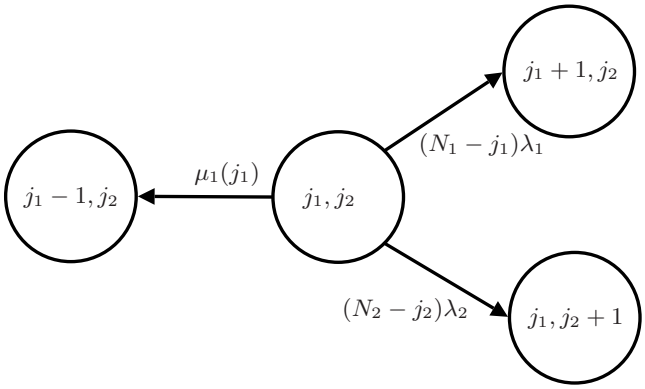

Figure 5: Transitions in the model with priorities, case $j_1 > 0$.

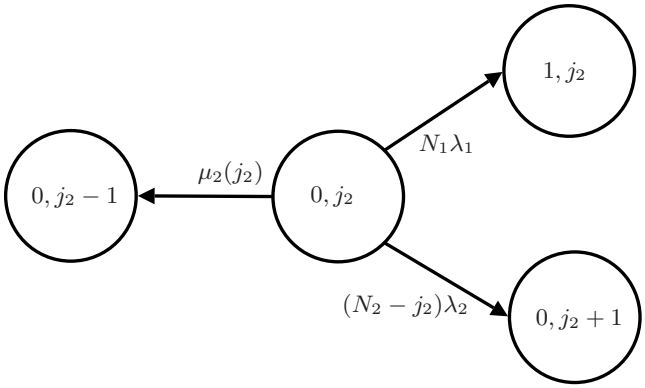

Figure 6: Transitions in the model with priorities, case $j_1 = 0, j_2 > 0$.

$$Th_{C_d,N_1} = \frac{\sum_{\substack{(j_1,j_2)\in E \\ j_1>0}} p(j_1,j_2) \min\left\{C, d \cdot \left(\left[\frac{j_1 - 1}{dT}\right] + 1\right)\right\} \cdot \mu_{GPRS}}{\sum_{(j_1,j_2)\in E} j_1 p(j_1,j_2)}$$

Throughput of non priority traffic:

$$Th_{C_d,N_2} = \frac{\sum_{\substack{(j_1,j_2)\in E^* \\ j_1=0, j_2>0}} p(j_1,j_2) \min\left\{C, d \cdot \left(\left[\frac{j_2 - 1}{dT}\right] + 1\right)\right\} \cdot \mu_{GPRS}}{\sum_{(j_1,j_2)\in E} j_2 p(j_1,j_2)}$$

Assuming that $n_{i,\max} < N_i$, we can also calculate the blocking probabilities. For ease of implementation, we chose a different approach here, that is, to obtain the blocking probability as the ratio between blocked arrival transitions and total arrival transitions, instead of the calculation of section 3.1. The results are:

Blocking probability of priority traffic:

$$B_{C_d,N_1} = \frac{\sum_{j_1=n_{1,max}} p(j_1,j_2) \cdot (N_1 - j_1)}{\sum_{j\in E} p(j_1,j_2) \cdot (N_1 - j_1)}$$

Blocking probability of non priority traffic:

$$B_{C_d,N_2} = \frac{\sum_{j_2=n_{2,max}} p(j_1,j_2)\cdot(N_2-j_2)}{\sum_{j\in E} p(j_1,j_2)\cdot(N_2-j_2)}$$

If there are two priority classes, but one of them (for instance, the lower priority class) is composed of several types of traffic (i.e. Web, Wap, etc.) we can still use this model to obtain approximate results. We summarize the different types of traffic by choosing an equivalent arrival rate and flow size for the aggregated class. The arrival rate of the aggregated traffic will be given by:

$$\lambda_{eq} = \sum_{i=1}^{k} \frac{N_i}{N}\lambda_i$$

where N_i is the number of users of each type within the class, N the total number in the class and λ_i the arrival rate of type i flows.

Meanwhile, the mean flow size in the aggregated class will be given by:

$$E(\sigma_{eq}) = \frac{1}{\sum_{j=1}^{k} N_j\lambda_j} \sum_{i=1}^{k} N_i\lambda_i E(\sigma_i)$$

In this case, the weights in the sum correspond to the probabilities that a flow comes from type i.

5. MEASUREMENT METHODOLOGIES AND TRAFFIC ANALYSIS IN THE GSM/ GPRS/ EDGE NETWORK

The performance evaluation raised by the previous models assumes a deep knowledge of voice and data traffic characteristics. This imposes the need of accurate measurement methodologies that allow both to apply these models to real situations and to verify their validity.

An important issue to consider in performance evaluation over data networks is the problem of "time-scales". Different factors have different impact depending on the considered time scale: packets, flows, sessions, etc. In this sense, measurement methodologies must clearly define the relevant time-scale to be used. Traditional measurements conducted by network operators use "long time-scales" (even larger than sessions, e.g. 1 hour), allowing to catch only average users' behaviour. While these are useful for general evaluation, smaller time-scales may help evidence particular problems. The data traffic model introduced in section 3 proposes a flow level analysis. This analysis was carried out at the IP level, sniffing operator's data traffic at the *Gi interface.* [3] Traditional methods were applied in the case of voice traffic.

5.1 Data traffic analysis

As stated in section 3, the data traffic model proposes a differentiation between activity (ON periods) and inactivity periods (OFF periods) along a user's session. A user's session is identified by the IP address assigned by the cellular system. Each session is compossed of different flows of traffic, each of them identified by the traditional 5-tuple

[3]IP connection interface to external networks, e.g. Internet

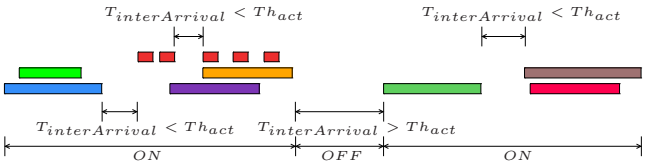

Figure 7: Data traffic of a user's session

(IP origin and destination, port origin and destination and protocol). An ON period consists of all agregated flows with inter-packet-arrival time ($T_{interArrival}$) smaller than certain threshold Th_{act}. Figure 7 shows the data traffic of a standard user's session.

The times elapsed between activity periods are defined as the OFF periods (and times between flows of the same ON period will be called mini-OFF periods). Given the ON and OFF periods, the type of traffic is identify by the average data size transmitted at the ON period ($E(\sigma)$) and the average duration of the OFF period ($E(\tau)$).

The ON/OFF period identification was achieved by packet inter-arrival time inspection. For eack type of traffic, this procedure consists in finding the threshold Th_{act} that separates contiguous activity periods. More precisly, considering two contiguous packets i and $i+1$:

- if $T_{interArrival} < Th_{act}$, both packets $i+1$ and i belong to same activity period
- if $T_{interArrival} > Th_{act}$, packet i belongs to one activity period and packet $i+1$ to the next one.

This technique was adjusted in a test radio base station, working under a controlled traffic situation to clearly identify the Th_{act} threshold for each type of traffic. Measurements carried out at this test bed were also used to verify the relevance of the proposed GPRS/EDGE model. The evaluation was done at throughput level, comparing model's results (Th_{model}) against two other estimations: throughput from the operator system's registers (Th_{cont}) and throughput obtained from the traffic capture (Th_{cap}). The latter is computed as the average ratio between transmitted bytes and duration of all ON periods within a user's session:

$$Th_{capture} = \frac{1}{N}\sum_{i=1}^{N} \frac{B_{ON}^i}{T_{ON}^i}$$

where B_{ON}^i and T_{ON}^i are the transmitted bytes and duration of the i-nth ON period and N the number of activity periods identified.

Table 1 presents the results obtained for different types of traffic, number of mobiles and cell configuration:

T_{cap} tends to underestimate the value of throughput because of the mini-OFF periods. On the other hand, T_{cont} considers average values for 15 minutes' periods along with some particular computation techniques which bias the estimation. Despite these differeneces, T_{model} values are consistent with other estimations, showing the relevance of the proposed model.

Values obtained from the test bed experience where applied to real traffic captures within the operator's network. Figure 8 presents the Th_{model} per user of WAP traffic, considering a mixture of WEB (25%) and WAP (75%) users

Traffic	TS	N	T_{cont}	T_{cap}	T_{model}
PTT (down)	2 CS2	5	6.0	4.5	6.9
PTT (down)	2 CS2	4	6.0	5.7	7.4
MMS (up)	1 CS2	6	11.4	7.2	7.0
MMS (up)	1 EDGE	6	40.4	46.0	39.0
WAP (down)	1 CS2	6	6.0	6.3	6.5
WAP (down)	2 CS2	6	13.5	12.2	18.4

Table 1: Throughput estimation - test bed experience. N is the number of mobiles in the experiment

within the cell and ficticious cell configuration with 4 CS-2 data dedicated time slots, $T_{TBF} = 3$ and .... These results can be used by the operator in different situations, from performance evaluation of different types of services to cell dimensioning. In the example, a cell of these characeristics may tolerate up to 20 users with a reazonable experienced quality (assuming that a WAP service should have at least 15 *kbps* for good performance).

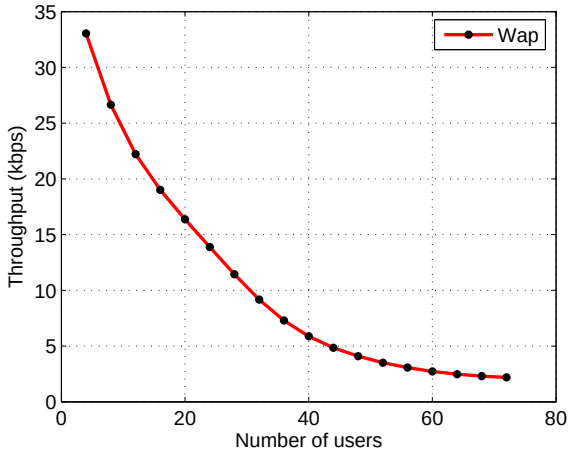

Figure 8: Per user throughput for WAP traffic

An interesting feature of data traffic arised from the ON/OFF modelling: *traffic utilization rate* (ratio between $E(\sigma)$ and $E(\tau)$) of certain types of traffic (e.g. WEB, WAP) present a huge variance. At first glance this observation is not surprising, as different services within the same type of traffic have different characteristics (i.e. mail and FTP applications, both classified as WEB type of traffic). However, the influence of this variation over performance is significant.

Section 6 studies the impact of traffic utilization rate, considering two different users' profiles: an *average user* (average rate) and a *high utilization user* (average rate within the biggest rates of the traffic).

5.2 Voice traffic analysis

The proposed GSM model was validated by direct comparison of full-rate and hafe-rate traffic intensity estimation against real per-cell slots' utilization registered by the operator. Considering average arrival and service rates and thresholds' configuration (HRTh and PACKTh) provided by the network operator, full-rate and half-rate traffic intensity was computed for three different cells. Figure 9 presents the results obtained for a 24hs period analysis over each cell (values have been anonymized). The model correctly estimates both full-rate and half-rate traffic intensities. The biggest differences occur under heavy load conditions, where callbacks' augmentation drift away the Poissonian arrivals' hypothesis. Even so, results show the model adjusts network's reality.

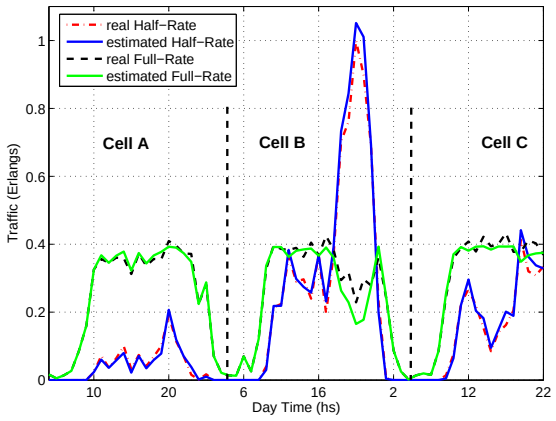

Figure 9: Half-rate and full-rate traffic intensity estimation

6. PERFORMANCE AND QUALITY OF SERVICE ISSUES

This section analyzes the cell performance, when some cell configuration parameters and some properties of the traffic are modified.

6.1 Performance analysis of some GSM thresholds

The voice call assignment to time slots was discussed in section 3. In this section, the impact on the cell performance of the half rate threshold HRTh and the calls packing threshold (PACKTh) is analyzed.

6.1.1 Half rate threshold (HRTh)

The following assumptions have been done in order to analyze the HRTh threshold influence:

- The total traffic offered to the cell in Erlangs is fixed to one arbitrary value .
- The calls packing feature is disabled (PACKTh is disabled).

Figure 10 shows the full rate traffic and the call blocking probability against the threshold HRTh.

In this case, if it is necessary to have a call blocking probability lower than 2% (a typical design target) the full rate

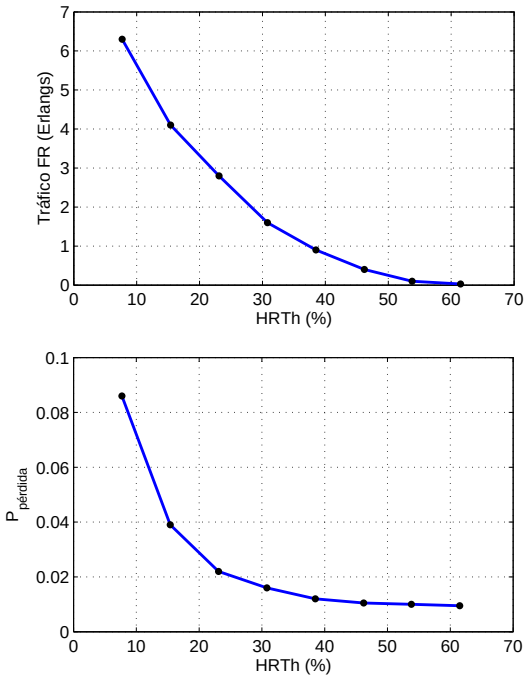

Figure 10: Full rate traffic and call blocking probability

traffic will be always below 12% of the total traffic offered to the cell..

Figure 11 shows the number of free slots that can be used for GPRS traffic. As it can be seen, when the threshold HRTh changes from 10 to 50 % the number of free slots is doubled.

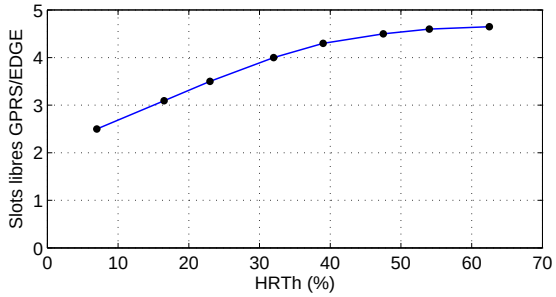

Figure 11: Free slots for GPRS traffic vs. HRTh.

6.1.2 *Packing threshold (PACKTh)*

In this section, the threshold HRTh is fixed in a typical value (35%) and the performance is analyzed under variation of the packing threshold (PACKTh).

When the threshold PACKTh is increased, the performance is modified in two ways:

- First, there are more free slots for GPRS traffic because the call packing feature frees slots.
- Second, there is an increase in the quality of service because the increase in the number of free slots makes the system come back more quickly to assign full rate calls.

The increase explained in the first point is decreased by the effect of the second point (more calls are assigned full rate). Both effects results then in a quality of service improvement and a small gain in free slots for GPRS traffic.

At last the increase in the PACKTh value generates a small increase in the call blocking probability. Figure 12 shows this effect.

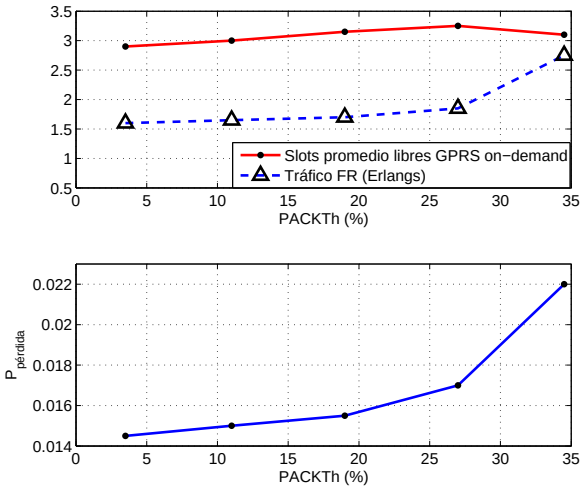

Figure 12: Full rate traffic, free slots and blocking probability.

6.2 User profile influence on cell performance

In section 5.1 it has been explained that different users (of the same type of connection) use the network in different ways. The utilization rate has strong variations for different users profile.

For example, in the case of WAP connections, there are some users that only chat and others that mainly use the network to download music, videos, etc. The analysis can be made with the mean user anyway, and the results will be very good. However, the actual performance will not be as good. This means that in the case where there are big variations in user profiles the mean user is not a good representation.

Figure 13 shows the throughput per user (for WAP users) for average users (top) and mix of both profiles (bottom). This figure clearly shows the strong influence of users' profiles on cell dimensioning.

6.3 Performance evaluation under traffic priorization policies

In section 4 the traffic priorization model was described. In this section a toy example to illustrate how the performance can be improved for a certain priority traffic is given. Two types of traffic are considered, WEB and WAP. The percentage of each type of traffic is constant and the total number of users is changed.

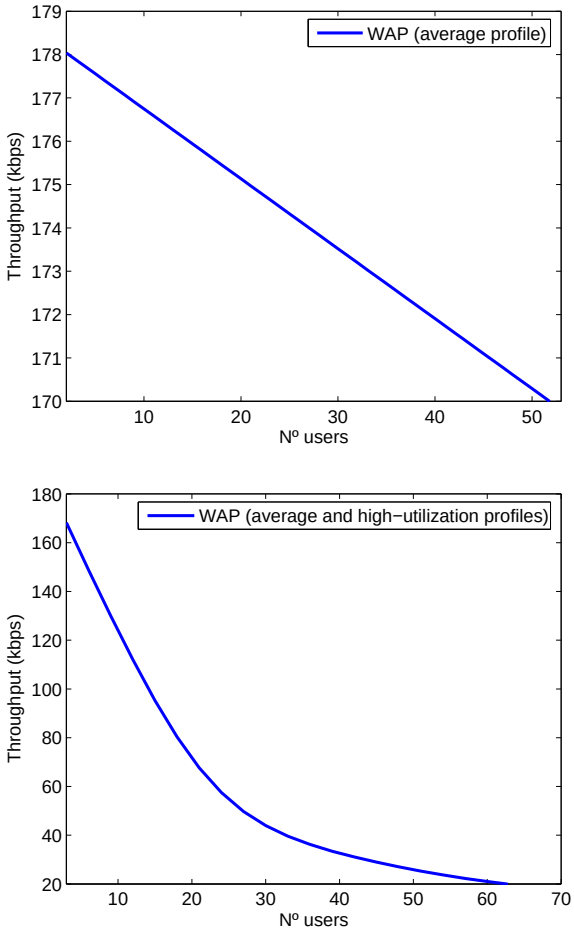

Figure 13: Throughput per user in different mix of users profiles

- 50% of WAP traffic and 50% of WEB traffic
- 2 radios
- 4 slots with EDGE MCS-9
- 10 slots PDCH CS-2
- $T_{TBF} = 2$
- 3 slots for download in each mobile.
- voice traffic is not considered.

Figure 14 shows the throughput per user of each type of traffic. As can be seen the prioritized traffic obtains more throughput than the unprioritized one. WAP traffic use the system as if the WEB one was not in the system. The WEB traffic use less time the system resources and so it has less throughput.

Figure 15 shows the blocking probability for the example described before. In the figure can be observed that the blocking probability for the prioritized traffic is always near zero. However, for the WEB traffic the blocking probability goes to one very rapidly from a certain number of users.

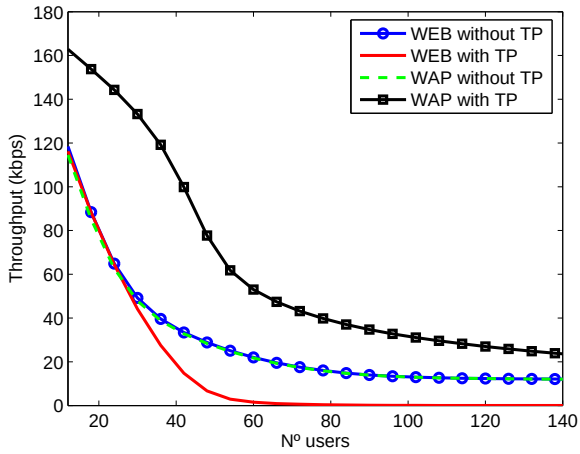

Figure 14: Throughput variation for the model with and without traffic priorization.

This behavior is typical for prioritization systems. Naturally, the increase in the WEB blocking probability happens for a lower number of users in the prioritized case than in the none prioritized one.

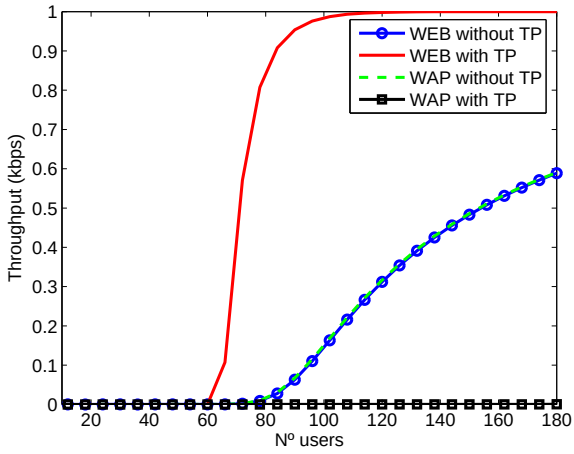

Figure 15: Blocking probability for WAP and WEB traffic in the model with and without prioritization.

It can be concluded that the deployment of a traffic prioritization system can result in a strong increase in the throughput of the prioritized traffic. This increase is achieved with prejudice to the throughput of the non prioritized traffic and some degradation in the call blocking probability. This type of models are recommended for cases where the flows of the traffic with priority have a very short duration and represent a small percentage of the total traffic in the system.

7. CONCLUSIONS

In this work, a model for GSM/GPRS/EDGE cell dimensioning and performance evaluation is presented. This model includes several relevant characteristics of network

equipment that have not been considered in previous studies, such as time-slot assignment strategies (both in GSM and GPRS/EDGE), different slot transmission rates, etc. Individual models for GSM and GPRS/EDGE are developed and their interaction is analysed. These models allow the evaluation of end-users' applications by means of average throughput and blocking probability computation. A generalization of previous model is also introduced, taking into account traffic priorization between different applications.

Presented models are validated against real data provided by a GSM/GPRS/EDGE network operator and by specially designed tests conducted over a test radio base station. Obtained results show the accuracy of the model and illustrate how it can be applied to analyze the impact of diverse design parameters on network performance.

The GSM/GPRS/EDGE performance evaluation problem is complex and many aspects still remain open for future study. More accurate measurement methodologies should be considered for model's parameters estimation. The development of others QoS models represents another possible direction.

8. REFERENCES

[1] 3rd Generation Partnership Project. http://www.3gpp.org.

[2] H. Dahmouni, B. Morin, and S. Vaton. "Performance Modelling of GSM/GPRS Cells with Different Radio Resource Allocation Strategies". In *IEEE Wireless Communications and Networking Conference*, 2005.

[3] H. Dahmouni, D. Rossé, B. Morin, and S. Vaton. "Analytical Model for preformance Evaluation of GPRS/EDGE Multi-Service Networks". In *7th. IFIP International Conference on Mobile and Wirelesss Communications Networks*, 2005.

[4] H. Dahmouni, D. Rossé, B. Morin, and S. Vaton. "Impact of data traffic composition on GPRS performances". In *19th International Teletraffic Congress*, 2005.

[5] C. H. Foh, B. Meini, B. Wydrowski, and M. Zuckerman. "Modeling and Performance Evaluation of GPRS". In *IEEE Vehicular Transport Conference*, 2001.

[6] L. Kleinrock. *"Queueing Systems, Volume 1: Theory"*. Wiley-Interscience, 1975.

[7] A. C. B. Kochem and E. L. Bodanese. "Providing QoS over GPRS: Admission Control, Radio Resource Reservation, and Scheduling". In *ITU & ITC Workshop for Developing Countries at the 18th International Teletraffic Congress (ITC-18)*, 2003.

[8] B. Li, L. Li, B. Li, and X.-R. Cao. "On Handoff Performance for an Integrated Voz/Data Cellular System.". *Wireless Networks*, 9(4):393–402, 2003.

[9] C. Lindemann and A. Thümmler. "Performance Analysis of the General Packet Radio Service". *Computer Networks*, 41(1):1–17, 2003.

[10] J. Roberts and L. Massoulié. "Bandwidth Sharing and Admission Control for Elastic Traffic". In *ITC Specialist Seminar*, 1998.