A Markovian Model for Mobile Cellular Networks with QoS Differentiation

Georges Nogueira
Université Pierre et Marie Curie, Paris6
Paris, FRANCE
Email: georges.nogueira@lip6.fr

Bruno Baynat
Université Pierre et Marie Curie, Paris6
Paris, FRANCE
Email: bruno.baynat@lip6.fr

Ahmed Ziram
Nortel
Chateaufort, FRANCE
Email: aziram@nortel.com

***Abstract*— This paper presents a realistic and accurate analytical model to dimension mobile cellular networks with QoS differentiation. QoS per applicative flow is commonly defined in GPRS/EDGE or 3G systems where streaming applications with real time properties and elastic data applications have to share radio resources. The need for accurate and fast-computing tools is of primary importance to tackle complex and exhaustive dimensioning issues. In this paper, we present a generic QoS analytical model developed in the context of EDGE networks but that can be adapted to a different technology. We develop a Markovian model that takes into account the QoS differentiation between real time and non-real time classes and gives expressions for all the required performance parameters. We compare our model with simulation and show its accuracy.**

I. INTRODUCTION

Mobile devices such as cell phones, digital assistants or laptops are increasingly used to transfer data over cellular wireless networks such as GPRS/EDGE, 3G and 4G networks. These commercial wireless networks carry data traffic for a variety of applications such as multi-media messaging (MMS), web browsing, and advanced applications like video-streaming or push-to-talk. Since radio resource remains the critical resource, operators need to manage networks in a way that provides both a comfortable QoE (Quality of Experience) for subscribers and an efficient bandwidth usage. For these purposes, Packet Flow Context (PFC) has been normalized in 3GPP recommendation [2] providing QoS differentiation between real time and non-real time application flows and service continuity between EDGE and 3G/4G networks. In this context, it is of primary importance for network engineers to have a radio link dimensioning tool that allows predicting the impact of traffic growth within and between classes.

In this paper we study the radio link in GPRS/EDGE networks (now denoted (E)GPRS) with PFC differentiation. GPRS (General Packet Radio Service) is an overlay to GSM networks that allows end-to-end IP-based packet traffic and EDGE (Enhanced Data rates for Global Evolution) is its improvement allowing higher throughputs and integrating the QoS differentiation provided by PFC. This 2G+ wideband network is being deployed worldwide, and because of its low cost and the good performance achievable, it is a key coverage solution to provide nationwide wireless data services complementary to 3G/4G networks.

Many works study the problem of performance analysis in (E)GPRS networks either by simulation or analytical modeling. Both have advantages and drawbacks. On the one hand, the accuracy of simulation results is obtained at the expense of long processing time (prohibitive to be involved in a dimensioning optimization process) and the analysis of the performance results for dimensioning the system inputs is often a difficult task (see e.g. [5], [17], [20], and [21]). On the other hand, analytical modeling gives faster results and a better understanding of the intrinsic system behavior, but relies on strong and non-realistic assumptions. Many works propose analytical models assuming a single type of data traffic with classical circuit-based assumptions that are not adapted to wireless data networks [6], [8], [9], [12], [13]. Other works succeed in providing simple models, but still for systems with a single type of data traffic and without QoS differentiation [4], [15], [18]. Finally, some papers propose models integrating different traffic classes and/or QoS differentiation [10], [11], [22], but still do not integrate all the specificities of PFC mechanisms.

As a response, we need an analytical model that it is both accurate and realistic. We develop such an efficient Markovian model that provides performance parameters with a very good accuracy as validated by simulation. It makes the tool perfectly suitable for the expected dimensioning issues as it avoids the use of time consuming simulations.

This paper is organized as follows. Section II presents the system description. In Section III, we first develop independent models for each class. We then develop an analytical model for the complete PFC system in Section III-C. Section IV finally presents validation results.

II. SYSTEM DESCRIPTION

A. System assumptions

We consider a single (E)GPRS cell submitted to data traffic with QoS differentiation. Our study tackles the analysis of the bottleneck, i.e. the radio link, and focuses on the downlink assumed to be the critical resource because of data traffic asymmetry. As a short reminder, (E)GPRS is a packet overlay on the circuit-based GSM system. With GSM, on each frequency carrier a 200 kHz bandwidth is shared between 8 users. Each user is given a circuit, also called time-slot because of Time-Division mutliplexing scheme (TDMA). With (E)GPRS, a mobile station can use several time-slots

simultaneously to have a higher throughput. Time-slots are shared between mobiles with a granularity of 20 ms (a so-called "radio block"). Every 20 ms the RRM (Radio Resource Manager) allocates the time-slots to the mobiles having an on-going transfer.

We make the following assumptions:

- The radio resource is divided into two independent parts according to the so-called "Complete Partitioning" policy [8], one dedicated to voice and one dedicated to data. Here, we only focus on the data part and assume that there is a fixed number T of time-slots dedicated to (E)GPRS traffic in the cell. Obviously the classical Erlang formulas apply for voice.
- All (E)GPRS mobiles have the same reception capability for data traffic. They are denoted "$(d+u)$" corresponding to the mobile multi-slot class, where d is the maximum number of time-slots that can be allocated for a mobile per TDMA frame in downlink, and u is the maximum number of time-slots that can be allocated in uplink. Note that, as we are interested in the modeling of the downlink traffic, only the parameter d is relevant for the model. Nowadays, most (E)GPRS mobiles are $(4+1)$ or $(4+2)$ and thus, can use at most $d = 4$ time-slots per TDMA frame in downlink.

Our system is characterized by the following parameters:

- t_B, the system elementary time interval equal to the radio block duration, i.e. $t_B = 20$ ms.
- x_B, the number of data bytes transferred over one time-slot. x_B/t_B is the throughput offered by the RLC/MAC layer to the above LLC (Logical Link Control) transport layer. The value of x_B depends on the radio modulation and coding scheme (MCS). For EDGE we have:

EDGE coding scheme	MCS1	MCS2	MCS3
x_B (in bytes)	22	28	37

MCS4	MCS5	MCS6	MCS7	MCS8	MCS9
44	56	74	112	136	148

  Note that MCS depends on the radio conditions and allows the mobile station to adapt its transmission rate to the radio link quality. In our model, we assume that all the mobiles use a given MCS that can correspond to an average radio link quality and that can also include non ideal radio conditions [4].
- tbf_{max}: the maximum number of mobiles that can simultaneously have an active downlink TBF (Temporary Block Flow) whatever the QoS class. TBF is the RLC/MAC transfer entity that is mandatory for a mobile to be active in the system. On a single TDMA, assuming uplink and downlink flows occur concurrently, (E)GPRS specifications give:

$$tbf_{max} = \min(32, 7T), \tag{1}$$

  because of the (E)GPRS system limitations on the signalling capabilities [4] (no more than 32 TFIs (Temporary Flow Identity) per TDMA and 7 USFs (Uplink State Flag))[1] per uplink time-slot).

B. QoS specifications

We study a system that provides QoS differentiation of data flows with mobiles running applications with different traffic characteristics. This service differentiation aims at providing QoS that matches the application need and an efficient use of the radio resources. Each data flow is associated with a QoS profile corresponding to its application type. This QoS description is known as Packet Flow Context (PFC) in (E)GPRS systems (see [1] and [2]). PFC both provides a differentiated QoS on the radio side for pure EDGE subscribers and service continuity to 3G/4G networks. PFC differentiation covers two main functionalities for the QoS management: connection admission control and resource allocation for the active connections.

The definition of PFC classes is split into two sets: some mobiles perform interactive applications having elastic traffic properties (e.g. email, web, FTP). This set of application is managed as best effort traffic and thus, the downloading durations of data elements depends on the system load. Data flows associated with these applications with no real time constraint are denoted as NRT connections.

The other set of mobiles performs streaming applications with real time constraints (e.g. audio/video streaming, conversational applications). We assume that session durations of these applications do not depend on the quantity of resource they received and thus, are independent from the system load (for instance, a video-streaming session duration only depends on its content characteristics). Data flows associated with these applications are denoted as RT (Real-Time) connections. Because of live streaming necessities, these applications require a guaranteed throughput to maintain real time QoE. This guaranteed throughput is performed by reserving GBR (Guaranteed Bit Rate) time-slots per TDMA frame for each RT connection (GBR is not necessarily an integer value, as it corresponds to an average value over time). If this reservation is not possible (in lack of available resources) the new RT connection can be degraded and managed by the system as NRT in a best effort way. We denote by RTd these degraded connections that can be upgraded back to RT connections if available resource is freed.

Note that, as described below, the RRM will first allocate a maximum of MBR time-slots per unit of time for each NRT (or RTd) connections to limit their impact on RT performance. Finally, in case of extra resource, it is equally divided among all ongoing connections.

Finally, a part of the TDMA frame can be reserved for NRT and RTd connections in order to ensure a minimum throughput guaranty to low priority mobiles. We denote Min_{NRT} this number of dedicated time-slots per TDMA frame. (Note again that Min_{NRT} is not necessarily an integer.

[1]TFI is the TBF identifier coded in 5 bits; USF defines the number of mobiles that can be multiplexed on one time-slots.

1) Admission control: The RRM decides to admit a new connection in relation to its PFC class and the available remaining resources. If the system limit defined by relation (1) is not reached, any connection demand is proceeded in the following priority order:

- For RT traffic, a connection demand is accepted as RT connection if the guaranteed bit rate GBR can be met for each RT connection. RT connections have a preemptive priority over NRT connections. Thus, the admission of a new RT connection can result in the preemption of one or several NRT connections in lack of remaining resources. If the guaranteed bit rate cannot be met for the new RT demand, it is degraded and admitted as RTd connection. RTd connections have no longer throughput guarantees and are managed as NRT connections without any priority. As soon as a RT connection ends, the RRM randomly selects any RTd connection and upgrades it as RT for the rest of its session duration;
- For NRT traffic, connection demands are accepted withtout any minimum throughput requirement.

2) Resource allocation: Every time-step t_B, the RRM allocates the resources. It first fulfills the GBR requirements for each RT connections. RT connections thus obtain their corresponding guaranteed bitrate GBR, even in congestion, thanks to the reserved time-slots at admission control step. Next, NRT and RTd connections fairly share the remaining resources left by RT connections up to the maximum bit rate MBR defined in the NRT and RTd PFC profile. In case of extra resource, it is equally allocated between RT, RTd and NRT connections up to the maximum download capacity d.

C. Resource management

In order to model the system, we need to define the systems acceptation limits for each PFC profile depending on the current state of the system, i.e. the number of concurrent connections for each class at a given time-step. We denote by n_{RT} (resp. n_{RTd} and n_{NRT}) the number of RT connections (resp. RTd and NRT connections) at a given time. n_{max}^{RT} is the maximum number of RT connection (as RT connections have a preemptive priority over NRT connections, it corresponds to the maximum number of GBR units fitting into T time-slots), $n_{max}^{RTd}(n_{NRT})$ is the maximum number of RTd connections (that only depends on the number of competing NRT connections, as RTd connections only exist when $n_{RT} = n_{max}^{RT}$), and $n_{max}^{NRT}(n_{RT}, n_{RTd})$ is the maximum number of NRT connections (that depends both on the number of RT connections determining the remaining resources and the number of competing RTd connections sharing the remaining resources). As $n_{RTd} = 0$ if $n_{RT} \neq n_{max}^{RT}$, the maximum number of NRT connections is only related to the sum $n_{RT} + n_{RTd}$ and will thus be simply denoted as $n_{max}^{NRT}(n_{RT} + n_{RTd})$. The detailed expressions of these limits are given in Appendix A.

III. Analytical modeling

As a first step in the modeling of the whole system, we first consider each population independently. It is obvious that there is a strong dependence between the system parameters implying a strong correlation between the performance of the PFC classes. As a first step towards the modeling of the complete system, we first present the traffic and system asumptions independently for each class, and developp dedicated single-class models.

A. Real-Time class

We first consider a system only containing RT mobiles that generate a streaming traffic.

1) Real-Time characteristics: As discussed in Section II-B, a new connection demand is admitted as a RT connection if the available resource is sufficient to guarantee GBR time-slots per TDMA to the mobile (in addition to the remaining on-going RT connections). If the remaining resource is less than GBR time-slots per time-step, a new demand is degraded and admitted as a RTd connection with no throughput guarantees (provided the system signaling limit tbf_{max} is not reached, in which case it is rejected). As the number of active connections increases, the throughput of RT connections remains over the GBR time-slots per t_B and the throughput of RTd connections decreases (because of the sharing of the remaining resources). We assume that users of RTd connections can tolerate the quality degradation and do not stop prematurely their streaming, even if the throughput they obtain is less than what they expect. We also assume that the duration of the streaming is not affected by the degradation. As a consequence, the streaming duration has the same characteristics and can be modeled in the same way for both RT and RTd connections (as well as for RTd connections that are potentially selected by the RRM to be permanently upgraded to RT). Finally, it is important to emphasize that the real time nature of the traffic implies that the duration of any accepted connection (RT or RTd) is independent of the system load.

2) Real-Time traffic modeling: RT streaming traffic is modeled as follows. We assume that there is a fixed number N_{RT} of RT mobiles in the system. Each of them is supposed to generate an infinite length session of ON/OFF traffic. ON periods correspond to the streaming activity proceeded through RT or RTd connections. ON period durations are supposed to be exponentially distributed with a rate μ_{RT} equal to the inverse of the average streaming duration t_{on}^{RT}. OFF periods correspond to the inactive period between two streaming sessions. The OFF period durations are supposed to be exponentially distributed with a rate λ_{RT} equal to the inverse of the average inactive period duration t_{off}^{RT}. These traffic assumptions (finite population, infinite length sessions and memoryless distributions) are discussed in [4], [14], [16].

3) Real-Time model: The system is modeled by a linear Continuous-Time Markov chain where a state i corresponds to the total number of concurrent connections (RT and RTd). Let us recall that n_{max}^{RT} is the maximum number of simultaneous RT connections that can be admitted in the system. As we assume here that there are no NRT mobiles in the system, the maximum number of simultaneous active RT and RTd connections, denoted by n_{max}^{RT+RTd}, is given by $n_{max}^{RT+RTd} =$

$n_{max}^{RT} + n_{max}^{RTd}(0)$ (see Appendix A). Finally, as the admission of RTd connections only occurs after state n_{max}^{RT}, we can easily express the number of RT and RTd connections at state i as follows: $n_{RT} = \min(i, n_{max}^{RT})$ and $n_{RTd} = \max(0, i - n_{max}^{RT})$.

A transition out of a generic state i to a state $i+1$ (for $0 \leq i \leq n_{max}^{RT+RTd} - 1$) occurs when a new streaming request is accepted. This transition is performed with a rate $(N_{RT} - i)\lambda_{RT}$, corresponding to the arrival of one RT mobile among the $(N_{RT} - i)$ in OFF period. Note that we do not need to pay a particular attention to states before or after n_{max}^{RT}, as any connection demand arriving when the limit n_{max}^{RT+RTd} is not reached, is accepted as a RT before n_{max}^{RT} and as a RTd connection after n_{max}^{RT}.

A transition out of a generic state i to a state $i-1$ (for $1 \leq i \leq n_{max}^{RT+RTd}$) occurs when a streaming connection ends. For $i < n_{max}^{RT}$ it simply corresponds to the end of a RT connection (among the i active RT connections). For $i \geq n_{max}^{RT}$ it corresponds either to the end of one of the n_{max}^{RT} RT connections or to the end of one of the $i - n_{max}^{RT}$ RTd connections. In the former case, the RRM randomly selects a RTd connection to be instantaneously upgraded, thus recovering its throughput guaranty. Finally, as we assume that the streaming duration does not depend on the type of connection (RT or RTd) and is exponentially distributed, in both cases the transition from state i to $i-1$ is performed with a rate $i\,\mu_{RT}$. Note that we can easily account for a more general case where degradation affects the streaming duration by adjusting the departure rates from any state $i > n_{max}^{RT}$ in an appropriate way (e.g. by only decreasing the departures rates of the $n_{max}^{RT} - i$ RTd connections and thus modifying accordingly the global departure rate from state i).

The Markovian model is thus a birth and death process that turns out to be equivalent to the Engset model for circuit switched systems [7]. The steady-state probabilities $p_{RT}(i)$ of having i simultaneous active (RT or RTd) connections are thus given by:

$$p_{RT}(i) = \frac{\rho_{RT}^{i}\, N_{RT}!}{(N_{RT} - i)!} p_{RT}(0) \quad \text{for } 0 \leq i \leq n_{max}^{RT+RTd}, \quad (2)$$

where ρ_{RT} is given by $\rho_{RT} = \lambda_{RT}/\mu_{RT}$ and $p_{RT}(0)$ is obtained by normalization.

B. Non-Real-Time class

We now consider a system only containing NRT mobiles that generate an elastic traffic.

1) Non-Real-Time characteristics: As detailed in Section II, NRT connection demands will be systematically accepted as long as the signaling limit tbf_{max} is not reached. As a NRT connection corresponds to the download of a data element, its duration depends on the available resource. Thus, as opposed to RT connections, the duration of NRT connections depends on the system load. As a consequence, we characterize a NRT connection by a size (in bytes), as opposed to a RT connection that is characterized by a time (in seconds).

2) Non-Real-Time traffic modeling: NRT traffic is modeled as follows. We assume that there is a fixed number N_{NRT} of NRT mobiles that are sharing the total bandwidth of the cell (as we assume that there are no RT mobiles). Each of them is doing an ON/OFF traffic. ON periods correspond to the download of an element through NRT connections. The size of downloaded elements are supposed to be exponentially distributed with a mean of x_{on}^{NRT} bytes. Then we define the avarage data rate per time-slot as $\mu_{NRT} = x_B/(x_{on}^{NRT} t_B)$. OFF periods correspond to the reading time between two downloading ON periods. OFF period durations are supposed to be exponentially distributed with a rate λ_{NRT} equal to the inverse of the mean OFF period duration t_{off}^{NRT}.

3) Non-Real-Time model: The system is modeled by a linear Continuous-Time Markov chain where a state j corresponds to the total number of concurrent NRT connections, limited to a maximum given by $n_{max}^{NRT} = n_{max}^{NRT}(0)$ (see Appendix A).

A transition out of a generic state j to a state $j+1$ (for $0 \leq j \leq n_{max}^{NRT} - 1$) occurs when a new data download request is accepted, i.e. when a new NRT connection is accepted. This transition is performed with a rate $(N_{NRT} - j)\lambda_{NRT}$, corresponding to the arrival of one NRT mobile among the $(N_{NRT} - j)$ in OFF period. Note that a blocking event can occur if a new downloading request arrives when the system is in the state n_{max}^{NRT}.

The transition out of a generic state j to a state $j-1$ occurs when an active data download ends. This transition is performed with a rate $j\,\mu_{NRT}$, corresponding to the departure rate of one active NRT connection among the j active ones. When the system is in state j, the mobiles can use up to T time-slots for data transmission. Now, because of the maximum downloading capacity d, if $jd < T$, each mobile only receives a maximum of d time-slots per time-step. Thus, the available bandwidth is not fully utilized by NRT connections. A transition rate from state j to state $j-1$ corresponds to one of the j NRT connection that completes its transfer. This transition rate is then equal to $jd\mu_{NRT}$. On the other hand, if $jd \geq T$, the allocator has to share the T time-slots among the j NRT connections, and the transition rate from state j to state $j-1$ is then equal to $T\mu_{NRT}$. The generic transition rate from a state j to a state $j-1$ (for $1 \leq j \leq n_{max}^{NRT}$) is thus: $\min(j\,d, T)\mu_{NRT}$.

This NRT model gives equivalent performance parameters as the "Erlang-like" model described in [4], even if this last has been developed from Discrete-Time Markov chains. This model has also been described in [15] as part of a more complete (E)GPRS system with a Partial Partitioning scheme between voice and data (but without any QoS differentiation between data users). The steady state probabilities $p_{NRT}(j)$ can be derived easily from the birth and death structure of the Markov chain as follows:
for $0 < j \leq j_0$:

$$p_{NRT}(j) = \frac{N_{NRT}!}{j!d^j(N_{NRT} - j)!} x_{NRT}^{j} p_{NRT}(0), \quad (3)$$

for $j_0 < j \leq n_{max}^{NRT}$:

$$p_{NRT}(j) = \frac{N_{NRT}!}{j_0! d^{j_0} T^{j-j_0} (N_{NRT}-j)!} x_{NRT}^{j} p_{NRT}(0), \quad (4)$$

where $j_0 = \lfloor T/d \rfloor$ is the maximum value of j such that $jd < T$, x_{NRT} is given by $x_{NRT} = \lambda_{NRT}/\mu_{NRT}$, and $p_{NRT}(0)$ is obtained by normalization.

C. Complete system modeling

In order to model the complete system with both RT and NRT populations, we combine the two single class models presented in Section III into a multidimensional Markov chain. A first approach would be to develop a 3-dimensional Continuous-Time Markovian model where each dimension is associated to one of the three connection types (RT, RTd and NRT). But, as described in Section III-A, there are RTd connections in the system only when the limit n_{max}^{RT} for the number of RT active connections is reached. As a consequence, the 3-dimensional Markov chain is made of two orthogonal planes, as illustrated on the left part of Fig. 1, and can equivalently be transformed into a 2-dimensional Markov chain given in right part of Fig. 1. This model combines one vertical dimension for both RT and RTd, and one horizontal dimension for NRT. By now, each state of the chain is a couple (i, j) where i is the number of RT and RTd active connections, and j is the number of NRT active connections. Because of the limiting conditions on the number of active mobiles of each type (see Appendix A), each state (i, j) is such that:

$$0 \leq \quad i \quad \leq n_{max}^{RT} + n_{max}^{RTd}(j), \quad (5)$$
$$0 \leq \quad j \quad \leq n_{max}^{NRT}(i). \quad (6)$$

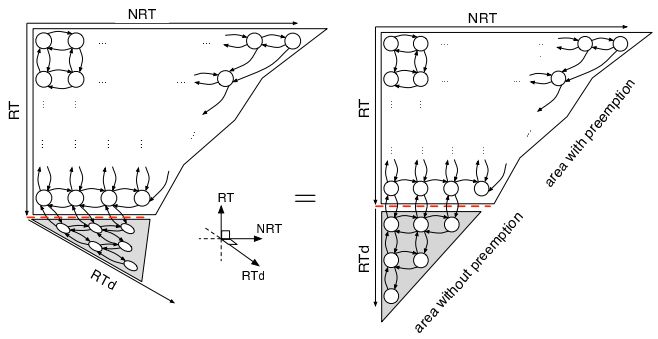


Fig. 1. 3-dimensional to bidimensional CTMC.

1) Combined resource sharing: We first do not take into account the maximum MBR constraint on NRT and RTd connections defined in Section II, and consider that NRT and RTd connections can use all the available resources left by RT connections up to their maximum download capacity d. This additional constraint and the modifications it involves are described in the next subsection (extra resource sharing correction).

Now, by combining the two single-class models, we have to pay a particular attention to the proportion of the resource available for NRT connections, as their performance strongly depend on it. This part of the resource actually depends on the number of concurrent RT and RTd connections. First, let us denote by $T_{RTd+NRT}(i)$ the remaining resource after the RRM allocates GBR time-slots for RT active connections:

$$T_{RTd+NRT}(i) = T - \min(i, n_{max}^{RT}) GBR. \quad (7)$$

Then, the RRM equally divides the $T_{RTd+NRT}(i)$ time-slots between both the j NRT and the k RTd connections (where $k = \max(0, i - n_{max}^{RT})$). Consequently, the proportion $T_{NRT}(i, j)$ of the remaining resource used by the j NRT connections can be expressed as:

$$T_{NRT}(i, j) = \frac{j}{k+j} T_{RTd+NRT}(i). \quad (8)$$

As the departure rate of NRT connections depends on the actual available resource $T_{NRT}(i, j)$, the horizontal transition out of a generic state (i, j) to state $(i, j-1)$ is now: $\min\left(j\, d, T_{NRT}(i, j)\right) \mu_{NRT}$. As explained in the NRT single class model of Section III-B, a transition rate from state (i, j) to state $(i, j+1)$ is $(N_{NRT} - j)\lambda_{NRT}$. In the same way, vertical transitions from states (i, j) to state $(i-1, j)$ and $(i+1, j)$ have the same expressions as the ones given in Section III-A for the single-class RT model, since the resource available does not impact the performance of RT or RTd connections. The resulting Markov chain is illustrated in Fig. 2 showing transitions on a generic state (i, j).

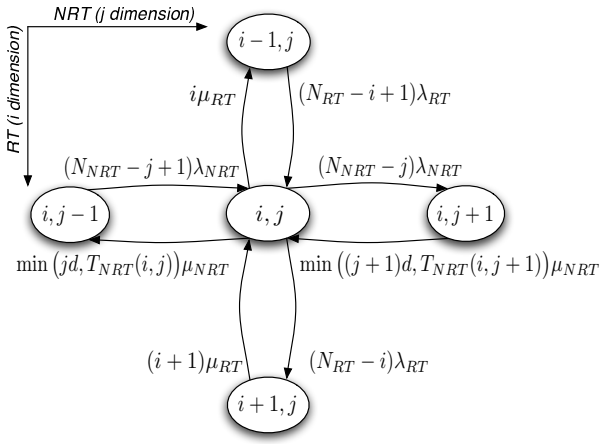


Fig. 2. Transitions on a generic state (i, j).

Let us now recall that RT connections have a preemptive priority over NRT connections and let us first consider a state (i, j) such that $i < n_{max}^{RT}$. A new RT connection demand will always be accepted as RT connection, and because of the preemptive priority, its admission can result in rejecting on-going NRT connections, if the remaining resource left after the admission is no longer sufficient to proceed all of them. As illustrated in Fig. 3, rejections occur if a new RT connection demand arrives when the system is in one of the limiting states $\left(i, n_{max}^{NRT}(i+1) + 1\right)$ to $\left(i, n_{max}^{NRT}(i)\right)$. Thus, a single RT connection admission can reject several NRT connections

at a time (up to $n_{max}^{NRT}(i) - n_{max}^{NRT}(i+1)$).

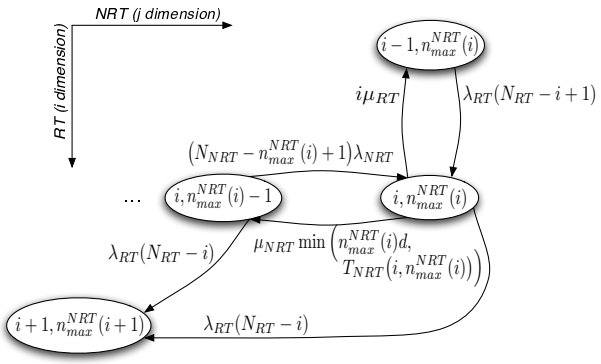


Fig. 3. Generic NRT preemption transitions.

Now, If we consider a state (i, j) such that $i \geq n_{max}^{RT}$, a new RT connection demand will only be accepted as a RTd connection if $i < n_{max}^{RT} + n_{max}^{RTd}(j)$, as we consider that RTd connections have no priority over NRT connections. As a consequence, in the 2-dimensional Markov chain, there are no diagonal transitions when $i \geq n_{max}^{RT}$.

The steady-state probabilities $p(i, j)$ of this 2-dimensional Continuous-Time Markov chain can be obtained using any numerical technique (see [19] for a list of possible methods). Then, all the performance parameters can be derived easily from the steady-state probabilities (see Appendix B for detailed expressions).

2) Extra resource sharing correction: We now take into account the maximum bit rate MBR constraint on NRT and RTd connections defined in Section II. As described in Section II-C, the last step of the resource allocation algorithm is designed to take into account the possible redistribution of extra resource. In other words, if each RT connection gets its guaranteed bit rate GBR and each RTd or NRT connection obtains its maximum bit rate MBR, and there remains available resources, these resources are equally redistributed between all (RT, RTd and NRT) connections.

Once again, the duration of any RT or RTd connection is assumed to be independent of the quantity of resources given to mobiles during their steaming activity. As a consequence, departure rates of the vertical RT model are not affected by the extra resource. In the case of NRT connections, higher throughputs allow connections to end up earlier their download. As a result, extra resource has an impact on the departures rates of the horizontal NRT models.

Let us denote by $\tilde{T}(i, j)$ the available extra resource left after allocating GBR time-slots per time-step to each RT connection, and by limiting each RTd and NRT connection to MBR time-slots per time-step:

$$\begin{aligned} \tilde{T}(i,j) &= T - \max(i, n_{max}^{RT})GBR \\ &\quad - \max(j, j+i-n_{max}^{RT})MBR. \end{aligned} \tag{9}$$

Now, let $\delta(i, j)$ be the portion of extra resource obtained by any single connection:

$$\delta(i,j) = \frac{\tilde{T}(i,j)}{i+j}. \tag{10}$$

In order to take into account the extra resource in the horizontal NRT models, we simply have to modify the transition rates from any state (i, j) to state $(i, j-1)$, for $0 < j \leq n_{max}^{NRT}(i)$, as follows:

$$\min\left(jd, j(MBR + \delta(i,j)), T_{NRT}(i)\right)\mu_{NRT}. \tag{11}$$

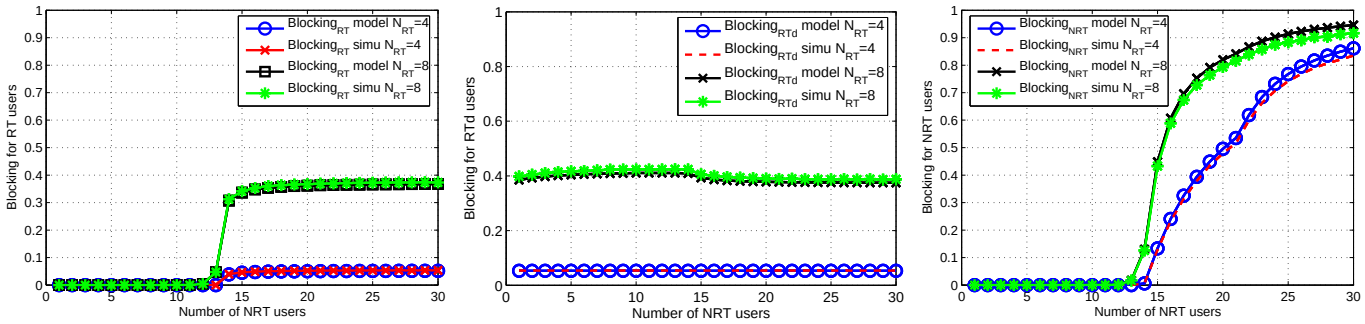
Note that any other reallocation policy can be used and would result in a straightforward modification of the rates given in relation (11).

IV. Validation and Performance Study

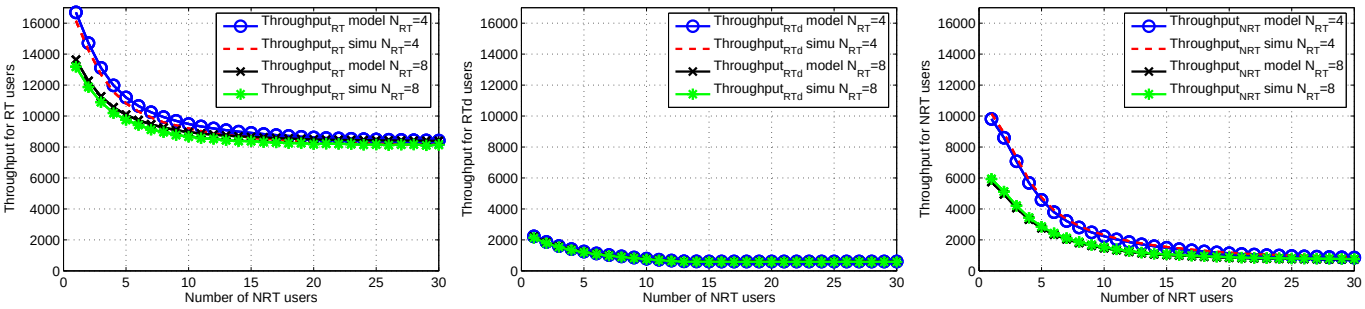
First, we validate the analytical model by comparison with simulation. Simulations have been performed with an home-made event driven simulator developed using the CNCL library [3]. This simulator is based on the same assumptions as our analytical model, i.e. infinite ON/OFF sessions and memoryless distributions for both ON and OFF. Nevertheless, the simulator captures the detailed behavior of the radio resource allocator and thus performs the exact PFC differentiation as described in Section II-C.

We consider several configurations of a cell containing a predefined population of RT mobiles $N_{RT} = \{4; 8\}$ and a varying number of NRT mobiles $N_{NRT} = [1; 30]$. All mobiles of a same PFC class generate the same traffic (see Table I for detailed parameters). We show in Fig. 4 the blocking probability P_r for RT and NRT mobiles, the degradation probability Pr_{RTd} for RTd connections, and the average throughput per user $\bar{X}$ for each connection type. Results derived from the analytical model and simulations are compared. As can be seen on the figures, the curves corresponding to the model and those corresponding to simulation are very close. The maximum relative error never exceeds 5% for any performance parameter in any configuration. In the case of detailed steady-state probabilities, we can find some higher errors, but always affecting data points with a very low contribution on the overall distribution and thus on average performance. The different configurations (corresponding to N_{NRT} varying from 1 to 30) have been obtained in several days with the simulation tool (on a standard single-core 2GHz processor) and in few minutes with the analytical model. We have performed many other experiments that gave similar results. As a conclusion to this validation, we can say that the model captures very precisely the behavior of the system.

We now provide a brief analysis of the obtained performance. Fig. 4(a) presents the blocking for RT mobiles, i.e. RT connection demands that cannot be admitted even as RTd because of lack of available resources. This curve shows that RT blocking is very sensitive to NRT population. In addition, we see in Fig. 4(b) that NRT population has a slight influence on degradation probability. This comes from the fact that the elastic behavior of NRT connections makes them stay longer in the system as the traffic load increases. Consequently, the



(a) Blocking for RT users. (b) Degradation of RT users. (c) Blocking for NRT users.



(d) Throughput for RT users. (e) Throughput for RTd users. (f) Throughput for NRT users.

Fig. 4. Validation of the analytical model through comparison with simulation.

TABLE I
PARAMETER VALUES FOR VALIDATION.

Parameter	Value	Description
T	6	Number of dedicated TS for (E)GPRS
d	4	Maximum number of TS used by a mobile per TDMA frame
x_B	112 bytes	Payload per radio block (MCS7)
GBR	1.5 TS	Guaranteed bit rate for RT connections
MBR	1 TS	Maximum bit rate for NRT connections
Min_{NRT}	0.15 TS	Reserved part of the TDMA for NRT and RTd users
N_{RT}	8	RT mobiles population
t_{on}^{RT}	180s	ON average duration for RT mobiles
t_{off}^{RT}	300s	OFF average duration for RT mobiles
x_{on}^{NRT}	8000 bytes	ON average page size for NRT mobiles
t_{off}^{NRT}	300s	OFF average duration for NRT mobiles

more NRT mobiles in the system, the more likely new RT connections demands that should be admitted as RTd are rejected. The curves in Fig. 4(d, e, f) show the average instantaneous throughput obtained by RT, RTd, and NRT connections. As expected, RT connections have a decreasing throughput with an asymptotic value equal to GBR time-slots per t_B seconds (here equal to 8400 bit/s). As the NRT population increases, the quantity of extra resource shared between all connections decreases and then, the RT throughput reaches its minimum guarantee (see Fig. 4(e)(f)). Unexpectedly, NRT connections reach higher throughputs than RTd connections. This can be explained by the fact that NRT connections stay longer in the system because of their elastic characteristics, whereas RTd connections only occur in congestion situations and thus receive in average a lower throughput. Note however that even if they have a lower instantaneous throughput, RTd connections can eventually be upgraded to RT connection and recover their guaranteed bit rate. Finally, both RTd and NRT throughputs reach an asymptotic minimum value corresponding to the maximum number of connections that can share the available resources.

V. CONCLUSION

We provide a realistic and accurate analytical model for cellular networks with QoS differentiation based on PFC mechanisms. We develop a Markovian model for multi-class traffic systems derived from simpler classical birth-and-death processes. Our model provides the expected efficiency and accuracy necessary for complex performance and dimensioning analyses. We are investigating application of this analytical modeling methodology to 3G systems. In addition, we have worked on advanced decomposition techniques allowing us

to derive even simpler models that provide closed-form expression for all the performance parameters [16]. Having a computationally fast module for packet-based traffic modeling is an essential asset for these advanced technologies.

APPENDIX

A. (E)GPRS system limitations

Let us remind that up to 7 TBF can be multiplexed on each time-slot and up to 32 TBF in the whole TDMA (see Section II-A).

RT connection limit:
As the RT connections have a preemptive priority over NRT connections, we can define n_{GBR}, the maximum number of GBR units fitting into the T time-slots:

$$n_{GBR} = \left\lfloor \frac{T - Min_{NRT}}{GBR} \right\rfloor, \tag{12}$$

and then, w.r.t. relation (1), n_{max}^{RT} the maximum number of RT connections is:

$$n_{max}^{RT} = \min(tbf_{max}, n_{GBR}, N_{RT}). \tag{13}$$

NRT connection limit:
We first assume that there are no RTd connections and n_{RT} RT connections. Let us denote by BW_{RT}, the resource used by these n_{RT} RT connections at a given time-step:

$$BW_{RT} = n_{RT}\, GBR, \tag{14}$$

and by T_{RT} the exact number of (E)GPRS time-slot used by RT connections as:

$$T_{RT} = \lfloor BW_{RT} \rfloor. \tag{15}$$

Let T_{NRT} be the corresponding number of time-slot entirely available for NRT connections:

$$T_{NRT} = T - \lceil BW_{RT} \rceil. \tag{16}$$

As the RRM allocates contiguous time-slots for RT connections, the last time-slot allocated for a RT connection may be also used by NRT connections. Then, we express n_{occ}^{RT}, the number of TFI identifiers used by RT connections in this shared time-slot as:

$$n_{occ}^{RT} = \left\lceil \frac{BW_{RT} - T_{RT}}{GBR} \right\rceil, \tag{17}$$

and we denote by n_{occ}^{NRT}, the available TFI identifiers for NRT connections in this shared time-slot:

$$n_{occ}^{NRT} = \max(0, 7 - n_{occ}^{RT}). \tag{18}$$

We can now define $n_{max}^{NRT}(n_{RT}, 0)$, the maximum number of NRT connections assuming n_{RT} RT connections and no RTd connection:

$$\begin{aligned} n_{max}^{NRT}(n_{RT}, 0) &= \min(N_{NRT}, 32 - n_{RT}, \\ &\quad 7\,T_{NRT} + n_{occ}^{NRT}). \end{aligned} \tag{19}$$

Finally, we can give the general expression $n_{max}^{NRT}(n_{RT}, n_{RTd})$ assuming n_{RT} RT connections and n_{RTd} concurrent RTd connections as:

$$\begin{aligned} n_{max}^{NRT}(n_{RT}, n_{RTd}) &= \min(N_{NRT}, 32 - n_{RT}, \\ &\quad 7\,T_{NRT} + n_{occ}^{NRT} - n_{RTd}). \end{aligned} \tag{20}$$

RTd connection limit:
As the RRM manages RTd connections as NRT, the resource occupation and the signaling limit are the same for both connection types. Nevertheless, there are RTd connections in the system only when $n_{RT} = n_{max}^{RT}$. Thus, $n_{max}^{RTd}(n_{NRT})$ the maximum number of RTd connections is given by:

$$\begin{aligned} n_{max}^{RTd}(n_{NRT}) &= \min(N_{RT} - n_{max}^{RT}, \\ &\quad 32 - n_{max}^{RT} - n_{NRT}, \\ &\quad 7\,T_{NRT} + n_{occ}^{NRT} - n_{NRT}). \end{aligned} \tag{21}$$

B. Performance parameters

RT blocking probability:
RT connection demands that cannot be admitted even as RTd (because of lack of available resource) are rejected. This event occurs with a rate λ_{rej}^{RT}:

$$\lambda_{rej}^{RT} = \sum_{i=n_{max}^{RT}}^{n_{max}^{RT+RTd}} p\big(i, n_{max}^{NRT}(i)\big)\,(N_{RT} - i)\lambda_{RT}. \tag{22}$$

The global arrival rate of RT connections is defined as:

$$\lambda_{all}^{RT} = \sum_{i=0}^{n_{max}^{RT+RTd}} \sum_{j=0}^{n_{max}^{NRT}(i)} p(i,j)(N_{RT} - i)\lambda_{RT}. \tag{23}$$

Thus, we can derive P_r^{RT}, the RT blocking probability as:

$$P_r^{RT} = \frac{\lambda_{rej}^{RT}}{\lambda_{all}^{RT}}. \tag{24}$$

RT degradation probability:
RT connection demands can also be admitted as RTd. This event occurs with a rate λ_{deg}^{RT}:

$$\lambda_{deg}^{RT} = \sum_{i=n_{max}^{RT}}^{n_{max}^{RT+RTd}-1} \sum_{j=0}^{n_{max}^{NRT}(i)-1} p(i,j)(N_{RT} - i)\,\lambda_{RT}. \tag{25}$$

Then, we can derive P_{deg}^{RT}, the RT degradation probability:

$$Pr_{RTd} = \frac{\lambda_{deg}^{RT}}{\lambda_{all}^{RT}} \tag{26}$$

NRT blocking probability:
NRT connection demands can be rejected in lack of available signaling identifiers. This event occurs with a rate λ_{rej}^{NRT}:

$$\lambda_{rej}^{NRT} = \sum_{i=0}^{n_{max}^{RT+RTd}} p\big(i, n_{max}^{NRT}(i)\big)\,\big(N_{NRT} - n_{max}^{NRT}(i)\big)\lambda_{NRT}. \tag{27}$$

Finally, RT connection admissions may result in NRT rejections. The number of such rejections per unit of time is given by λ_{pre}^{NRT}:

$$\lambda_{pre}^{NRT} = \sum_{i=0}^{n_{max}^{RT}-1} \sum_{j=n_{max}^{NRT}(i+1)+1}^{n_{max}^{NRT}(i)} p(i,j) .(j - n_{max}^{NRT}(i+1))(N_{RT} - i)\lambda_{RT}. \quad (28)$$

The global arrival rate of NRT connections is defined as:

$$\lambda_{all}^{NRT} = \sum_{i=0}^{n_{max}^{RT+RTd}} \sum_{j=0}^{n_{max}^{NRT}(i)} p(i,j)(N_{NRT} - j)\lambda_{NRT}. \quad (29)$$

Thus, we can derive P_r^{NRT}, the RT blocking probability:

$$P_r^{NRT} = \frac{\lambda_{rej}^{NRT} + \lambda_{pre}^{NRT}}{\lambda_{all}^{NRT}}. \quad (30)$$

RT average instantaneous throughput:

$$\bar{X}_{RT} = \sum_{i=1}^{n_{max}^{RT+RTd}} \sum_{j=0}^{n_{max}^{NRT}(i)} p(i,j)\frac{x_B}{t_B} . \max\big(GBR, \min(GBR + \delta(i,j), d)\big). \quad (31)$$

RTd average instantaneous throughput:

$$\bar{X}_{RTd} = \sum_{i=n_{max}^{RT}+1}^{n_{max}^{RT+RTd}} \sum_{j=0}^{n_{max}^{NRT}(i)} p(i,j)\frac{x_B}{t_B} . \min\big(\frac{T - n_{max}^{RT} GBR}{\max(0, i - n_{max}^{RT}) + j}, MBR + \delta(i,j), d\big). \quad (32)$$

NRT average instantaneous throughput:

$$\bar{X}_{NRT} = \sum_{i=0}^{n_{max}^{RT+RTd}} \sum_{j=1}^{n_{max}^{NRT}(i)} p(i,j)\frac{x_B}{t_B} . \min\big(\frac{T - \min(i, n_{max}^{RT}) GBR}{\max(0, i - n_{max}^{RT}) + j}, MBR + \delta(i,j), d\big) \quad (33)$$

REFERENCES

[1] 3GPP TS 23.060 : GPRS Service description: Stage 2 - http://www.3gpp.org.

[2] 3GPP TS 23.107 : Quality of Service (QoS) concept and architecture - http://www.3gpp.org.

[3] CNCL (Communication Networks Class Library) - http://www.comnets.rwth-aachen.de/".

[4] B. Baynat and P. Eisenmann. "Towards an Erlang-Like formula for GPRS/EDGE network engineering". In *IEEE International Conference on Communications (ICC)*, June 2004.

[5] J. Cai and D. Goodman. "General Packet Radio Service in GSM". In *IEEE Comm. Magazine*, pages 122–131, October 1997.

[6] Y. Chung, D. Sung, and H. Aghvami. "Steady State Analysis of Mobile Station State Transitions for General Packet Radio Service". In *Proc of IEEE PIMRC*, September 2002.

[7] T. O. Engset. "On the calculation of switches in an automatic telephone system" - Tore Olaus Engset: The man behind the formula. First published in Norwegian (1915), 1998.

[8] Fang and D. Ghosal. "Performance Modeling and QoS Evaluation of MAC/RLC Layer in GSM/GPRS Networks". In *Proc. of IEEE International Conference on Communications*, May 2003.

[9] C. H. Foh, B. Meini, B. Wydrowski, and M. Zukerman. "Modeling and Performance Evaluation of GPRS". In *Proc. of IEEE VTC*, May 2001.

[10] G. Haring, H. Hlavacs, A. Kamra, and M. Bansal. "Modelling Resource Management for Multi-Class Traffic in Mobile Cellular Networks". In *Proc. of the 35th HICSS*, January 2002.

[11] C. Lindemann and A. Thummler. "Performance Analysis of the General Packet Radio Service". In *Computer Networks*, pages 1–17, January 2003.

[12] M. Mahdavi, R. Edwards, and P. Ivey. "Performance Evaluation of a Data Subsystem in GSM/GPRS Using Complete Sharing of Available Bandwidth". In *London Communications Symposium*, 2001.

[13] S. Ni and S. Haggman. "GPRS performance estimation in GSM voice and GPRS shared resource system". In *Proc. of IEEE Wireless Communication and Networking Conference(WCNC)*, pages 1417–1421, 1999.

[14] G. Nogueira. P.h.D. Thesis - University Pierre et Marie Curie - Methodes analytiques pour le dimensionnement des reseaux cellulaires, 2007.

[15] G. Nogueira, B. Baynat, and A. Ziram. "An Erlang-like law for GPRS/EDGE engineering and its first validation on live traffic". In *Proc. of International Conference on International Conference on Performance Evaluation Methodologies and Tools (ValueTools)*, October 2006.

[16] G. Nogueira, B. Baynat, and A. Ziram. "An Efficient Analytical Model for QoS Engineering in Mobile Cellular Networks". In *Proc of IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM)*, June 2008.

[17] M. Oliver and C. Ferrer. "Overview and Capacity of the GPRS (General Packet Radio Service)". In *Proc. of PIMRC'98*, 1998.

[18] S. Pedraza, J. Romero, and J. Muoz. "(E)GPRS Hardware Dimensioning Rules with Minimum Quality Criteria". In *Vehicular Technology Conference*, pages 391–395, 2002.

[19] W. J. Stewart. *An Introduction to the Numerical Solution of Markov Chain"*. Princeton University Press, New Jersey, 1994.

[20] P. Stuckmann and F. Muller. "GPRS Radio network capacity considering coexisting circuit switched traffic sources". In *European conference on wireless technologie*, 2000.

[21] P. Stuckmann and O. Paul. "Dimensioning Rules for GSM/GPRS Networks". In *Proc.of the Aachen Symposium on Signal Theory*, pages 169–174, 2001.

[22] U. Vornefeld. "Analytical Performance Evaluation of Mobile Internet Access via GPRS Networks". In *European Wireless*, February 2002.