Analysis of the Effects of Outages on the Quality of Service of GPRS Network Systems

F. Tataranni*, S. Porcarelli*, F. Di Giandomenico** and A. Bondavalli***

* University of Pisa, Computing Engineering Dep., via Diotisalvi 2, I-56126, Pisa, Italy
** IEI/CNR, Via V. G. Moruzzi 1, I-56010 Ghezzano, Pisa, Italy
*** University of Firenze, Dip. Sistemi e Informatica, Via Lombroso 6/17, I-50134 Firenze, Italy

Abstract

The General Packet Radio Service (GPRS) extends the Global System Mobile Communications (GSM) by addressing packet-oriented traffic. Availability is the most important dependability requirement for such Communication Systems as GPRS. Focusing on the contention phase, where users compete for channel reservation, this paper analyses the GPRS with the objective to understand its behaviour under critical conditions, as determined by periods of outages, which significantly impact on the resulting dependability. In fact, during outages (service unavailability), users trying to access the service accumulate, leading to an overload of the system. When the system resumes its operations, the accumulated users determine a higher probability of collisions on resources assignment (and therefore a degradation of the QoS perceived by the users). Our analysis, performed using a simulation approach, allowed to gain insights on the impact of outages on the QoS and of the overload that GPRS systems have to face after outages.

1 Introduction

GPRS (General Packet Radio Service) has been developed to enhance the Global System Mobile Communications (GSM) system with the introduction of services based on a packet switching technique. These services provide a more efficient use of the radio resources, by accommodating data sources that are bursty in nature, at more convenient costs for subscribers. Typical examples of applications producing bursty traffic are Internet applications, e.g. World Wide Web, FTP and e-mails.

Work on GPRS has started in 1994, and a standardisation of the GPRS specification has been recently performed by ETSI (European Telecommunications Standard Institute). Analyses of the GPRS expected behaviour have been performed, essentially focusing on measures like throughput, delay for the transmission of a frame from the source to the destination, and a measure of the blocking phenomenon due to contentions on the random access attempts to get the resource (channel) for data transmission (e.g., [1], [2], [5], [10]). At the same time, GPRS, like other networked systems, is availability-critical, and significant effort is being devoted by systems suppliers to get always higher competitive availability figures. Availability is defined as the property of "readiness of usage" [6], measured as the delivery of correct service with respect to alternation of correct-incorrect service. However, in the case of communication systems, like GPRS, whose services are continuously required by users, the mere estimation of availability in terms of intervals of times the system is operative with respect to those in which the system is halted is not a satisfactory measure to know. In fact, it can be easily observed that, during the period of stoppage (outage), users requesting services accumulate, waiting to have their request(s) accepted as soon as the system is up again. Therefore, an overload occurs at system restart; the high number of requests leads to a higher probability of collisions to get access to system resources, with a negative impact on the offered quality of service (QoS). The system requires some time to absorb the congestion before getting back to the "normal" behaviour, in which the "normal" QoS is provided to users.

The goal of this work is to analyse the behaviour of the GPRS in presence of outages, in order to estimate the degradation of the QoS as perceived by users and to understand the relevant phenomena. More specifically, we concentrate on the behaviour of the GPRS during the contention phase where users compete for the channel reservation using a random access procedure. As already observed, the intervals of time following the end of periods of outages are characterised by overloads that are responsible for a degraded QoS that gradually improves, moving towards the reference average values. Intuitively,

0-7695-1101-5/01 $10.00

the duration of these intervals and how degraded is the QoS in the meantime seem to depend mainly on the duration of the outage, although other system conditions may play an important role as well. A proper understanding of the evolution of such characteristics is therefore a valuable source of information, especially when policies for contrasting unavailability are considered. The approach followed towards our goal has been to define a model of the GPRS and analyse it using a simulation method.

Indicators of the QoS degradation have been identified and analysed using a transient type of analysis. Also, "normal" QoS provided by the system has been measured through steady-state type of analysis, to get average reference indicators, necessary to appreciate the degradation of GPRS services implied by unavailability periods. Our study allows achieving useful insights on the influence of internal and external parameters on the figures of merit in such critical conditions. Also, useful suggestions can be derived for the system provider on appropriate settings to balance between system performance/cost and user satisfaction.

The rest of the paper is organised as follows. Section 2 gives an overview of the GPRS architecture. Section 3 introduces the relevant figures of merit we have identified for our objective, describes our assumptions and presents a first model of the GPRS behaviour in presence of outages. Section 4 deals with the definition of the settings for the numerical evaluation. Section 5 is then devoted to illustrate the results of the experiments performed to evaluate the effects of outages. In Section 6, an extension of the model is proposed, consisting in the addition of a queue mechanism, and the impact of such new feature on the identified indicators is analysed, also performing comparisons with those obtained by system configurations not employing it. Conclusions are found in Section 7.

2. GPRS Overview

The GPRS introduces a packet oriented data service for GSM, with a more efficient packet switching allocation mechanism. An important goal of the GPRS technology is to allow GSM license holders to share physical resources on a dynamic, flexible basis between packet data services and other GSM services. We briefly recall here the main characteristics of the GPRS [3] [4].

To introduce GPRS in the existing GSM infrastructure, additional elements are needed to provide support for packet switching, namely: *Service GPRS Support Node* (SGSN) and *Gateway GPRS Support Node* (GGSN). The SGSN controls the communications and mobility management between the mobile stations (MS) and the GPRS network. The GGSN acts as an interface between the GPRS network and external packet switching networks such as Internet, or GPRS networks of different operators. Between GPRS Support Nodes (i.e., SGSN and GGSN), an IP based backbone network is used. The Base Station Subsystem (BSS) is shared between GPRS and GSM network elements, to maintain compatibility and to keep low the investments needed to introduce the GPRS service. The ISO/OSI structure of the system is shown in Figure 1.

The Sub Network Dependent Convergence Protocol (SNDCP) provides functionalities to map different network protocols onto logical link supported by the Logical Link Control (LLC) layer; this last is responsible for moving user data between mobile stations and the network. The Radio Link Control (RLC) layer allows to transmit data across the air interface.

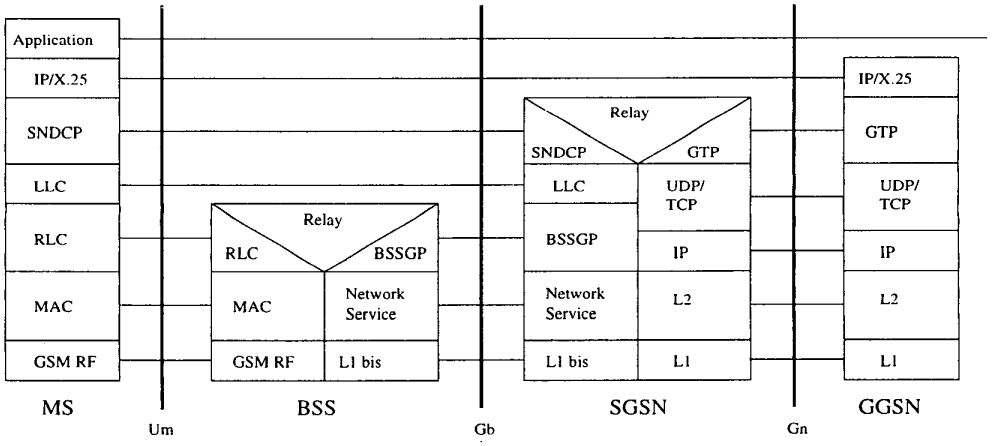

Figure 1: The ISO/OSI structure of GPRS

A Medium Access Control (MAC) layer is introduced, to control data transmission in packet oriented mode. The RLC/MAC layer will ensure the concurrent access to radio resource among several MSs. Each RLC block is divided in four normal bursts that have the same structure as GSM radio bursts, since GPRS shares the same physical layer as GSM. The GPRS allows several "Logical channels" to share physical channel (called Packet Data CHannel, PDCH) through time division multiplexing. PDCHs are associated with a single time slot of a TDMA frame (composed of 8 time slots). In a cell that directly supports GPRS, a Master PDCH is allocated, to provide control and signalling information to start data transfer both in up-link and in down-link, and to handle the users mobility. A MPDCH accommodates a logical channel for up-link transmission of channel request: the Packet Random Access Channel (PRACH). When a mobile station needs to transmit, it has to send a channel request to the network. The access method, based on a *Random Access Procedure*, can cause collisions among requests by different MSs, so it may become a bottleneck of this system. This is the specific aspect of GPRS addressed in this work, which therefore deserves a more detailed description [4]. The MSs get the access control parameters by listening to the Packet Broadcast Control CHannel (PBCCH). Such parameters are the number of maximum re-transmissions M, the persistence level P and the parameters S and T. The MS is allowed to make a maximum of M+1 attempts to send a Packet Channel Request message. At the beginning of the procedure a timer is set (to 5 sec). At the expiry of this timer, the procedure, if still active, is aborted and a failure is indicated to the upper layer. The first attempt to send a Packet Channel Request can be initiated at the first possible TDMA frame containing PRACH. For each attempt, the mobile station extracts a random value R, and only if R is bigger than, or equal to, the persistence level P the station is allowed to send a Packet Channel Request. After a request is issued, the MS waits a time, which depends on S and T; if it does not receive the Packet Down-link Assignment (or a Packet Queuing) in this time, a new attempt is tried, if it is still allowed to make one, otherwise a failure is notified to the upper layer. The parameters S and T are also used by a MS to determine the next TDMA frame in which a new attempt is possible, should the previous be unsuccessful and a new attempt still allowed. Traffic packet data channels, called slave PDCH, are needed in a cell to transport users data and transmission signalling, as acknowledged and non-acknowledged message. For what concerns data transfer, up-link and down-link channels allocation is completely independent and a MS can operate up-link and down-link data transfer simultaneously.

3. Modelling the GPRS behaviour

We focused on the GPRS behaviour during the contention phase performed by users when making channel request; as previously discussed, this is a critical part, being a potential bottleneck of the system. Other GPRS critical phases, such as bottlenecks in up-link at points where traffic from several cells converge, as well as contention on down-link response messages, are not considered in this work. Accounting for them, and devising ways to include this study in a more complete framework, are interesting directions for next studies.

Before introducing our model, derived using Stochastic Activity Networks (SAN) [7], we define the relevant figures of merit and the assumptions made in our study.

The effects of outages on the GPRS services have been studied through specific indicators of the QoS as perceived by end users, which are:

- the time necessary for the system to reach its steady-state behaviour, following the end of an outage; this measure is indicated as *recovery time*;
- a measure of the congestion induced by the outage in the system, as an indication of the service degradation, both during the outage and during the recovery time.

We chose to model the GPRS as a closed system with a constant number of users, who are forced to wait for the satisfaction of the requests they ask for. This choice is not realistic but conservative: it determines a higher load on the system, inducing a less favourable situation. If an outage lasts long enough, all users will be contending the resources and the system, when restarting, will have to face the highest overload possible.

Both indicators are then measured by observing the marking of the place representing the users which had their last request satisfied and are not trying to get another service. By comparing the number of tokens in the place *active* (see section 3.2) with the expected number of tokens contained in it when the system is in steady state, one can perceive the degradation of the service, whereas the recovery time is measured as the time necessary to reach the steady-state marking.

3.1 Assumptions

The model has been defined under the following assumptions concerning the configuration of the GPRS and the users behaviour:

1 only one cell has been taken into account, containing a constant number of users, whose contexts are permanently retained (therefore no "attach" and "detach" procedures, to register and delete users information respectively, are considered in our study);

2 all users belong to the same priority class and have the same request rate, i.e., they are indistinguishable from the point of view of generated traffic;
3 user requests fit in one LLC frame (1600 bytes) and, from the user's viewpoint, once a request has been made, he cannot abort it but has to wait until the service is provided;
4 the radio channel is considered faultless, meaning that no re-transmissions are necessary at the LLC and RLC levels. To keep consistent, the coding scheme considered is the CS-1, the most robust coding scheme among the four accounted for by the standard;
5 at most one radio frequency is devoted to the GPRS traffic (8 time slots);
6 only one MPDCH, for signalling and control information, is assumed, carrying 1, 2, or 4 PRACHs;
7 each traffic channel is allocated to a single user at a time, who will retain it until the completion of his data transmission; concurrent usage of traffic channels and multi-slot assignments to a single user are not considered;
8 it is not allowed to queue the request through an Access Grant Reservation. So in case no traffic channel is available, random access requests cannot be accepted. (This assumption will be released in section 6).

The first four assumptions have been made for the sake of model simplicity; relaxing them would not invalidate the modelling approach followed, but would add significant complexity to the derived model. The other assumptions are congruent with typical GPRS configurations which are being currently considered by suppliers.

3.2 The model

Figure 2 shows a first model, briefly explained in the following.

- Tokens in the place *active* represent those users that have sent successfully their up-link data. After some time, accounted for by the timed transition *to_req*, a user issues a new request and a token is moved from *active* to the place *new_req*.
- The block starting with the instantaneous activity *req* and ending with the input gate *control* represents the dynamics of the random access procedure. *req* states the maximum number of attempts a user is allowed to make in sending an Access Burst. It has one case for each possibility; the associated probabilities have been derived on the basis of the parameters M, P, S, T and the timer. Tokens in places *ready1, ..., ready8* represent the number of users allowed to make a maximum of 1,....,8 attempts, respectively. The instantaneous activities *check_p1,...,check_p8* model the persistence level. If the user passes the persistence level, he can send an Access Burst and moves into the place try_i, otherwise he moves into the correspondent place $fail_i$. Should a user consume all his assigned attempts to make his request, or should the time-out regulating the maximum allowed time for making a request (set to 5 sec) expire, the user is moved into the place *block*. A blocked user will do a new attempt after a time sampled from the timed activity *b_to_n*, having exponential rate and taking into account Automatic Retransmission Time (ART). The place *w5* and the activity *wait_5* take into account those users that haven't been assigned any attempt, because they will always fail the persistence level. According to the standard specification, they have to wait 5 seconds before moving in the place *block*.
- The instantaneous transition *check_capture* checks, stochastically, if there is a successful receipt of one Access Burst; if yes, a token is placed in *one_accepted*, otherwise in *all_discarded*. The instantaneous transition *who_is_passed* fires when there is a token in *one_accepted* and it allows to choose which level the accepted Access Burst comes from, placing a token in one of the places *p1,...,p8* (each Access Burst at each level has the same probability to be the accepted one). The input gate *control* and the activity *control_act* properly update the places recording the residual tries made available to the other concurrent requests (places *ready1,...,ready8, try1,...,try8, fail1,...,fail8, wait_a0,...,wait_a7* and *p1,...,p8*).
- When there is a successful receipt of one Access Burst and there is a free channel (that is at least a free pair between *ch1-a1,....,ch7-a7*), the output gate *choose_channel* puts a token in one of the places *ch1,...,ch7*. The timed activities *su1,...,su8* simulate the set-up time of a radio link to send user data. The timed activities *exp1,...,exp7* and *d1,...,d7* simulate the data send time. Their values depend on the scenario we consider (see Section 4).
- The subnet enclosing the timed activities *PRACH_available* and *slot_available*, and the places *en* and *enable*, models the multiframe on the MPDCH.
- The subnet including places *work, out_serv* and *ok* is used to represent the occurrence of outages, and the consequent repair of the system. A token in the place *work* represents the correct running of the network. The firing of the timed activity *outage* represents the occurrence of an outage. This activity has a deterministic time, chosen in such a way to model the occurrence of an outage after the network has already reached the steady-state. A token in the place *out_serv* represents the unavailability of the network. The output gate *control_fault* simulates the effect of the fault.

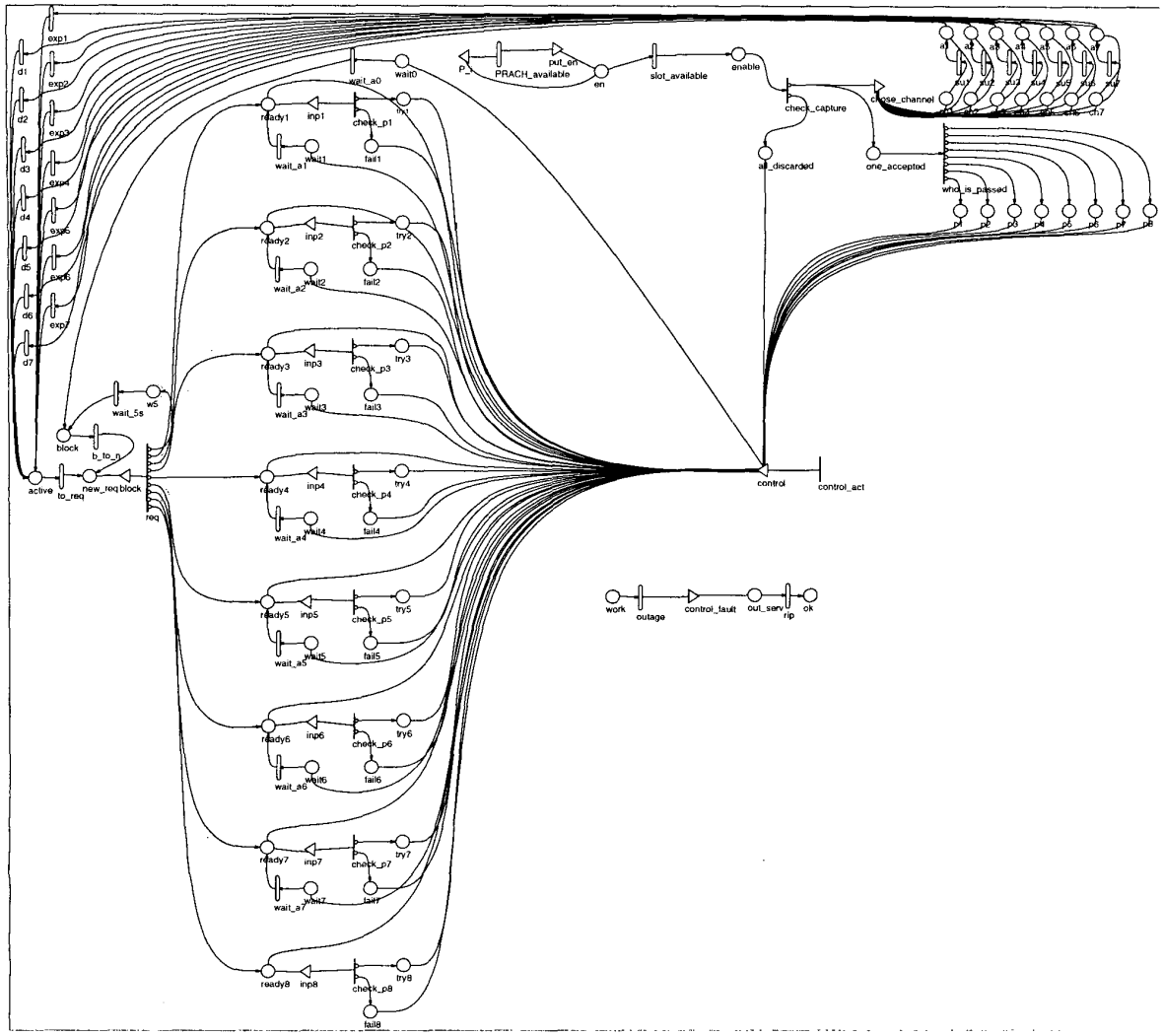

Figure 2: SAN model of the Random Access Procedure of the GPRS system

In our analysis, it consisted in the inhibition of the immediate activity *req* and the moving of the tokens of the whole net in the place *new_req*. This form of outage could be caused, for example, by a malfunction in the antenna of the cell itself, which therefore stops to send and receive signals. The time necessary to have all the tokens in *new_req* represents the outage duration to reach the worst system conditions (i.e., the maximum congestion). The timed activity *rip* represents the time necessary to repair the system and moves a token in the place *ok*.

4. Settings for the numerical evaluation

The model presented in Section 3 has been numerically solved by using the simulator provided by the UltraSAN tool [8]. The nature of the measures and the order of magnitude of the results we are looking for make a simulation approach appropriate for studying the system. At the same time, we could represent real system conditions better than analytical approaches do (we could choose distribution functions resembling the occurrence of specific phenomena, and not be forced to the exponential distribution).

We identified two representative scenarios to exercise the model on. They differ in workload characteristics and user behaviour, although the overall system load they offer to the system is the same. In both scenarios, we suppose that each user packet fits into one LLC frame. This way, only one random access is needed to send user data relative to a service request.

In the first scenario, users' data are generated using the Railway model described in the ETSI document for evaluation criteria [9]. This is an exponential negative distribution, truncated at the maximum length of 1000 bytes, and having an approximate average of 170 bytes. We suppose that users inter-request time, that is the time between two users requests for data transmission, is 10 seconds on average. This time, accounted for in the model by the transition *to_req*, includes the time needed to download the requested data and the time a user needs to make the next request, mainly based on examining the service provided at the previous request. This scenario seems to be adequate for typical Web browsing applications. In the second scenario, data traffic is generated using a uniform distribution in the interval [1000, 1600] bytes. Moreover, requests are less frequent than in the previous case (users inter-request time is 76.4 seconds on average), but each request involves more data to be transferred; short e-mails or filled forms in the WEB could be examples of applications fitting these characteristics.

Table 1 reports the model transitions that depend on the workload characterisation, together with their settings. For the sake of brevity, we do not explain how these values are derived; full details are in [11].

Transitions	Scenario 1	Scenario 2
exp1,...,exp7	Exponential with rate 1/(0.147712 sec.)	disabled
d1,...,d7	Deterministic with firing time 0.812416 sec.	Uniform within [0.812416 - 1.29248] sec.
to_req	Exponential with rate 1/(10 sec.)	Exponential with rate 1/(76.4 sec.)

Table 1: Transitions characterisation and settings in the two scenarios

Table 2 summarises the other main parameters involved in the evaluation, together with their assigned default values. A more comprehensive settings definition, including, e.g., the probabilities of the cases of the model transition *req*, is in [11].

Among the values specified in the ETSI document [4], we chose mean value for the parameters P, S and T, and the maximum allowable for M. Investigations to evaluate the sensitivity to variations of these parameters could be performed in next studies.

We dedicated one radio frequency to GPRS, meaning that up to 8 PDCHs can be allocated, one of which is always used as master (without possibility to carry users data). A previous work [12] has deeply investigated on proper configurations of PRACHs and S-PDCHs, which lead to pretty good expected behaviour of the system under "normal" system conditions. Here, we take advantage of this previous study in setting up PRACHs and S-PDCHs in the two considered scenarios (specifically, the choice has been made so as to keep the expected probability that a user request is not successful around the value of 0.1 or less).

Symbol	Description	Value
M	Max number of re-transmissions	7
P	Persistence Level value	7
S	TDMA frame number to next try	76
T	Spreading number to next attempt	14
ART	Automatic Retransmission Time	0.1 sec
M-PDCH	# of Master Packet Data CHannel	1
S-PDCH	# of Slave Packet Data CHannel	2,3
U_A	# of active users in the cell	60, 150
PRACH	# of Packet Random Access CHannel per multiframe	2, 4
t_out	outage duration	5-300 sec

Table 2: Relevant parameters and their default values

The two values for U_A (the number of users in the cell) considered in the subsequent analysis have been 60 and 150. The outage duration, t_out, is the other main varying parameter in our study, ranging in the interval of a few seconds to a few minutes ([5 - 300 sec]). Actually, we extended the variation of t_out till the maximum value of interest in our configurations, i.e., the value of outage duration which leads to the maximum congestion in the network, with no users in the place *active*.

5 Numerical Evaluation

This Section presents the numerical evaluation performed solving the SAN model in a transient period, extending from the outage occurrence to an interval of time following the system repair.

The analysis focuses on studying the number of users whose requests have been satisfied, that is by looking at the variations of the number of tokens in the place *active* in Figure 2. Average values (as determined by the steady-state analysis), indicative of the "normal" system conditions, have also been determined and used as reference values against which the effects of outages occurrences can be appreciated. In the next figures, such steady-state values have been enclosed in dotted lines at a distance of ± 3%.

At growing values of the outage duration t_out, more and more tokens move from *active* to *new_req*, meaning that more users are making new service requests. Some time is necessary to complete this process (which depends on how frequently active users issue new requests). Therefore

the longer is the outage, the higher the number of active users affected by the outage, yielding a higher degradation of the QoS.

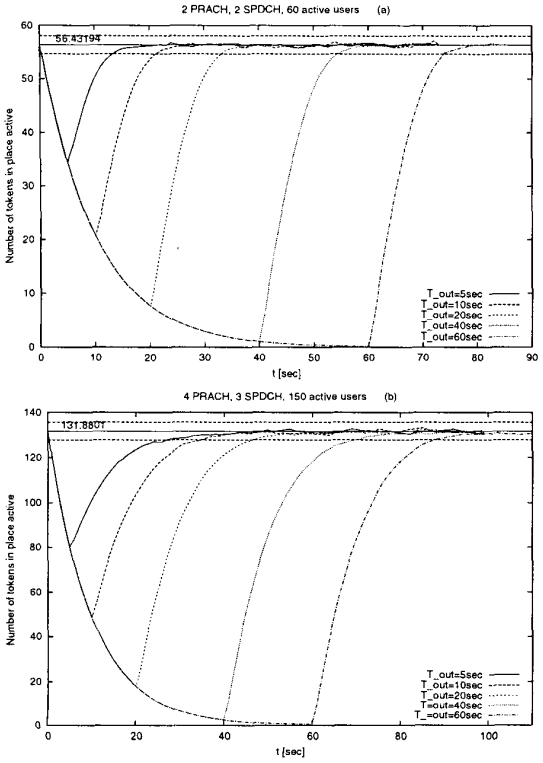

Figure 3: Scenario 1- Behaviour during outages and recovery, with 60 users (a) and 150 users (b)

Figures 3.(a) and 3.(b) relate to scenario 1 and plot the number of tokens in place *active* at different times, for 60 and 150 active users, respectively; they give a direct indication of both the QoS degradation during the outage and the time necessary to bring the system back to the expected behaviour after the outage. In both of them, several curves have been plotted for different values of t_out, the duration of the outage. The decreasing line on the left of the two figures traces the variations of tokens in the place *active* from the instant when the outage occurs (time 0). It can be observed that the decrease rate is highly dependent on the number of users in *active*; however, the time to empty this place is the same in both cases. This is not surprising, since users issue requests with the same rate. Then, the increasing lines show the time necessary for obtaining the expected number of tokens in the place *active* (determined by a steady state analysis under normal system conditions and indicated by the upper horizontal line), for different durations of the outage. It can be noted that the recovery time does not change much for the different values of t_out (slightly lower values are observed for lower t_out), while it appears to be more sensitive to the number of active users.

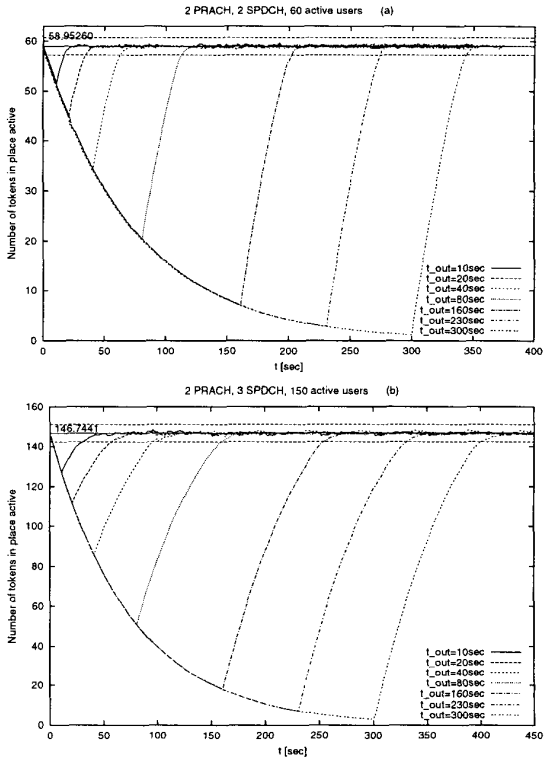

Figure 4: Scenario 2- Behaviour during outages and recovery, with 60 users (a) and 150 users (b)

Figure 4 contains the same information when scenario 2 is applied. It can be immediately observed that the degradation time is much longer than previously due to the different behaviour of the users (the users inter-request time is higher). The recovery time is also much longer compared to scenario 1, since the accumulated workload is higher. Moreover, in this case it appears to be more sensitive to the outage duration. Summarising, in this scenario the system degradation is slower than in scenario 1, and also the time necessary to come back to the "normal" working conditions is longer than before.

To better appreciate the influence of the outage duration on the resulting recovery time in scenario 2, Figure 5 points out the time to reach the steady state vs. the duration of outage for the two configurations of Figure 4. It can be observed that the recovery time varies significantly for low values of the t_out, becoming almost independent from it when the outage duration gets high values (in the

Figure, greater than 250 seconds). In fact, when outages last this long, almost all users have started new requests (place *active* is almost empty, as Figure 4 shows) and the network has practically reached its worst case overload. From this point on the recovery time cannot get worse.

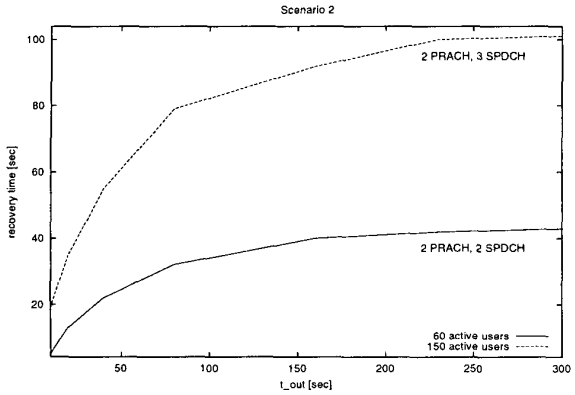

Figure 5: Time to steady state vs. outage duration - scenario 2.

6 Extended model

In this Section, we extend the model presented in Section 3 by accounting for queuing procedures, a feature included by the GPRS standard to improve system performance.

Collisions reduce channel occupation and increment both the blocking probability and the access delay. To reduce the effect of collisions, a queue of pending channel requests can be introduced. When an access burst is correctly received by the network, but no traffic channels are available, the network can send a "packet queuing notification" message to notify to the mobile station that its channel request has been correctly received and that a "packet up-link" assignment message will be sent as soon as possible. In this way, a new access burst from the same mobile station is avoided, thus reducing the collision probability for other users requests. According to the standard, when a user receives an "access grant" message, a timer (T3162) of 5 seconds is set; if no "packet up-link" assignment message is received by the timer expiration, the mobile station has to restart with a new request for channel reservation.

We retain the same assumptions listed in subsection 3.1, except for assumption 8, which is relaxed here.

6.1 Introducing the queuing mechanism

To account for the queue mechanism, new elements have been added to the model in Figure 2. Instead of using an infinite queue and a timer to limit the presence of a user in the queue as prescribed by the standard, we used a finite queue, whose length has been appropriately chosen so that a user in the queue never waits more than 5 seconds to have the channel assigned. Then, users which do not fit in the queue have to proceed with a new channel request.

Figure 6 shows the subnet of the model in Figure 2 where the queue has been introduced. The main modifications on the previous model are the following:

- A token in the place *queue* represents a pending request waiting for up-link channel reservation.
- The immediate transition *check_capture* has been modified: a correctly received request is refused only if the queue is full and there is no available traffic channel.
- The input gate *choose_channel* puts a token in the place *queue* if all the available channels are busy.
- The immediate transition *q_control_a* fires when a channel is released and there are pending requests in the queue. When transition *q_control_a* fires, the input gate *q_control* moves a token from *queue* to a place *chn* (*ch1,ch2,..ch7*), corresponding to the available channel.

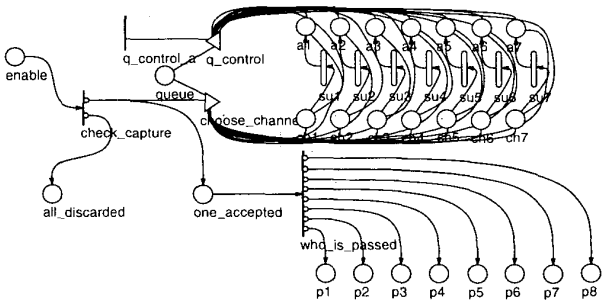

Figure 6: SAN of the queue mechanism

The maximum queue length, so as to allow users to remain in the queue no longer than 5 sec, has been determined separately for the two scenarios, since it depends on the service characteristics provided. The conservative assumption that users keep a channel busy for the time necessary to send the maximum allowed data packet length has been made. The resulting values for the queue length are 4 in scenario 1 and 2 in scenario 2 (to be multiplied by the number of available data channels). Again, more details are in [11], where it is also shown that such length for the queue is sufficient and no more significant improvements are obtained with longer queues.

6.2 Numerical evaluation

To determine the impact of the queue mechanism on the system behaviour, similar evaluations have been performed and compared with those previously discussed. The same two scenarios are therefore exercised; however, the system configurations in terms of PRACHs and S-PDCHs have been varied to better appreciate under which conditions the same system performance can be obtained when adopting the queue mechanism.

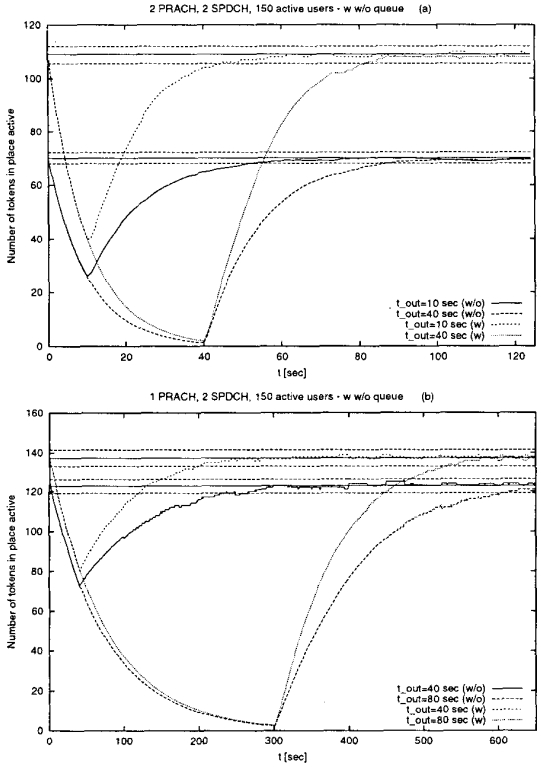

Figure 7: Effects of outages when accounting for the queue mechanism - scenario 1 (a) and scenario 2 (b)

Figures 7.(a) and 7.(b), relative to scenario 1 and scenario 2 respectively, compare the system behaviour, in presence of outage, with configurations employing and not employing the queue mechanism. Two couples of plots are shown in each figure, obtained varying the outage duration; t_out has been assigned a low and a high value with respect to significant outage duration in the two respective scenarios (note that the maximum outage duration to bring the system at its maximum congestion level, i.e., the time to empty the place *active*, is the same for both systems, with and without the queue, since it only depends on the number of U_A and on the users inter-request time). As a first observation, employing the queue leads to an increase in performance, quantified by the higher average marking of the place active, represented by the horizontal lines (again, enclosed by two dotted lines at a distance of ± 3%). This result, more relevant in case of scenario 1, confirms the benefits expected from the queue mechanism.

The rate with which tokens exit the place active during the outage (left descending curves) is, of course, higher for the configuration using the queue, since the average marking of this place (that is, at the time the outage starts) is higher for this configuration. However, the recovery time, necessary to bring the system to its average operational level after the end of the outage, is comparable for the two couples of plots in scenario 1 (Figure 7.(a)), and more favourable to the configuration with the queue in scenario 2 (Figure 7.(b)).

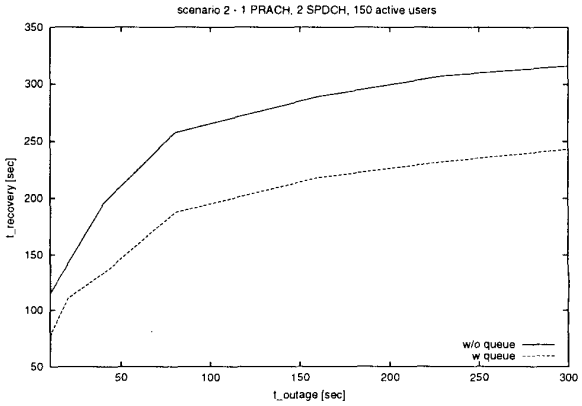

Figure 8: The effect of the queue on the recovery time at varying outage duration - scenario 2

Figure 8 completes our study. It shows the recovery time as a function of the outage duration in scenario 2, for two configurations which only differ for the queue mechanism. The better behaviour exhibited by the configuration which uses the queue, already pointed out in the previous figure 7.(b), is now extremely clear. For the same phenomenon already discussed when commenting Figure 5, the sensitivity of the recovery time to t_out is especially noticeable in the left part of the figure (low values of t_out).

Summarising, the results shown by this analysis re-enforces the benefits of adopting the queue mechanism, also as a measure to better cope with outage occurrences.

7. Conclusions

This paper has presented a study on modelling and analysis of the GPRS behaviour. The study has been focused on the critical Access Random Procedure of GPRS, where users compete for the channel reservation. An estimation of the effect of outages on the QoS as perceived by users has been performed. A first model has been derived and analysed, and a refinement of it has been proposed in a second step, by introducing a queue mechanism which significantly improves system performance.

Outages constitute a dependability-critical condition for communication systems as the GPRS, implying unavailability of the system and a consequent accumulation of users asking for the service. As a consequence, as soon as the GPRS comes back in operation, the high level of requests leads to a higher probability of collisions. We have therefore analysed the system during and in an interval after outages, to understand the degradation of the QoS and the time necessary to restore the steady-state behaviour, which has been in turn properly evaluated. Two representative scenarios characterised by different workloads have been selected and evaluated under varying values of the most relevant system parameters. The obtained results show very useful in devising GPRS configurations adequate to maintain an acceptable QoS also when the system is degraded, and to understand the availability bottlenecks.

Finally, several extensions of the study performed here have been devised. Among others, particularly interesting are: i) further extensions of the model of the GPRS accounting for other features allowed by current versions of the GPRS, e.g., the possibility to have concurrent usage of traffic channels and that of multi-slot assignments to a single user; ii) investigations on other forms of outages, caused by malfunctions in different system components, whose impact on the overall system degradation might be different; iii) to account for more sophisticated users behaviours; iv) to extend the analysis to other critical parts of the GPRS, such as bottlenecks in up-link at points where traffic from several cells converge and/or contention for getting down-link response. With respect to this last research direction, we have actually defined in this study a model structure to analyse random access procedures for resources contention. Therefore, to analyse other parts of the GPRS system where such contention problem arises, this model structure can be taken as the basic starting point, to be of course adjusted/extended to account for the specificity of the problem at hand.

References

[1] G. Brasche and B. Walke, "Concepts, Services and Protocols of the New GSM Phase 2+ General Packet Radio Service," *IEEE Communications Magazine*, August 1997, pp. 94-104.

[2] J. Cai and D. J. Goodman, "General Packet Radio Service in GSM," *IEEE Communication Magazine*, vol. October 1997, pp. 122-131, 1997.

[3] ETSI, "Digital Cellular Telecommunications System (Phase 2+); General Packet Radio Service (GPRS); Service Description; Stage 2," GSM 03.60 version 7.1.0 Release 1998

[4] ETSI, "Digital Cellular Telecommunication System (Phase 2+); General Packet Radio Service (GPRS); Mobile Station (MS) - Base Station System (BSS) Interface; Radio Link Control/Medium Access Control (RLC/MAC) Protocol," GSM 04.60 version 8.3.0 release 1999.

[5] C. Ferrer and M. Oliver, "Overview and Capacity of the GPRS," *9th IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications*, Boston, 1998, pp. 106-110.

[6] J. C. Laprie, "Dependability - Its Attributes, Impairments and Means," in *Predictably Dependable Computing Systems*, J.-C. L. B. Randell, H. Kopetz, B. Littlewood (Eds.), Springer Verlag, 1995, pp. 3-24.

[7] W. H. Sanders and J. F. Meyer, "A Unified Approach for Specifying Measures of Performance, Dependability and Performability," in *Dependable Computing for Critical Applications, Vol. 4 of Dependable Computing and Fault-Tolerant Systems*, (Eds..A. Avizienis, H. Kopetz, and J. Laprie), Springer Verlag, 1991, pp. 215-237.

[8] W. H. Sanders, W. D. Obal, M. A. Qureshi and F. K. Widjanarko, "The UltraSAN Modeling Environment," *Performance Evaluation Journal, special issue on Performance Modeling Tools*, vol. 24, pp. 89-115, 1995.

[9] GPRS Ad-Hoc ETSI/STG SMG2, "Evaluation Criteria for the GPRS Radio Channel," ETSI, 1996.

[10] P. Taaghol, R. Tafazolli and B. G. Evans, "An Air Interface Solution for Multi-rate General Packet Radio Service for GSM/DCS," *IEEE Vehicular Technology Conference VTC 97*, 1997, pp. 1263-1267.

[11] F. Tataranni, S. Porcarelli, F. Di Giandomenico and A. Bondavalli, "Modeling and Evaluation of the Effects of Outages on the Quality of Service of GPRS Systems," CNUCE-CNR Technical Report B4-2000-028, December 2000, also available on the web at: http://bonda.cnuce.cnr.it/Documentation/Reports/Techreports

[12] F. Tataranni, S. Porcarelli, F. Di Giandomenico, A. Bondavalli and L. Simoncini, "Modeling and Analysis of the Behavior of GPRS Systems," *IEEE 6th International Workshop on Object-oriented Real-Time Dependable Systems*, Roma, Italy, January 2001.

Acknowledgements

This work has been partly performed in the framework of a cooperation between CNUCE/CNR and the Modelling and Simulation Team at the Motorola Technology Center Italy in Torino.