Modeling and Performance Evaluation of GPRS

Chuan Heng Foh[1], Beatrice Meini[2], Bartek Wydrowski[1] and Moshe Zukerman[1]

[1]ARC Special Research Centre for Ultra-Broadband Information Networks,
EEE Department, The University of Melbourne,
Parkville, Vic. 3010, Australia
{chuanhf, b.wydrowski, m.zukerman}@ee.mu.oz.au

[2]Dipartimento di Matematica,
Universita' di Pisa, via Buonarroti 2,
56127 Pisa,
Italy
meini@dm.unipi.it

Abstract

This paper provides an accurate model of the General Packet Radio Service (GPRS). GPRS is modeled as a single server queue in a Markovian environment. The queueing performance of data packets is evaluated by matrix geometric methods. The arrival process is assumed to follow a two state Markov modulated Poisson process (MMPP), and the service rate fluctuates based on voice loading. The analytical results are confirmed by simulation.

1. Introduction

The rapid growth of the Internet has prompted a need for wireless data access to the Internet. Although Global System for Mobile Communications (GSM) systems provide a fixed rate data service, they result in inefficient use of bandwidth for data users due to the bursty nature of data traffic.

To carry data traffic more efficiently, the use of dynamic channel allocation for data packets was implemented in GSM, known as General Packet Radio Service (GPRS) [1]. Fundamentally, GPRS is based on the *hybrid switching* [2] principle with two types of traffic: (i) voice calls and (ii) packet data. In practice, in accordance with current design and traffic management policies, voice calls have priority over data packets and data packets which cannot immediately be transmitted are queued at the source.

In modeling GPRS, we assume for simplicity that all awaiting data packets in all remote sources are in a single server queue (SSQ). The bandwidth available to these data packets is dependent upon the number of voice calls in the system.

To capture the bursty nature of data traffic, we model the arrival process as a Markov Modulated Poisson Process (MMPP). It is shown that MMPP can be used to produce bursty traffic [3], and we further demonstrate the fundamentally significant short range dependent property of MMPP.

The single server queue is modeled as a queue in a Markovian environment to reflect both the MMPP arrival process and the effect of voice loading on the packet data queueing performance. We use matrix analytic methods [4] to obtain the numerical results of the delay of data packets. To verify our analytic model, we develop a simulation program that models the operation of GPRS with more details. It includes the signaling time required for data packets, and the non-exhaustive use of a time slot of a GSM TDMA frame.

This paper is structured as follows. In Section 2, GPRS is outlined. The analytic model is described in Section 3. In Section 4, the performance evaluation of GPRS is presented. Finally in Section 5, we verify the analytical results with the simulation results.

2. General Packet Radio Service

GPRS is a GSM packet radio service. GSM shares the radio spectrum resource by performing both frequency-division multiple access (FDMA) and time-division multiple access (TDMA). FDMA divides the 25Mhz spectrum into 124 carrier frequencies spaced 200khz apart. A certain number of these frequency bands is allocated to a base station of a cell. Each of these frequency bands is further divided in time. Eight channels are created by dividing time into eight time slots. A TDMA frame is formed by packing the eight time slots.

In the fixed rate data service of GSM, a data user is permanently allocated two time slots from every TDMA. One of the time slot is used for uplink and another for downlink. Due to the bursty nature of data packets, the fixed time slot allocation is inefficient.

GPRS employs the concept of capacity on demand by dynamically allocating a certain number of time slots to a data user whenever there are data packets waiting for transmissions. Furthermore, uplink and downlink are allocated separately so that more efficient bandwidth usage can be reached for asymmetric data traffic. For a cell that supports GPRS, all the radio resources are shared by both GSM voice users and GPRS data users. The unused time slots from voice traffic may be reallocated to carry GPRS traffic.

There are four coding options available for GPRS: CS-1, CS-2, CS-3 and CS-4 capable of transmitting 9.05k, 13.4k, 15.6k, and 21.4k bits in every second respectively. They are tailored to suit different channel conditions.

TABLE 1. Parameters of the MMPP traffic in Fig. 1

	(a)	(b)	(c)	(d)
λ_0	0.01	0.01	0.01	0.01
λ_1	1	0.1	0.05	0.05
r_0	10^{-8}	10^{-8}	10^{-5}	0.5
r_1	$2\cdot10^{-9}$	$2\cdot10^{-9}$	10^{-6}	0.005

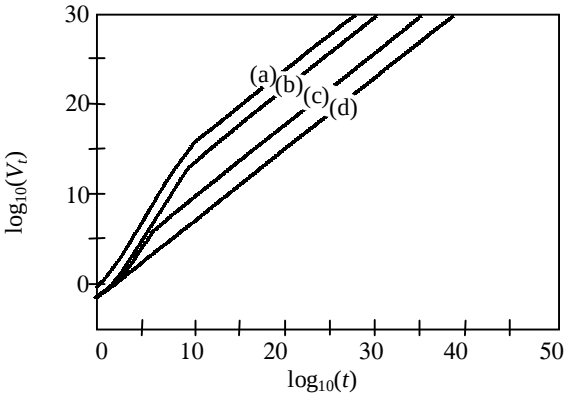

Fig. 1. The Variance Time Curve of MMPP

Readers are referred to [1] for more details of GPRS specifications.

3. Queueing Model

The packet data arrival traffic model

The arrival process is assumed to follow a two state MMPP [3]. A two state MMPP is an alternating Markovian process with two arrival states, where the arrival process in arrival state m is a Poisson process with rate: λ_m, m=0,1. The sojourn time in each arrival state is exponentially distributed with the mean sojourn time in arrival state 0 and 1 being r_0^{-1} and r_1^{-1} respectively. The values of the mean and the variance of MMPP traffic are given in [3]. Let N_t be the amount of work arriving during time interval (0, t). For MMPP, the mean of N_t, denoted E[N_t], is:

$$E[N_t] = \frac{\lambda_0 r_1 + \lambda_1 r_0}{r_0 + r_1} \cdot t$$

and the variance of N_t, denoted V_t, is:

$$V_t = E[N_t] \cdot \left(1 + \frac{2(\lambda_0 - \lambda_1)^2 r_0 r_1}{(r_0 + r_1)^2 (\lambda_0 r_1 + \lambda_1 r_0)} - \frac{2(\lambda_0 - \lambda_1)^2 r_0 r_1}{(r_0 + r_1)^3 (\lambda_0 r_1 + \lambda_1 r_0) \cdot t} \left(1 - e^{-(r_0 + r_1)t}\right) \right)$$

It is well known that the autocorrelation of the arrival process has significant effect on queueing performance. It is also well known that real traffic (data and VBR video) exhibits long range dependence (LRD) [5]. However, since buffer size is limited so is the time period over which autocorrelation has effect. The larger the buffer size, the longer is the time period over which autocorrelation affects queueing performance. If the buffer size is equal to zero, autocorrelations has no effect on performance.

If we accept the view that for a given buffer size, the shape of autocorrelation curve, from a certain point onwards, does not affect queueing performance, we can use the MMPP, which is a short range dependent (SRD) process, to model LRD real traffic for the purpose of queueing performance evaluation.

For that purpose, we will consider the variance time curve of MMPP. In Fig. 1, we plot four variance time curves with different parameter sets of MMPP traffic. By comparing the four presented curves, the critical time interval of the traffic and the slopes of the curve within the critical interval can be controlled by the parameters of MMPP. For curves (a), (b) and (c), we obtained SRD traffic with different critical time intervals.

The voice traffic model

A Voice call is allocated a channel for the duration of the call. Regardless of whether there is silence or activity a slot is dedicated to the voice call for the duration of the call. Therefore admission of a voice call decreases the number of channels (servers) available for the data packets and the departure of a call increases the number of available channels (servers) available for the data packets.

The departure and arrival process of the voice traffic is modeled as an M/M/k/k process, where k is the number of voice channels present in the system. The voice arrival process is Poisson with parameter λ_v and the voice call holding time is assumed exponential with mean $1/\mu_v$.

4. Queueing Analysis

We have modeled GPRS as a queue in a Markovian environment, and we follow Neuts' analysis of such a queueing model as described in [4] pages 254-264. The infinitesimal generator is first introduced here to describe the system, then, we review Neuts' solution as applied to our queueing problem, and finally we show how the rate matrix R, required for Neuts' solution, is computed.

Infinitesimal Generator

We will assume that there are c channels (c does not include the channels allocated for signaling), out of total c channels we assign d channels exclusively for data packets. Therefore only c-d channels are available for circuit switched voice traffic. The state of the system under consideration is denoted by the three dimensional vector (i,j,m) where i is the number of data packets in the queue (including the one in service), j is the number of channels available for data packets, m is the arrival state (m takes the values 0 and 1). The service rate is always equal to $j\mu$, where μ is the service rate provided by one

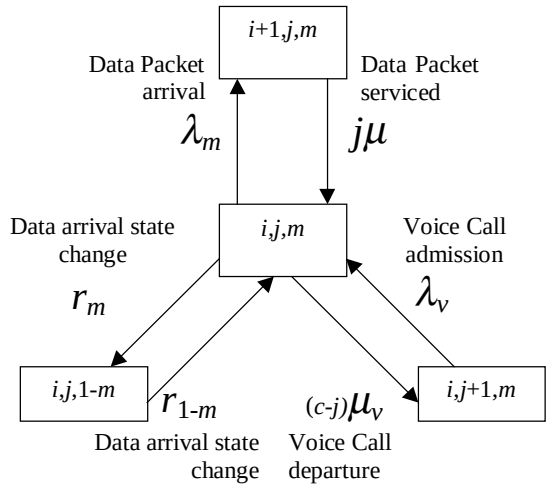

Fig. 2. State Transition Diagram

data channel, and $j = d, d+1, d+2, \ldots, c$. The packet data arrival rate is λ_m, $m = 0, 1$. All the possible state transitions are presented in Fig. 2. Hence, Fig. 2 defines the infinitesimal generator matrix G.

The sum of the entries in each row of G is 0. Let $\hat{h}_{i,j,m}$ be the probability of being in state (i, j, m), then the vector $\hat{h}$ is:

$$(\hat{h}_{0,0,0}, \hat{h}_{0,0,1}, \hat{h}_{0,1,0}, \hat{h}_{0,1,1}, \ldots,$$
$$\hat{h}_{0,c,1}, \hat{h}_{1,0,0}, \hat{h}_{1,0,1}, \hat{h}_{1,1,0}, \hat{h}_{1,1,1}, \ldots,$$
$$\hat{h}_{1,c,1}, \hat{h}_{2,0,0}, \hat{h}_{2,0,1}, \hat{h}_{2,1,0}, \hat{h}_{2,1,1}, \ldots).$$

The state transition balance equation is then $\hat{h}G = 0$. The steady state queue size distribution x_i is related to $\hat{h}$ as such $x_i = \sum_m \sum_j \hat{h}_{i,j,m}$. We now obtain x_i by another approach, using Neuts' analysis.

Neuts' solution

The steady state queue size distribution vector $\boldsymbol{x}$ can be solved analytically by considering the data packet SSQ's parameters being driven by a Markovian environment. The Markovian environment is comprised of two processes, the voice call arrivals and departures as well as the transition between data arrival states 0 and 1. These processes are independent, and determine the state of the environment j and m. The transition probability matrix of the Markovian environment is denoted Q and it is shown in Fig. 3.

The Markovian random environment is the stochastic process that determines the number of voice calls in the system and the packet data arrival state. As discussed, the packet data service rate fluctuates based on the voice calls in the system. For each state of the environment, there is an appropriate data service rate in vector $\boldsymbol{\mu}$ and packet data arrival rate in vector $\boldsymbol{\lambda}$.

The packet data service rates in each state of the environment (j, m) are:

$\boldsymbol{\mu}_{j,m} = \boldsymbol{\mu}_j = j\mu$, m=0,1 and $j=d, d+1, d+2, ..., c$.

Let vector $\boldsymbol{\mu}$ be defined by

$\boldsymbol{\mu} = (d\mu, (d+1)\mu, (d+2)\mu, \ldots, c\mu)$.

The packet data arrival rates in each state of the environment (j, m) are:

$\boldsymbol{\lambda}_{j,m} = \lambda_m$, m=0,1 and $j=d, d+1, d+2, ..., c$.

Let vector $\boldsymbol{\lambda}$ be defined by

$\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_0, \lambda_1, \lambda_0, \lambda_1, \ldots. \lambda_1)$.

For example, if j=10 and m=1, with c=22 and d=1, there are 22 channels with 1 reserved for data packets, and there are 12 voice calls in the system as there are 10 packet data channels available. The packet data arrival state is 1. The data service process then has parameter 10μ (since 10 channels are allocated to serving data) and the data arrival process has parameter λ_1.

Assuming the queue is stable (mean arrival rate < mean service rate), to determine the queueing performance of the GPRS system, the stationary probability vector $\boldsymbol{x}=(x_0,x_1,x_2,\ldots)$, which describes the probability distribution of the queue size, needs to be determined.

Using Neuts' formalisation of the M/M/1 queue in a Markovian environment [4], this problem is translated into finding the minimal solution for the matrix R, which satisfies the matrix equation

$$R^2\Delta(\mu) + R(Q - \Delta(\lambda + \mu)) + \Delta(\lambda) = 0$$

where Q is the transition probability matrix of the Markovian environment, and $\Delta(z)$ the diagonal matrix of the vector $\boldsymbol{z}$.

The matrix R can be evaluated using a cyclic reduction

	d,0	d,1	d+1,0	d+1,1	…	c,1
d,0	$-r_0-(c-d)\mu_v$	r_0	$(c-d)\mu_v$			
d,1	r_1	$-r_1-(c-d)\mu_v$		$(c-d)\mu_v$		
d+1,0	λ_v		$-\lambda_v-r_0-(c-d-1)\mu_v$	r_0		
d+1,1		λ_v	r_1	$-\lambda_v-r_1-(c-d-1)\mu_v$		
…						
c,1						$-\lambda_v-r_1$

Fig. 3. The Matrix Q

algorithm described in the next subsection. The stationary probability vector $\boldsymbol{x}$ of the stable queue is given by

$$x_k = \pi(I-R)R^k \quad for\ k \geq 0$$

where π is the stationary probability vector of Q (Eq. (6.2.5) from [4]). The vector π is given by solving $\pi \cdot Q = 0$ by, for example, successive relaxation.

The mean queue size is computed using the stationary probability vector $\boldsymbol{x}$. The mean data packet delay is found using Little's law. Packet delay is the time from the generation of the packet to the time the last bit is sent. The mean delay is thus obtained as follows:

$$mean\ packet\ arrival\ rate = \frac{\lambda_0 r_1 + \lambda_1 r_0}{r_0 + r_1}$$

$$mean\ queue\ size = \sum_{i=1}^{i=\infty} x_i \cdot i$$

$$mean\ delay = \frac{mean\ queue\ size}{mean\ packet\ arrival\ rate}.$$

Computation of the rate matrix R

In this subsection we describe the algorithm for the computation of the rate matrix R. The matrix R is the minimal nonnegative solution of the matrix equation

$$A_0 + A_1 R + A_2 R^2 = 0$$

where $A_0 = \Delta(\mu)$, $A_1 = Q - \Delta(\lambda+\mu)$, $A_2 = \Delta(\lambda)$. Here, minimal nonnegative solution means that for any other nonnegative solution $\hat{R}$, the matrix R is entrywise less than the matrix $\hat{R}$.

This effective numerical method that works in the case where the associated M/M/1 queue is transient or positive recurrent [4]. The method is based on the use of cyclic reduction and is extensively discussed in the early paper [7] and later in [8, 6, 9]. Here we will recall the algorithm and its convergence properties and we refer the reader to [7, 6] for details and proofs.

This method has nice properties, namely it is quadratically convergent, that is the approximation error e_j at the j-th step is such that $e_j \leq c\sigma^{2^j}$ for suitable constants $c>0$ and $0<\sigma<1$, and has a very low computational cost. Moreover all the computations involved in this scheme are numerically stable, since they consist of sums and products of nonnegative matrices and inversions of M-matrices (see [7]).

The algorithm consists of generating the four sequences of matrices $\{A_0^{(j)}\}_{j\geq 0}$, $\{A_1^{(j)}\}_{j\geq 0}$, $\{A_2^{(j)}\}_{j\geq 0}$, $\{\hat{A}^{(j)}\}_{j\geq 0}$ according to the following recurrences:

$$\begin{aligned}
A_0^{(j+1)} &= -A_0^{(j)}\left(A_1^{(j)}\right)^{-1}A_0^{(j)}\\
A_1^{(j+1)} &= A_1^{(j)} - A_0^{(j)}\left(A_1^{(j)}\right)^{-1}A_2^{(j)} - A_2^{(j)}\left(A_1^{(j)}\right)^{-1}A_0^{(j)}\\
A_2^{(j+1)} &= -A_2^{(j)}\left(A_1^{(j)}\right)^{-1}A_2^{(j)}\\
\hat{A}^{(j+1)} &= \hat{A}^{(j)} - A_2^{(j)}\left(A_1^{(j)}\right)^{-1}A_0^{(j)},\ j \geq 0,\\
A_0^{(0)} &= A_0,\ A_1^{(0)} = A_1,\ A_2^{(0)} = A_2,\ \hat{A}^{(0)} = A_1.
\end{aligned} \tag{1}$$

In [2, 1] it is shown that these sequences of matrices are such that:

1. for any $j\geq 0$ it holds
$$\hat{A}^{(j)}R + A_2^{(j)}R^{2^j+1} = -A_0;$$
2. if the associated M/M/1 queue is positive recurrent, then the sequences $\{A_0^{(j)}\}_{j\geq 0}$ converges in norm to zero as σ^{2^j}, where $\sigma = \max\{|\alpha|:\det(A_0+\alpha A_1+\alpha^2 A_2) = 0,\ \alpha\in C,\ |\alpha|<1\}$ is the largest modulus eigenvalue of R;
3. if the associated M/M/1 queue is transient, then the sequence $\{A_2^{(j)}\}_{j\geq 0}$ converges in norm to zero as σ^{2^j}, where $\sigma = 1/\min\{|\alpha|:\det(A_2+\alpha A_1+\alpha^2 A_0) = 0,\ \alpha\in C,\ |\alpha|>1\}$;
4. the sequence $\left\{\left(\hat{A}^{(j)}\right)^{-1}A_2^{(j)}R^{2^j+1}\right\}_{j\geq 0}$ converges in norm to zero as σ^{2^j}, where $0<\sigma<1$ is one of the two values defined above.

From these properties, an approximation of R is given by $-\left(\hat{A}^{(j)}\right)^{-1}A_0$, for a moderately large value of j. Moreover, in order to stop the computation, we check the minimum value between the norms of $A_0^{(j)}$ and $A_2^{(j)}$: if this value is smaller than a threshold value ε, then $-\left(\hat{A}^{(j)}\right)^{-1}A_0$ will be an approximation of R.

The resulting algorithm can be synthesized by the following scheme:

Algorithm: Cyclic reduction for computing *R*

Input The matrices A_0, A_1, A_2 and an error bound $\varepsilon > 0$ for the stopping condition.

Output An approximation $\tilde{R}$ of R.

Computation

1. Set $j = 0$, $A_0^{(0)} = A_0$, $A_1^{(0)} = A_1$, $A_2^{(0)} = A_2$, $\hat{A}^{(0)} = A_1$.
2. Compute $A_0^{(j+1)}$, $A_1^{(j+1)}$, $A_2^{(j+1)}$, $\hat{A}^{(j+1)}$ by means of (1) and $r = \min\left\{\left\|A_0^{(j+1)}\right\|, \left\|A_2^{(j+1)}\right\|\right\}$, where, for a matrix $B = (b_{h,k})_{h,k=1,\ldots,n}$, $\|B\| = \max_{h=1,\ldots,n}\sum_{k=1}^{n}\left|b_{h,k}\right|$.
3. If $r \geq \varepsilon$ set $j = j + 1$ and go back to step 2; otherwise, set $\tilde{R} = -\left(\hat{A}^{(j+1)}\right)^{-1}A_0$.

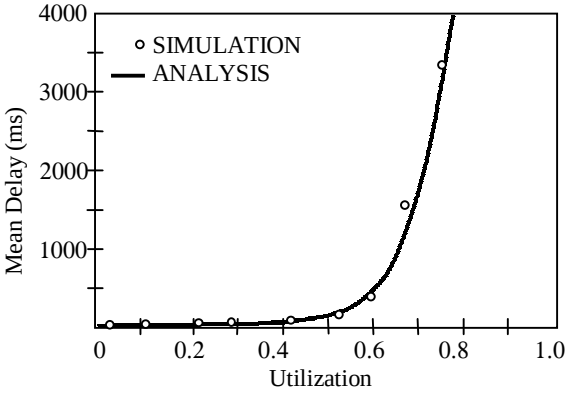

Fig. 4. Mean delay versus utilization

5. Simulation Testing

The analysis was performed for a typical cell. In a typical cell, there are 24 available channels, two of which are often reserved for signaling and Short Message Service (SMS). Of the 22 available, one is reserved exclusively for packet data traffic, leaving 21 for voice traffic. The mean packet length was set to 512 bytes, a typical TCP packet, making the mean data service rate μ=512*8/45.25=1/90.519 packets/burst period. The mean voice call holding time was taken to be 36000 burst periods long (3 minutes), and the call arrival rate was chosen so that the cell is 30% utilized by voice calls λ_v=0.3·(c-d)·μ_v. Table 2 lists all of the parameters used. These parameters were used to obtain a result using the analytical approach as well as by simulation.

TABLE 2. Model Parameters

Parameter	Value	Meaning
λ_0	variable	Packet arrival rate state 0
λ_1	$\lambda_1=5\lambda_0$	Packet arrival rate state 1
r_0	0.00001	State 0 to 1 transition rate
r_1	0.000001	State 1 to 0 transition rate
μ	1/90.519	Mean service rate
c	22	Total channels
d	1	Data reserved channels
μ_v	1/36000	Voice call 1/(hold time)
λ_v	0.3·(c-d)·μ_v	Voice call arrival rate

The simulation includes signaling delays not modeled in the analysis. To account for resource allocation, each data packet generated suffers a fixed delay of 1 GSM TDMA frame. If a transmission (data or voice) ends during a time slot, the remaining time of this time slot cannot be used for further voice or data traffic.

Packet delay results obtained using the parameters in Table 2 are shown in Fig. 4 for both the analysis and simulation. The agreement between the analytical model and simulation validates the analytical approach.

Interestingly, it appears that the additional overheads that the simulation accounted for had an insignificant impact on the mean queueing delay. Resource allocation for a packet transmission is suffered only by the first TDMA frame of the packet. For packets spanning many frames, this results in insignificant additional delay. The non exhaustive use of time slots also did not contribute to any significant delays. This is also true as long as voice or packet transmissions last over many time slots.

Fig. 4 indicates that in order to operate GPRS with reasonable packet delays the utilization must be kept below 60%.

6. Conclusion

The focus of this work was to present a simple and accurate analytical model of the GPRS system. The use of MMPP to obtain SRD traffic, captured the bursty nature of packet data traffic. Matrix geometric techniques were used to obtain queueing performance results. Simulation testing agreed with the analytical model results.

References

[1] R. Kalden, I. Meirick and M. Meyer, "Wireless Access Based on GPRS," *IEEE Personal Communications*, April 2000.

[2] M. Zukerman, "Circuit allocation and overload control in a hybrid switching system," *Computer Networks and ISDN Systems*, vol. 16, no. 4, 1989, pp. 281-298.

[3] H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE JSAC*, vol. SAC-4, no. 6, September 1986.

[4] M. F. Neuts, "Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach," *The Johns Hopkins University Press*, Baltimore, MD, 1981.

[5] R. G. Addie, M. Zukerman and T. D. Neame, "Broadband traffic modeling: simple solutions to hard problems," *IEEE Communication Magazine*, August 1998, pp. 88-95.

[6] D. A. Bini, L. Gemignani, and B. Meini, "Computations with infinite Toeplitz matrices and polynomials," *Linear Algebra Appl.*, 2001.

[7] D. A. Bini and B. Meini, "On the solution of a nonlinear matrix equation arising in queueing problems," *SIAM J. Matrix Anal. Appl.*, 17:906-926, 1996.

[8] D. A. Bini and B. Meini, "Improved cyclic reduction for solving queueing problems," *Numerical Algorithms*, 15:57-74, 1997.

[9] C. He, B. Meini, and N. H. Rhee. "A shifted cyclic reduction algorithm for QBD's," *SIAM J. Matrix Anal. Appl.*, 2001.