Radio Resources Dimensioning According to Different Allocation Strategies in GSM/GPRS Networks

GEORGETA BUDURA, CORNEL BALINT, ADRIAN BUDURA
Communication Department
“Politehnica” University of Timisoara
Bd. V. Parvan. No. 2
ROMANIA
georgeta.budura@etc.upt.ro, cornel.balint@etc.upt.ro

Abstract: - The presence of mixed traffic (i.e., voice and data) in GSM/ GPRS networks raises many problems regarding dimensioning aspects. An important and critical element of these networks is represented by the radio resources. In this paper we study three models which characterize the mixed traffic process in a cell when the available radio resources are shared according different strategies between voice and data. All the strategies we consider are based on the Partial Partitioning allocation scheme. Our combined FR (full-rate) and HR (half-rate) voice traffic technique and the DHR (dynamic allocation technique) voice technique implement an optimal use of resources based on half-rate mobiles allocation capability and re-packing mechanisms. We also define performance parameters suitable for dimensioning purposes, like blocking probabilities, average throughput per user and average total throughput. We present a series of experiments which prove that the presented voice traffic allocation techniques improve the GPRS throughput.

Key-Words: - GSM/GPRS, modeling, Erlang-B law, packing, blocking probability, Engset law, throughput

1 Introduction

The integration of GPRS and EGPRS into the GSM system raises many problems. The traditional GSM system operates based on a circuit-switching technology. The GPRS network allows an end user to send and receive data in packet transfer mode hence, it optimizes the use of radio resources since, unlike the circuit switched mode, no connection between the mobile and the network is established when there is no data flow in progress. Based on the PS technology with a variable throughput the GPRS network is designed for supporting several types of data traffic, such as Wap, Web, E-Mail, etc.

Sharing resources between different users and different services is a key concept for GSM/GPRS resources dimensioning. The multiservice traffic model must lead to effective dimensioning methods, in order to avoid congestion or network over-dimensioning.

Lots of previous works have studied traffic modeling in GSM/GPRS networks trying to establish different performance parameters needed for dimensioning purposes. Dimensioning rules have to be developed to plan and estimate the radio capacity needed for the predicted amount of voice and data users when the radio resources are shared between circuit and packet switched services.

It is well known that GSM operators have dimensioned their networks for voice service in terms of offered voice traffic and blocking probability. The reference model for such systems is the Erlang-B formula [1]. This formula gives the proportion of calls that are blocked as a simple function of system capacity and voice traffic intensity. As opposed to this model, data traffic is characterized by bursts and is highly application dependant. A communication session may last for an extended period of time with intermittent packet transmissions.

Another major problem of dimensioning GSM/GPRS networks is the choice of strategy to partition the available cell capacity between traditional GSM and new GPRS services. The Radio Resources Manager (RRM) is in charge of optimizing the usage of radio resources, based on a specific resource sharing algorithm.

Three main static resources sharing schemes can be distinguished in literature [2]:

- In the first scheme, called Complete Sharing (CS), all radio channels are shared between voice and data.
- In the second one, named Complete Partitioning (CP), time-slots are divided into two sets and each type of traffic is allowed to use only its dedicated set.
- The third scheme, known as Partial Partitioning (PP), contains the following channel sets: one set shared between voice and data traffic and two sets reserved for the strict usage of a dedicated traffic type: voice or data.

The PP scheme, offers many advantages: first, reserving a set of time-slots for each type of traffic allows guaranteeing, as in CP, a minimum QoS for each type of traffic. Second, the PP scheme is more efficient than CP, which is not suitable for maximizing radio utilization, especially when dealing with a highly varying demand. Due to these advantages, the PP resource sharing algorithm is widely implemented in a large number of deployed GSM/GPRS networks.

In order to improve the number of voice users that the network can manage, different strategies can be applied when assigning the voice time-slots.

Several papers have been published on traffic modeling and performance evaluation in GSM/GPRS networks. The major works in this field are based on analytical models using queuing theory and continuous-time Markov chains, and assuming an infinite number of users in the cell [1], [3] - [6].

In [7]-[9] analytical models based on discrete-time Markov chains have been proposed and a single type of traffic is considered. It is assumed to be generated by a finite number of users and modeled by an Erlang-like law. In [2] Dahmouni et al. present an approach for dimensioning GPRS networks based on the modified Engset model.

Recently published papers address the problem of improving the quality of service and the performance parameters based on different strategies regarding the radio resources allocation [10] - [13].

In this paper we implement three models for mixed voice and data traffic and define several performance parameters in order to compare the models and to establish a suitable method for dimensioning radio resources in GSM/GPRS networks.

2 The GSM/GPRS System

Our paper considers a single cell which supports two types of traffic: GSM voice calls and GPRS data flows. In the GPRS technology a mobile station can use several time-slots simultaneously for one application, in order to perform its transmission with a higher throughput. Each time-slot can be shared among several users by assigning different Temporary Flow Identities (TFI) to the mobile phones. Up to 32 TFI’s can be allocated per TDMA frame. In addition to time-slot partitioning, the GPRS system also allows for time-slot aggregation: for a single mobile user the system can allocate up to \(d\)-time-slots simultaneously for downlink and up to \(u\)-time-slots simultaneously for uplink, depending on the mobile station’s capability class \(d+u\). The choice of the number of TBF’s that a PDCH can have in uplink and downlink depends on the operator’s choice.

Our study focuses on the radio resources allocator which distributes the downlink radio channels among voice calls and GPRS data flows.

When modeling our system we consider the following parameters:

- \(TS\): the number of time-slots, partitioned into a contiguous set of \(TS_v\) time-slots dedicated to voice calls, \(TS_{VD}\) time-slots shared between voice and data and \(TS_d\) time-slots dedicated to GPRS.

- \(d\) (resp. \(u\)): is the number of time-slots that can be used simultaneously for downlink (resp. uplink) traffic. All GPRS mobiles have the same radio capability, denoted \(d+u\).

- The RLC radio block size represents an important parameter for the GPRS system. In the downlink, IP packets are fragmented and encapsulated into LLC frames by the SGSN. The payload size of each radio block depends on the coding schemes, i.e., the applied radio error protection. The GPRS standard defines four Coding Schemes. For each coding scheme a data rate parameter, \(\mu_{GPRS}(kbits / s)\) is associated.

- We assume that in the case when there is no available channel, a voice call will be lost on arrival.

- Voice calls have a preemptive priority over data flows on the shared part of the TDMA due to the fact that they generate the largest amount of revenue in most actual operating systems. As a consequence, if all \(TS_v\) time-slots dedicated to voice are occupied and all \(TS_{VD}\) time-slots are in use with at least one of them allocated to data, then one time-slot assigned to the GPRS traffic in the shared part of the TDMA will be reallocated to voice on the arrival of a GSM request.

3 System Models

In the case of the Partial Partitioning strategy the available time slots (TS) of the TDMA are partitioned into \(TS_v\) time-slots dedicated to voice, \(TS_d\) time-slots dedicated to data and \(TS_{VD}\) time slots shared between voice and data with a total preemptive priority of voice over data on the shared
The systems we have implemented contain different models dedicated to voice and data and are also able to manage the interaction between voice and data according to the priority of voice calls over data calls. We report on the performance of three system models using the same data model and three different voice models.

### 3.1 Data Traffic Model

The data traffic is considered at the flow level timescale. A session is modeled as a series of flows (page downloads) separated by inactivity periods (think times) with no data transfer. The traffic generated by the users represents an ON/OFF process [10].

Data traffic is modeled under the assumption that there is a fixed number \( N \) of data mobiles in the cell. Each mobile is doing an ON/OFF traffic with an infinite number of pages:

- **ON periods** correspond to the download of an element like a WAP, a WEB page, an email, a file, etc. Its size is characterized by a discrete random variable \( X_{on} \), with an average value \( E[\sigma] \).
- **OFF periods** correspond to the reading time of the last downloaded element, which is modeled as a random variable \( T_{off} \) with an average value of \( E[\tau] \) seconds.
- The maximum number of GPRS users in active transfer is given by:
  \[
  n_{max}(TS_D) = \min(N, 32, mTS_D)
  \]
  where \( m \) is the maximum number of users that can use a single time-slot.

The data traffic process is based on the Engset model [2] and includes particular specifications as presented in Fig.1. This stochastic process describes the number of active users at any point in time and represents a finite state space.

![Engset model applied to data traffic process](image)

**Fig.1** The Engset model applied to data traffic process

From Fig.1 we can compute the transition rate from state \( j \) to \( j+1 \), \( \lambda_j \), as:

\[
\lambda_j = (N-j)\lambda_D = (N-j)\frac{1}{E[\tau]},
\]

for \( j = 0, 1, \ldots, n_{max} - 1 \)

Furthermore, we can express the transition rate of the death process as:

\[
\mu_j = \min(jd,TS_D)\mu_D = \min(jd,TS_D)\frac{\mu_{GPRS}}{E[\sigma]},
\]

for \( j = 1, \ldots, n_{max} \)

Because of the maximum download capacity \( d \) of each GPRS mobile, two situations can be distinguished:

1. If \( jd < TS_D \), the available bandwidth is not fully utilized by data mobiles. As a consequence the transition rate from state \( j \) to state \( j-1 \), given by the generated transfer of one mobile, is \( jd\mu_{GPRS} / E[\sigma] \);
2. If \( jd \geq TS_D \), the allocator has to share the \( TS_D \) time-slots among the \( j \) data mobiles and the transition rate from state \( j \) to state \( j-1 \) is \( TS_D\mu_{GPRS} / E[\sigma] \).

Let \( p_D(j) \) be the steady-state probability that \( j \) users are in active transfer. According to the Engset model this can be expressed as the closed form below:

\[
p_D(j) = p_D(0)\frac{C_n^j}{\prod_{i=1}^{j-1}\min(d_i,TS_D)}\rho_D^j,
\]

where \( \rho_D = E[\sigma] / (E[\tau]\mu_{GPRS}) \). We can observe that the steady-state distribution depends only through the ratio \( E[\sigma] / E[\tau] \) on the data traffic parameters \( E[\sigma] \) and \( E[\tau] \).

### 3.2 Voice Traffic Models

#### 3.2.1 Classical Erlang Model Applied for Cells with Partial Partitioning

We apply the classical Markov chain model for voice considering that new voice calls arrive according to a Poisson process with rate \( \lambda_v \) and call durations are exponentially distributed with mean \( \mu_v \). The model steady-state probabilities are given as function of voice traffic intensity \( \rho_v = \lambda_v / \mu_v \) by relation (5):

\[
p_v(t) = \frac{\rho_v}{t!} \sum_{i=0}^{TS_v + TS_D} \frac{P_v^i}{t!}, \quad t \in [0,TS_v + TS_D]
\]

It is assumed that for each call a full time slot is given. This strategy is known as full rate assignment. According to relation (5) we have computed the Erlang-B [1] formula that gives the call blocking probability necessary to dimension the cell in order to guarantee a minimum QoS for voice traffic.
\[ B_{r,pp} = \frac{\prod_{i=0}^{TS_r + TS_{wp}}}{(TS_r + TS_{wp})!} \sum_{i=0}^{TS_r + TS_{wp}} \frac{\rho^{i}}{i!} \] (6)

3.2.2 Combined FR and HR voice traffic model

A common strategy, called half-rate assignment, is to share the same time slot between two users. This increases the number of calls in the system in cases when there are few time slots available. The strategy implementation needs to define a threshold \( HRTh \), such that when the number of free time slots is less than \( HRTh \) the calls are assigned at half rate [10].

The problem associated with an HR assignment is that it can force the system into a state in which many slots are assigned at half-rate, but each slot to only one user. Because this time-slots are already allocated to voice calls and due to the priority of voice over data they are not available to data. This could induce inefficiencies in the system when the PP strategy is implemented.

The scheme we have proposed in [12] allocates half-rate capable mobiles to full-rate or half-rate channels according to the existing traffic situation in the cell. If the number of time slots (traffic channels) in the cell is above a predefined threshold, half-rate capable mobiles are allocated to full-rate channels. Otherwise half-rate capable mobiles will be given a half-rate time-slot. The transitions between the possible states are presented in Fig.2. Once a mobile has been allocated to a full/half rate time-slot, the mobile will operate in this mode until the call is terminated. The procedure drawback is the creation of so-called partially allocated time-slots which are time slots occupied by only one half-rate call. To make an optimal use of resources and to avoid the rejection of a call from a mobile that has only full rate capabilities, a re-packing procedure is applied.

In this case, it is necessary to repack two time-slots, each occupied by only one half-rate call, into a single time-slot with two half-rate calls.

For the proposed repacking model it is not possible to find out an analytic expression for the stationary distributions of the process. Knowing all the states of the system, the possible transitions between these states and their transition rates, it is possible to construct the \( Q \)-matrix of the process and find the stationary distribution by numerically solving the equation \( \mathbf{P} \mathbf{Q} = 0 \) with the normalization condition \( \sum_{n} P_{f}(n) = 1 \). We denoted by \( P_{f}(n) \) the probability of the current state of our model, where \( n \) depends on the following three coordinates: \( n(t) = (n_{f}(t), n_{h}(t), n_{i}(t)) \).

\[ p_{f}^{*}(s) = \sum_{\{i,j,k\}|j+k=s} P_{f}(i,j,k) \] (7)

3.2.3 Full DHR voice traffic model

The full DHR procedure is based on the so called half-rate operation feature of mobiles. The DHR technique always allocates half-rate capable mobiles to half-rate channels. Mobiles that are not capable of half-rate will always be allocated to a full-rate channel [13].

The current state equilibrium distribution \( P_{f\alpha}(n) \) of this model depends, again, on three coordinates: \( n(t) = (n_{f}(t), n_{h}(t), 0; 1) \).

In this case, the system is a product-form type network [14] and the equilibrium distribution is given by:

\[ p_{f\alpha}(n) = \frac{\rho^{n_{f}}_{f} \rho^{n_{h}}_{h}}{n_{f}! n_{h}!} / \sum_{n_{f}} \rho^{n_{f}}_{f} \rho^{n_{h}}_{h} \] (8)

where \( \rho_{f} \) and \( \rho_{h} \) represent the traffic for full-rate mobiles and the traffic for half-rate mobiles.
respectively, and are defined by:
\[
\rho_{V_i} = \frac{\lambda_{V_i}}{\mu_V} = \frac{\lambda _V}{(1-c_h)} \frac{1}{\mu_V}
\]
\[
\rho_{D_i} = \frac{\lambda_{D_i}}{\mu_D} = \frac{\lambda_D}{c_h} \frac{1}{\mu_D}
\]
(9)

The coefficient \( c_h \) represents the ratio of users capable to use half-rate voice coding.

For calculating the probabilities given by equation (8) we apply the one dimensional recursion formula independently published by Kaufman and Roberts. The recursion is applied to the total number of time-slots occupied by voice in a cell \( q, q \in [0, Q] \), \( Q = TS - TS_D \):
\[
q \rho_{V,i} (q) = \rho_{V,i} p_{V,i} (q-1) + \frac{1}{2} \rho_{D,i} p_{V,i} (q-\frac{1}{2}), \quad q = \frac{1}{2}, \frac{3}{2}, \ldots Q
\]
(10)

For the blocking probability we have used the expression proposed by Ross [14]:
\[
B_{V,i} = (1-c_h) [p_{V,i} (Q-\frac{1}{2}) + p_{V,i} (Q)] + c_h p_{V,i} (Q) = (1-c_h) p_{V,i} (Q-\frac{1}{2}) + p_{V,i} (Q)
\]
(11)

4 Performance parameters and experimental results
The basic idea behind constructing the mixed voice and data traffic model relies on two assumptions: (1) the voice calls are independent of GPRS connections and (2) the voice and data traffic evolve at different time scales. Based on the distributions given by equations (4),(5),(7) and (10) we have computed the average performance parameters of the system.

The average performance parameters are determined as follows:

- the average throughput per user can be expressed as [2]:
\[
X_{u,pp} = \sum_{i=0}^{TS - TS_D} p_{V}(s) \sum_{i=0}^{\min(TS - TS_D, TS - s)} (12)
\]
where \( p_V(s) \) represent the steady-states probabilities for voice calls and \( X_u = \sum_{j=1}^{n_{\text{max}}} p_D(j) \min(jd, TS_D) \mu_{\text{GPRS}} / \sum_{j=1}^{n_{\text{max}}} dp(j) \)
represents the data throughput per user when each service has its dedicated time-slots [12]. When the cell capacity is shared according to the PP schemes between voice and data, among the \( TS_{V,D} \) time-slots, those not used by the voice calls may be used for data traffic with a probability equal to the probability that \( TS - TS_D - s \) time slots are used by GSM users: \( P_V(TS - TS_D - s) \).

In equation (12) we have used three different expressions for \( p_v(s) \) according to equations (5), (7) and (10):

- for the total average throughput we propose the formula:
\[
X_{pp} = \sum_{i=0}^{TS - TS_D} p_{v}(s) \sum_{i=0}^{\min(TS - TS_D, TS - s)} (13)
\]
where (13) is similar to [2], as follows:
\[
B_{pp} = \sum_{s=0}^{TS - TS_D} p_{V}(s) B_{\min(TS - TS_D, TS - s)} (14)
\]
and represents the probability that \( TS_D \) time-slots are being used by \( n_{\text{max}} \) users among the other \((N-1)\) users.

We have implemented the models by simple programs written in Matlab considering the following parameters: \( E[\sigma] = 5KB, \ E[\tau] = 12s, \) GPRS mobile class: 4+1, CS2 coding scheme, \( TS = 8, TS_V = 3, TS_D = 1 \) and \( \rho_v = 5 \). Figures 3, 4, and 5 represent the performance parameters according to Eq. (12), (13), and (14). For the \( c_h \) coefficient we consider the values: \( c_h = 0.25 \) and \( c_h = 0.75 \).

5 Conclusion
We have implemented three mixed traffic models based on different strategies of allocating voice time-slots. Based on calculated performance parameters we conclude that combined FR-HR and Full DHR voice traffic models represent an improvement regarding resources utilization showed by a larger data throughput and a smaller data blocking probability. In addition the Full DHR technique is computational efficient implemented and it is based on \( c_h \) coefficient easy to be estimated in GSM networks.

For large \( c_h \) values, Full DHR technique shows better performances with reference to combined HR-FR.
Fig. 3 Performance parameter: Throughput per user

Fig. 4 Performance parameter: Total throughput

Fig. 5 Performance parameter: Blocking probability

References: