Analytical Model for Performance Evaluation of GPRS/EDGE Multi-Service Networks

H. Dahmouni, D. Rossé and B. Morin
Network Engineering and Tools
France Telecom Research & Development
6, avenue des usines
90007 Belfort cedex, France
{hamza.dahmouni, david.rosse, bertrand.morin}@francetelecom.com

S. Vaton
Department of Computer Science
ENST Bretagne
CS-83818
29 238 Brest cedex 3, France
sandrine.vaton@enst-bretagne.fr

Contact details of the first author:

Hamza DAHMOUNI
Network Engineering and Tools
France Telecom Research & Development
6, avenue des usines
90007 Belfort Cedex, France

Tel.: +33 384 54 43 66
Fax.: +33 384 22 07 34
hamza.dahmouni@francetelecom.com

Abstract
GPRS/EDGE networks are designed to transmit packet data. Several types of traffic with varying characteristics have to be supported by these networks. This paper proposes an analytical performance model for GPRS/EDGE networks in order to derive a simple relationship between cell capacity, traffic characteristics and performance. The model presented is based on the Modified Engset Model assuming a finite number of users generating infinite ON/OFF sessions and fair bandwidth sharing among active users. Exact and simple approximate performance formulas are derived from the model.

Keywords
GPRS, EDGE, data traffic, performance model, QoS, Engset.
Analytical Model for Performance Evaluation of GPRS/EDGE Multi-Service Networks

H. Dahmouni, D. Rossé and B. Morin
Network Engineering and Tools
France Telecom Research & Development
Belfort, France
hamza.dahmouni@francetelecom.com

S. Vaton
Department of Computer Science
ENST Bretagne
Brest, France
sandrine.vaton@enst-bretagne.fr

Abstract—GPRS/EDGE networks are designed to transmit packet data. Several types of traffic with varying characteristics have to be supported by these networks. This paper proposes an analytical performance model for GPRS/EDGE networks in order to derive a simple relationship between cell capacity, traffic characteristics and performance. The model presented is based on the Modified Engset Model assuming a finite number of users generating infinite ON/OFF sessions and fair bandwidth sharing among active users. Exact and simple approximate performance formulas are derived from the model.

Keywords—GPRS, EDGE, data traffic, performance model, QoS, Engset.

I. INTRODUCTION

GPRS/EDGE networks are designed to transmit packet data. The main challenge of this multi-service network in engineering is the integration and support of a wide variety of applications such as WAP, Web, e-mail, etc. These data services require different quality of service (QoS); the traffic corresponding to each service is highly bursty and can be characterized by a typical ON/OFF process [6]. Unlike GSM, the engineering of these networks is more difficult because the traffic generated by data users is highly dependent on the application and is very bursty. Moreover, the possibility of sharing a single time-slot between different services complicates the analysis.

Several research works have dealt with the performance evaluation of GPRS systems. The major works in this field are based on simulations (e.g. [13],[14]). Other works are based on analytical models using queuing theory and discrete or continuous-time Markov chains (e.g. [1],[9],[10],[12]). Most existing models assume an infinite number of users in the cell and thus model the arrival of traffic by a Poisson distribution. In [4], we developed an analytical model to evaluate the performance of GSM/GPRS systems, where a finite number of GPRS users generates the same data traffic. Our aim was to obtain simple and closed form formulas associating cell capacity, traffic characteristics and performance. In this model, traffic is modeled by an ON/OFF process following a general distribution (heavy-tailed) and a single service type is considered.

Recently, a multi-service model has been proposed in [2] for modeling high speed IP access links. A similar approach is used here to extend our basic model [4] in order to take into account multiple traffic types. The main contribution of this work is the precise formulations developed to predict a QoS level for multiple data traffic. Moreover, we give simple approximate performance expressions for the two distinct network states, low and high load states.

This multi-service performance model allows the analysis of the impact of the application mix (WAP, Web, etc.) on the QoS offered to different users. It also enables us to evaluate the impact of introducing a new services (e.g. video, P2P, etc.) on the QoS offered to already existing services, given that wireless services continue to evolve. Such a model is essential for an effective optimization of radio resource allocations.

The rest of the paper is organized as follows. Section II gives the main characteristics of GPRS/EDGE systems that have an influence on performance models. Section III describes the data traffic model. In Section IV, we present the mathematical description of the analytical model and the derivation of exact and approximate performance parameters for homogeneous traffic. Section IV presents an extension of the model to take into account heterogeneous data traffic. The last section provides a conclusion.

II. MAIN CHARACTERISTICS OF GPRS/EDGE SYSTEMS

In this section we present the different radio access characteristics of GPRS/EDGE systems that have an influence on performance models:

- The basic transmission unit of a PDCCH is called a radio block. The structure and the number of payload bits of a radio block depend on the coding schemes. The choice of coding scheme mainly depends on the quality of the radio link, i.e., Carrier-to-Interference (C/I) ratio. GPRS is based on GMSK modulation and proposes 4 coding schemes (CS1-CS4). EDGE introduces 8PSK modulation which allows operators to increase the data transmission rate on the air interface, and proposes 9 coding schemes (MCS1-MCS9). In the rest of the paper, we denote $\mu_{PDCCH}$ the transmission rate for one PDCCH.

- Several PDCCHs can be used simultaneously by a mobile to transfer data depending on the multi-slot class of the mobile handset. Nevertheless, each PDCCH can carry traffic for more than one connection simultaneously (a physical connection is
called a TBF). Thus, the data transmission rate also depends on the number of GPRS users multiplexed on the same PDCH. The number of TBFs that a PDCH can have in downlink and uplink depends on the operator’s choice. Each TBF is identified by a TFI of 5 bits length which limits the number of simultaneous connections to 32 per TDMA.

- Since data and voice users in the cell share the common transport media which has a limited capacity, the cell resources can be reserved statically or dynamically. However, a combination of both is also possible. In this paper we suppose that a fixed PDCH channel \( (C_d) \) is reserved statically for GPRS (or EDGE) data traffic.

III. DATA TRAFFIC MODEL

Unlike GSM voice traffic, data traffic is elastic because the transmission rate is variable and it adjusts according to the available capacity in the network. The analysis of this traffic can be represented on three scales corresponding to three entities: packet-level, flow-level which is a succession of packets of an application, and session-level which is a succession of flows of the same application.

The traffic generated by a session is usually characterized by an ON/OFF process where the ON-period corresponds to a flow or a succession of flows, e.g. send/reception of e-mail messages or a Web page load. The OFF-period (think-time) corresponds to the interval between the end of an ON-period and the beginning of the next one (see Figure 1).

![Figure 1. Example of WAP/Web traffic load within a user session](image)

Measurements of the size of documents such as Web pages, e-Mail message, etc. show that they present heavy-tail distributions like Pareto, Weibull or Log-Normal (e.g., [8],[11]). In [5] we presented a realistic traffic characterization at packet and flow levels, and provided a multi-service traffic model that is based on the statistical the analysis of GPRS traces measured on a live operational network. Table 1 shows traffic parameters for the most commonly-used services WAP and Web [5].

<table>
<thead>
<tr>
<th>TABLE 1. WAP and Web traffic parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Parameters</td>
</tr>
<tr>
<td>Flow size [byte]</td>
</tr>
<tr>
<td>Flows per page</td>
</tr>
<tr>
<td>Pages per Session</td>
</tr>
<tr>
<td>Page inter-arrival time [s]</td>
</tr>
</tbody>
</table>

IV. HOMOGENEOUS PERFORMANCE MODEL

In this section we present our basic model developed in [4]. In this model all data users generate identical characteristics for traffic and therefore receive the same QoS level. In addition we give a simple approximation of performance parameters for two distinct network states, low and high load states.

A. Model assumptions

We assume that a fixed number \( N \) of data users in the cell generates homogenous ON/OFF traffic, and that all customers have the same multi-slot class "d+u" where \( d \) and \( u \) are the numbers of time-slots that can be simultaneously used for the downlink and uplink traffic. Most customers have "\( d = 4 \) or 3" multi-slot capability in downlink and "\( u = 1 \) or 2" in uplink. We focus only on resources employed for downlink because data traffic is asymmetric and it is mostly downlink.

As the maximum number of time-slots per user is \( d \), \( s = \lfloor C_d/d \rfloor \) is the maximum number of active users such that the radio resource manager can allocate \( d \) time-slots for each one. If the number of users in active transfer is lower than \( s \), each user has \( d \) time-slots. Therefore, each one receives a service with rate \((d \mu_{PDCH})\). A fair sharing of total bandwidth is applied for all active users in transfer when the maximum number of time-slots cannot be allocated to all active users. When the number of active users is greater than \( s \), then each active user receives a service with the rate \( (C_d \mu_{PDCH})/j \) where \( j \) is the number of users in active transfer.

This bandwidth sharing model can be recognized as the generalized processor sharing (GPS) model considered by J. Cohen [3]. An interesting result of this model is that the stationary distribution of the stochastic process describing the number of active users in progress does not depend on any traffic characteristics except the averages; it depends only on the average of the ON-period and the average of the OFF-period. This is known as the insensitivity property which allows us to express the average performance only in terms of the average parameters.

B. Explicit adaptations of the Engset Model

The basic adaptation of the Engset model is to include service rate reduction (GPS bandwidth sharing). There are \( N \) users in total, each user alternating between the ON and OFF periods. We assume that the ON-period size has a general distribution with mean \( E[\sigma] \) and that the OFF-period duration follows a general distribution with mean \( E[\tau] \).

The mean size of an ON-period is \( E[\sigma] \), so when there is no radio resource congestion, the loading time is equal to \((E[\sigma]/d \mu_{PDCH})\). If there are more than \( s \) users in active transfer, all the \( C_d \) time-slots of the cell are used, then each user can be completed in an average time of \((E[\sigma]/C_d \mu_{PDCH})/j \) where \( j \) is the number of users in active transfer at a given time.

Among the total users, a random number \( J(t) \) are in ON-state at given time \( t \), where \( J(t) = 0,1,\ldots,n_{\text{max}} \). (The maximum number \( n_{\text{max}} \) of active users in transfer depends on the cell capacity, and it is limited by the maximum number \( m \) of TBFs per PDCH and the number of TFIs per TDMA.) This stochastic process describes the number of active users in progress and represents a finite state space. The insensitivity property allows us to consider an exponential distribution for the ON/OFF
process. So we can think of \( J(t) \) as a birth-and-death process with birth and death rates:

\[
\lambda_j = (N - j) E[\sigma] / E[\tau], \quad \text{for} \quad j = 0, 1, \ldots, n_{\text{max}} \quad (1)
\]

\[
\mu_j = \min(d, \frac{C_d}{j}) \mu_{\text{PDCH}} \quad \text{for} \quad j = 1, \ldots, n_{\text{max}} \quad (2)
\]

Let \( p(j) \) the steady-state probability that \( j \) users are in active transfer. It is obtained by solving the steady-state equilibrium equations of the birth-and-death process, and it is described by the closed form here below [4]:

\[
p(j) = p(0) \frac{N!}{(N-j)!} \prod_{i=1}^{j} \min(id/C_d,1) \left( \frac{D_d}{C_d} \right)^{i-1} \quad (3)
\]

where \( \rho = E[\sigma] / E[\tau] \mu_{PDCH} \) and \( p(0) \) is obtained by the normalization condition.

**C. Exact average performance**

A GPRS user perceives performance essentially through the average time necessary to transfer a document, so this time depends on the size of the document and the average throughput obtained by the user. In the following, we evaluate performance in terms of throughput obtained by each user in active transfer and blocking probability.

First, the average radio resource utilization is obtained by:

\[
U_{C_d,N} = \sum_{j=1}^{n_{\text{max}}} p(j) \min(jd/C_d,1) \quad (4)
\]

The main performance parameter of GPRS network engineering is the throughput obtained by each user. In our system, while the number of active users is \( j \), each active user receives an effective bandwidth \( r(j) \):

\[
r(j) = \frac{C_d}{j} \min(jd/C_d,1) \mu_{PDCH} \quad \text{for} \quad j = 1, \ldots, n_{\text{max}} \quad (5)
\]

The average aggregated throughput of the resource is given by:

\[
X_T = \sum_{j=1}^{n_{\text{max}}} p(j) j r(j) \quad (6)
\]

Let \( T_{ON} \) be the average time taken by the network to complete the delivery of an ON-period of size 1 bit. By Little’s law, the average number of active users equals the product of the average aggregate throughput \( X_T \) with the average time \( T_{ON} \). Thus, the average throughput \( X_{C_d,N} \) obtained by each user in active transfer is:

\[
X_{C_d,N} = \frac{1}{T_{ON}} \sum_{j=1}^{n_{\text{max}}} p(j) \min(jd/C_d,1) C_d \mu_{PDCH} \quad (7)
\]

Another important performance parameter for GPRS dimensioning is the blocking probability. The arriving GPRS user is allowed to transmit/receive data if a sufficient number of free resources are available, i.e., the number of users in active transfer is lower than \( n_{\text{max}} \), otherwise it is blocked. Thus, the blocking probability is the probability that \( C_d \) time-slots are being used by \( n_{\text{max}} \) users among the other \((N-1)\) users. It is given by the following expression:

\[
B_{C_d,N} = \frac{\Gamma - X_T}{\Gamma} = 1 - \frac{C_d}{\rho} \sum_{j=1}^{n_{\text{max}}} p(j) \min(jd/C_d,1) \quad (8)
\]

where \( \Gamma = \sum_{j=1}^{n_{\text{max}}} \lambda_j p(j) \) is the average number of new arrival.

![Figure 2. Average performance versus the number of users in the cell for \( C_d=4 \) and user capability \( (d=4 \text{ or 3}) \), \( \mu_{\text{PDCH}}=13.4\text{kb/s (CS-2)} \) and \( m=7 \).](image)

We can observe that as the number of GPRS users increases, the average throughput degrades rapidly (resp. slowly), blocking probability and radio utilization increase rapidly (resp. slowly) for high (resp. low) load \( \rho \), e.g., Web (resp. Wap). Furthermore, the multi-slot capability has a high influence on the throughput in the low load case. This is due to different levels of peak throughput when the number of users is small. On the other hand, for high cell load \( (U_{C_d,N} \approx 1) \) the multi-slot capability impact is negligible.

**D. Approximate average performance**

From figure 2, we can distinguish between two performance expressions corresponding to low and high load states. The cell is considered high loaded when all PDCHs are always fully used \( (U_{C_d,N} \approx 1) \), and low loaded when \( U_{C_d,N} < 1 \).

In a system with a finite source population the system load is equal to \( N\rho/C_d \) which is equivalent to offered traffic in Erlang. Thus, the high load state is assumed to occur when \( N\rho/C_d > 1 \), and the low load state when \( N\rho/C_d \ll 1 \).

From formula (3), we can derive simple asymptotic expansions of the performance expressions for both load states. Indeed, handset multi-slot capabilities are between 2 and 4 time-slots, the maximum number of PDCHs per TDMA being 8 time-slots. Hence, \( \min(id/C_d,1) = 1 \) for \( i > 2 \).

Thus, from (3) we derive the approximate performance formulas corresponding to formulas (7) and (8):
The use of these simple formulas allows us to determine the cell performances as a function of the input cell parameters and offered traffic per service.

V. HETEROGENEOUS PERFORMANCE MODEL

In this section, we extend our basic model for handling users with different types of traffic. In principle, several types of traffic such as WAP, Web, e-mail, etc. with different ON/OFF distributions have to be supported by the GPRS/EDGE networks. In what follows, M will be the total number of different traffic types; each traffic type has its own general distribution. E[\sigma_i] will be the average value for ON-period size and E[t_f] will be the average value for OFF-period duration for a traffic of type-k, k = 1, ..., M.

A. Two types of traffic

For simplicity’s sake, we claim that there are only two types of traffic, e.g., WAP and Web services. Let N_1 be the total number of Web users in the cell and N_2 be the total number of WAP users. Let j_i(t) be the number of Web users in active transfer and j_2(t) be the number of WAP users in active transfer at a given time t. A Generalized Processor Sharing scheduling is used for sharing bandwidth between all active users. So if the number of users in active transfer is less than s, each one receives a service with rate (d H_{PDCH}) ; if not, each one receives a service with rate (C_d H_{PDCH} / j_i) , j_i = j_1 + j_2. Thus, the effective service rate delivered to an individual user, while j users are in active transfer, remains given by the function r(j) defined by (5).

In the exponential design of user behavior, we claim that our process is a two dimensional birth-and-death Markov process on the state (j_1(t), j_2(t)) . The transition rate of this process is represented by the diagram of figure 3.

Let \Xi be the space of all feasible states,

\[ \Xi = \{(j_1, j_2) / 0 \leq j_i \leq N_i, i = 1, 2 \text{ and } 0 \leq j_1 + j_2 \leq n_{max}\} \]

At equilibrium, the steady-state balance equations are given by

\[ \begin{align*}
(N_1 - j_1) A \delta_{j_1} & + (N_2 - j_2) A \delta_{j_2} + j_1 r(j_1 + j_2) \delta_{j_1 + 1} \\
& + j_2 r(j_1 + j_2) \delta_{j_2 + 1} \rho_{j_1} \rho_{j_2} \\
& + (N_1 - j_1 + 1) A \delta_{j_1} \rho_{j_1 - 1} \\
& + (N_2 - j_2 + 1) A \delta_{j_2} \rho_{j_2 - 1} \end{align*} \]

for all \((j_1, j_2) \in \Xi\) and

\[ \delta_{k_1,k_2} = \begin{cases} 1 & \text{if } (k_1, k_2) \in \Xi \\ 0 & \text{if } (k_1, k_2) \notin \Xi \end{cases} \]

The stationary distribution of this system is given by the following closed product-form:

\[ p(j_1, j_2) = p(0,0) \frac{(j_1 + j_2)^{C_h / N_1} \rho_{j_1}^{A \delta_{j_1} / C_d} \rho_{j_2}^{A \delta_{j_2} / C_d}}{\prod_{i=1}^{d} \min(id / C_d , 1)} \]

where \( \rho_k = A_k \mu_{PDCH}^{-1} \) and \( p(0,0) \) is determined by the normalization condition.

Notice that this distribution depends on the traffic parameters \( E[\sigma_1], E[\sigma_2] \) only through the ratio \( A \). The use of this distribution allows us to determine analytically all the average performance of our system on function of the load \( \rho_i \), the available cell capacity for data traffic \( C_d \), the user capability \( d \) and the total number \( N_i \) of users.

B. Average performance evaluation

From the steady-state probabilities (11), the following average performance parameters can be derived.

As formula (7), according to Little’s law, the average throughput \( X_{C_i,N_i} \) obtained by each type-i user in active transfer is given as follows:

\[ X_{C_i,N_i} = \frac{\sum_{(j_1, j_2) \in \Xi} p(j_1, j_2) j_i r(j_1 + j_2)}{\sum_{(j_1, j_2) \in \Xi} j_i p(j_1, j_2)} \]

The type-i blocking probability \( B_{C_i,N_i} \) is given by:

\[ B_{C_i,N_i} = 1 - \frac{1}{A_i} \frac{\sum_{(j_1, j_2) \in \Xi} p(j_1, j_2) j_i r(j_1 + j_2)}{\sum_{(j_1, j_2) \in \Xi} N_i p(j_1, j_2)} \]
C. General case

In principle, several types of traffic with varying traffic characteristics have to be supported by the GPRS networks. We now consider that $M$ different services can be offered in the cells. Each service $k$ is characterized by its own distribution. Each traffic type has its own general distribution. $E[s_k]$ will be the average value for the ON-period size and $E[t_k]$ will be the average value for the OFF-period duration for a traffic of type-$k$, $k=1,...,M$. We assume that there is a fixed number $N_k$ of users generating traffic with type-$k$. In the exponential conception of user behavior, the system is modeled as an $M$-dimensional birth-and-death Markov process, with vector $\vec{j} = (j_1, j_2,...,j_M)$ representing the state of the system, i.e., the number of active users (in transfer) in the cell at a given time.

The steady-state probability is obtained by solving the steady-state equilibrium equation of the $M$-dimensional birth-and-death Markov process. It is given by the following closed product-form [7]:

$$p(\vec{j}) = p(\vec{0}) \prod_{k=1}^{M} \frac{C_{N_k}^{j_k} \rho_k^{j_k}}{\min(d, C_{d}/j_i)} \quad (14)$$

where $\rho_k = \frac{E[s_k]}{E[t_k]} \mu_{GPRS}$ and $p(\vec{0})$ is obtained by the usual normalization condition.

Using this expression, we can calculate the mean number of each type of user in active transfer, the average duration of an ON-period transfer for each service type. The average throughput obtained by each type of user in active transfer, as well as the blocking probability of each service type, can be obtained by applying relations (12) and (13).

As can be seen from figure 4, the mix traffic can be represented by an equivalent one with specific ON-period size and OFF-period duration [2]. Therefore, the approximate formulas (9) and (10) are also applicable on this equivalent traffic.

From these approximate formulas we can conclude that the system saturation level is strongly dependent on the presence of services with high ON-period size (e.g. Web service, FTP or a new service like P2P). For several active Web users, the total offered traffic is high, hence the 4 PDCHs are fully loaded, thus the average throughput decreases rapidly while the number of users increases. Furthermore, the presence of a service with a high load $\rho$ has a great influence on the performance of other services available in the system.

VI. CONCLUSION

In this work, we have developed an analytical model for GPRS/EDGE access networks. This model constitutes a data traffic model which presents a simplicity level comparable to the Engset model for GSM networks. Several extensions of this model have been implemented, in particular, the integration of voice and data services as well as the data multi-service case. The proposed model allows us to better measure the performance of the system according to traffic load by service. In particular, the approximate formulas derived from the model can be used to easily evaluate the impact of traffic composition on performance. Moreover, it makes it possible to predict the impact of the introducing a new service on the QoS level offered to already existing services.

REFERENCES