Accurate Availability Estimation of GPRS

Stefano Porcarelli[1], Felicita Di Giandomenico[2], Andrea Bondavalli[3], Massimo Barbera, Ivan Mura[4]

[1] Computer Engineering Department, University of Pisa, Via Diotisalvi 2, 56126 Pisa, Italy
stefano.porcarelli@guest.cnuce.cnr.it

[2] Istituto di Elaborazione della Informazione, CNR,Via G. Moruzzi 1, 56124 Pisa, Italy
digiandomenico@iei.pi.cnr.it

[3] Dipartimento di Sistemi e Informatica, University of Florence, Via Lombroso 6/17, 50134 Florence, Italy
andrea.bondavalli@cnuce.cnr.it

[4] Motorola Technology Center Italy, Via C. Massaia 83, 10147, Torino, ITALY
{Ivan.Mura, Massimo.Barbera}@motorola.com

Abstract

The General Packet Radio Service (GPRS) extends the Global System Mobile Communication (GSM), by introducing a packet-switched transmission service. This paper analyses the GPRS behavior under critical conditions. In particular, we focus on outages, which significantly impact the GPRS dependability. In fact, during outage periods the number of users trying to access the service grows proportionally over time. When the system resumes its operations, the overload caused by accumulated users determines a higher probability of collisions on resources assignment and therefore a degradation of the overall QoS. This paper adopts a Stochastic Activity Network modeling approach for evaluating the dependability of a GPRS network under outage conditions. The major contribution of this study lies in the novel perspective the dependability study is framed in. Starting from a quite classical availability analysis, the network dependability figures are incorporated into a very detailed service model that is used to analyze the overload effect that GPRS has to face after outages, gaining deep insights on its impact on user's perceived QoS. The result of this modeling is an accurate availability analysis, which takes into account not only the bare estimation of unavailability periods, but also the important congestion phenomenon following outages that contribute to service degradation for a certain period of time after operations resume.

Keywords: Availability, Outages, GPRS, Modeling and Evaluation, Stochastic Activity Networks, Simulation

Categories and Subject Descriptor: B.8.2 **[Performance and Reliability]:** Performance Analysis and Design Aids; C.4 **[Performance of Systems]:** Performance attributes; C.4 **[Performance of Systems]:** Reliability, availability, and serviceability.

1. Introduction

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

Work on GPRS started in 1994, and a standardization of the GPRS specification has been recently frozen by ETSI (European Telecommunications Standard Institute). Analyses of the GPRS expected behavior have been performed, essentially focusing on measures like throughput, delay for the end-to-end frame transmission, average number of attempts necessary to win contention on the random access channel (e.g., [1], [2], [5], [10]).

As it is for most telecommunication systems, and particularly for its “always connected” vision, GPRS classifies as an availability-critical system and a significant effort is thus being devoted by systems manufacturers to improve availability. 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 operational 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 requests accepted as soon as the system is up again. At system restart, the GPRS must face an overload; 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 get over the congestion before getting back to the “normal” behavior, in which the nominal QoS is provided to users.

This work contributes to the analysis of GPRS by providing a modeling approach suitable for investigating on the effects of outage periods on the service provision, with special attention on the user perception of the QoS. Two different levels of modeling are considered. The first one defines a GPRS *network* availability model, which focuses on the dependability of the various components of the GPRS infrastructure, while the other one defines a GPRS *service* dependability model, which builds upon the network availability model. The latter model is obtained from the model presented in [11]. It takes as input the detailed stochastic characterization of the outages that is obtained from the GPRS network availability model, and maps on it typical service requests pattern of GPRS applications. The resulting composed model allows bridging the gap between the classical network perspective that is commonly taken when studying the availability of telecommunications systems, and a user and application centric analysis of the dependability of services that can be provided through the packet data service of GPRS.

Both the two levels of modeling are built by resorting to the powerful modeling capabilities of Stochastic Activity Networks. Models are solved by simulation using the UltraSAN tool [8]. The tool allows defining in a very natural way the two-level hierarchy of models that compose the GPRS overall model. Indeed, the modular modeling approach of the tool facilitates the composition of the sub-models developed in the two levels of modeling.

This paper is structured as follows. In Section 2 we first introduce the aspects of the GPRS system that are relevant for the sake of the modeling and analysis conducted in the subsequent sections. We focus on the description of the architectural elements composing the Radio Access Network and the Core Network of GPRS, and we treat in specific detail the ALOHA random access procedure that plays a major role in the period immediately after an outage. Then, we present in Section 3 and Section 4 the GPRS network availability and service dependability model (together with a brief availability evaluation), respectively. The overall structure of the GPRS composed model, which represents the combination of the two modeling levels, is introduced in Section 5. Section 5 contains also the results of an evaluation campaign, aiming at quantitatively estimating the potential effects of outages on GPRS service dependability and at performing a sensitivity analysis to the most relevant parameters. Conclusions are in Section 6.

2. GPRS overview

The introduction of GPRS is a first step towards the full deployment of packet-data wireless networks. The use of the GSM circuit-switched transmission mode with data traffic, typically characterized by frequent alternation between activity and idle periods of the data source, results in an inefficient use of the scarce radio resources. In fact, in circuit switching allocation mechanisms, with high set-up time as in GSM, it is necessary to allocate a channel to a Mobile Station (MS) for all its transmission time without taking into account its real activity during this time. 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 make it possible for 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].

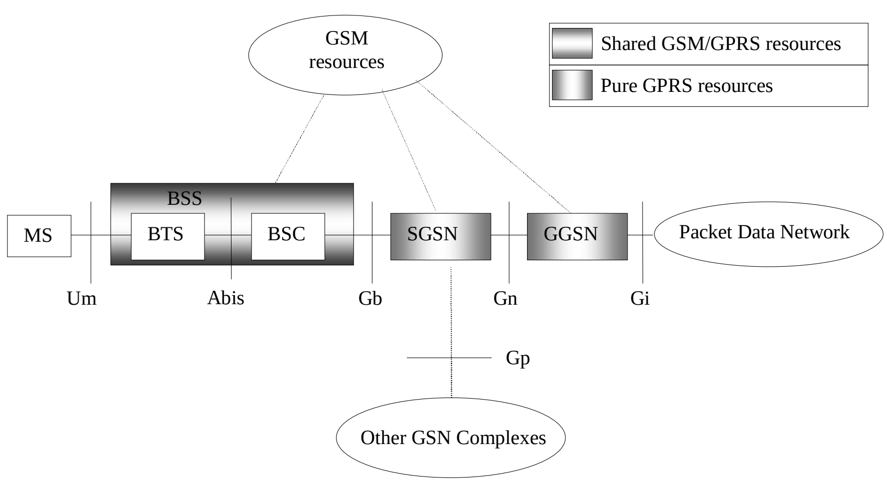

Figure 1: GPRS architecture

Figure 1 shows a snapshot of the GPRS system architecture. Besides the standard Base Tranceiver Station (BTS) and Base Station Controller (BSC) components already present in GSM, GPRS will include the two new network elements shown in Figure 1, which do not appear in the current GSM architecture, and which are collectively known as GPRS Support Nodes (GSN):

- The *Serving GPRS Support Node* (SGSN), at the same hierarchical level as the GSM component Mobile Switching Center (MSC), which keeps track of the individual MSs' location and performs security functions and access control. The SGSN is connected to the Base Station Subsystem (BSS) with Frame Relay;

- The *Gateway GPRS Support node* (GGSN), which provides inter-working with external packet-switched networks, and is connected with SGSN via an IP-based GPRS backbone network.

In order to access GPRS services, an MS first makes its presence known to the network by performing a GPRS attach. This operation establishes a logical link between the MS and the SGSN, and makes the MS available for SMS via GPRS, paging via SGSN and notification of incoming GPRS data. To send and receive GPRS data, the MS activates the packet data address it wants to use. This operation makes the MS known in the corresponding GGSN, and inter-working with external data networks can commence. User data is transferred transparently between MS and external networks with encapsulation and tunneling. Notice that the BSS is shared between GPRS and GSM network elements, to maintain compatibility and to keep low the investments needed to introduce the GPRS service. In fact, as far as the BSS subsystem is considered, the introduction of GPRS over an existent GSM network only requires a software upgrade.

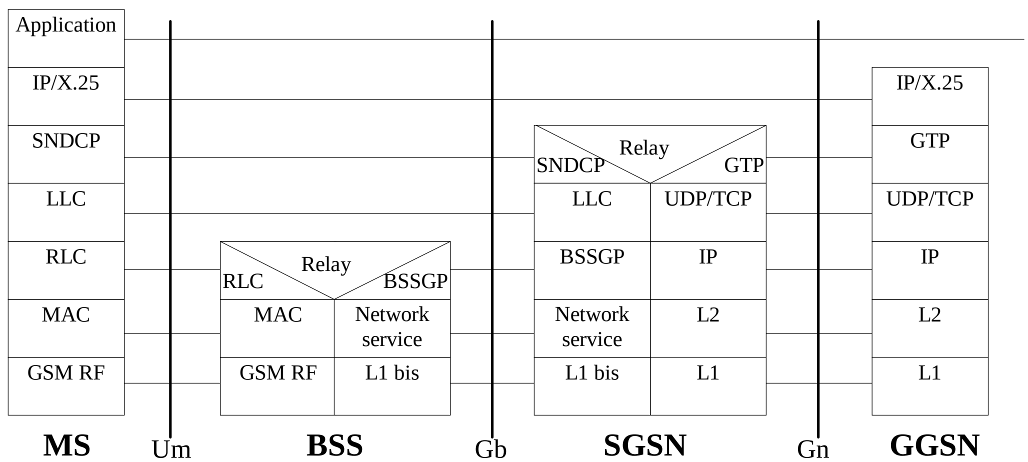

Figure 2: The ISO/OSI structure of GPRS

The protocol pile of GPRS is shown in Figure 2. 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, which is responsible for moving user data between MSs and the network. The Radio Link Control (RLC) layer allows transmitting data across the air interface. 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 MS. 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 by 8 time slots). In a cell that directly supports GPRS, a Master PDCH is allocated, to provide control and signaling information to start data transfer in both uplink and downlink directions, and to handle users mobility.

Among the logical channels that share the MPDCH, we focus on a specific channel dedicated to the uplink 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 through the PRACH. Since the network does not control the PRACH usage, the access method, based on a *Random Access Procedure*, may cause collisions among requests by different MSs. As it is observed during massive congestion events, such as New Year's Eve, the blocking on the PRACH may become a bottleneck of the system. This is the specific aspect of the 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 retransmissions 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 for a time, dependent on S and T. If it does not receive the Packet Down-link Assignment (or a Packet Queuing) a new attempt is tried, if it is still allowed to make one, otherwise a failure is notified to the upper layer. From parameters S and T, the MS also determines the next TDMA frame in which a new attempt is possible, should the previous one be unsuccessful and a new attempt still allowed.

Under normal workload conditions, this retry mechanism is able to make the MS request to reach the BSS subsystem with a very high probability. Once the MS request successfully reaches the BSS subsystem, traffic packet data channels, called slave PDCH, are allocated if available in the cell to transport users data and transmission signaling. For what concerns data transfer, up-link and downlink channels allocation is completely independent and a MS can operate uplink and downlink data transfer simultaneously. Should the selected cell be not immediately able to allocate the PDCHs, the MS request may be put in a queue to wait for the first available resources. In case the request cannot be accommodated, a reject message is sent to the MS.

3. Availability modeling of GPRS

This section introduces the network availability model of GPRS. It presents an architectural view of GPRS availability, by building a model that represents the failure/repair behavior of the end-to-end transmission path from the MS up to the external IP networks.

3.1. System availability related aspects

The availability of the overall GPRS system basically depends on the availability of the MS, BTS, BSC, SGSN and GGSN network elements; i.e., the entities providing the service in the external packet data networks, plus some other GSM network components, used to authenticate and localize the mobile user, such as the Home Location Register (HLR) and Visitors Location Register (VLR) databases. Moreover, the availability of the communication links between these units is also necessary for the system to be available from a user perception. These communication links include:

- Radio links between the MS and the BTS across the Um interface;
- E1 spans between a BTS and a BSC supporting the Abis interface;
- E1 spans from a BSC to a SGSN providing Gb interface;

- 100 BaseTs between SGSNs and GGSNs supporting Gn interface;
- E1 spans between SGSN and IP Backbone supporting the Gp interface;
- E1 spans between GGSN and IP Internet supporting the Gi interface.

To limit the scope of the modeling, we will consider one single GPRS cell, and do not take into consideration the mobility of users. These restrictions will limit the number of users that need to be taken into consideration. Moreover, the following assumptions have been made for the sake of building a manageable availability model of GPRS.

- The Mobile Station is fault free;
- The reliability of the radio links will not be taken into consideration in this model;
- The pure GSM network elements, such as HLR and VLR, are fault free, as well as all the links supporting the interface towards them;
- The external packet data network (Internet) is fault free, as well as the links supporting the Gi interface;
- The Gp interface is not supported.

Therefore, our availability model will consider the path shown in Figure 3 from the MS up to the GGSN, taking into consideration the reliability and the repairs of the networks elements along the path, as well as the links supporting the communications across the depicted interfaces.

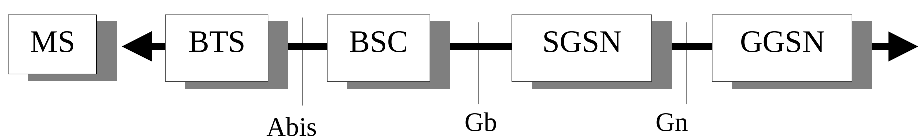

Figure 3: GPRS architecture under analysis

It is worthwhile remarking that several of the elements determining the network availability of GPRS are in fact redundant components. We will shortly describe in the following the various types of redundancy that may be deployed in the architecture of the system.

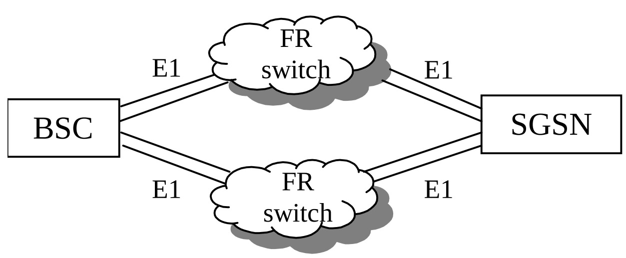

Figure 4: Redundant links between BSC and SGSN

The Gb interface between the BSC and the SGSN is deployed in a redundant fashion, with two Frame Relay switches as shown in Figure 4. This is done to tolerate the outages typically caused by network-independent events (links cut). Only one of the links is actively working, whereas the second one is a spare, ready to be put into operation in case the first link stops working.

Both the SGSN and the GGSN network elements are deployed in a redundant N+M configuration, in which N modules are actively working in a load-sharing mode, and M modules are held in a cold spare state. In case of failure of one or more active modules, the spare modules are enabled to switch over and take the role of active modules.

Failed elements or links are mainly repaired through replacement. It is important observing that links repair time is usually among the major contributors to telecommunications systems downtime.

3.2. Model definition

We build the availability model of GPRS by using a Stochastic Activity Network approach [7]. For each of the network elements introduced in the previous section, we will define a separate SAN subnet. Joining the various subnets through the JOIN operator of UltraSAN allows defining in quite a natural way the overall GPRS availability model.

3.2.1. SAN subnets of the non redundant elements

The simple SAN subnet in Figure 5 models the availability of the BTS network component. This same subnet (apart from obvious changes in the SAN elements) will be used to model all the other network elements for which no redundant deployment is foreseen, e.g. BSC, Abis E1 link, Gn E1 link.

The initial marking of place *BTS_up* is equal to 1, meaning that the modeled component is available in the initial state. Firing of activity *BTS_fail* represents a failure of the component. The firing time of *BTS_fail* follows a negative exponential distribution. At firing time, the token is removed from place *BTS_up* and put into place *BTS_down*, representing the unavailability of the component. Firing of *BTS_repair* represents the repair completion of the component, which occurs after a deterministic amount of time after the failure. The token is moved back to the *BTS_up* place when *BTS_repair* fires.

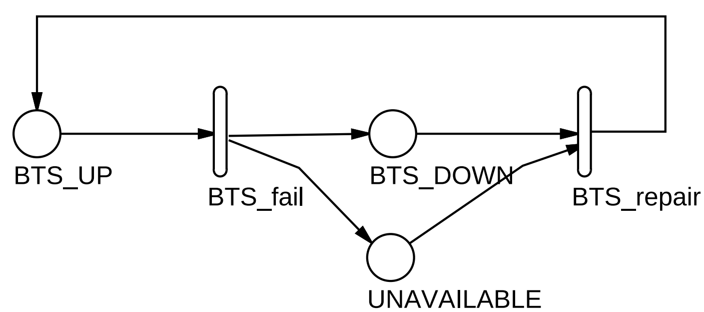

Figure 5: SAN model of the BTS

Place *UNAVAILABLE* is used to keep track of the availability of the BTS component together with that of the other GPRS components. In fact, for the non-redundant models, the *UNAVAILABLE* place is completely redundant, its marking being always equal to that of the *down* place of the subnet. The usefulness of this place will be explained in the following.

The firing time distributions and the parameters used to instanciate the distributions are shown in Table 1.

Activity	*Distribution*	*Parameter values*
BTS_fail	Exponential	rate → 1/*MTTF_BTS*
BTS_repair	Deterministic	value → *MTTR_BTS*

Table 1: Activity distributions for the BTS model

3.2.2. SAN subnet of the redundant Frame Relay links

The SAN subnet showed in Figure 6 models the availability of the redundant links across the Gb interface. The SAN in Figure 6 is quite similar to the one presented in Figure 5. The difference is found in the number of tokens that circulate in the model, which is equal to two to represent the double redundant Frame Relay link. Only one Frame Relay is active at the time. The active link fails according to the exponential failure rate of activity *GbLINK_fail*. When the active link fails, the second one becomes active and therefore subject to fail. To model that, each firing of *GbLINK_fail* removes only one token from place *GbLINK_up* and puts it into place *GbLINK_down*. Activity *GbLINK_repair* fires with a deterministic time. A repair removes all token present in place *GbLINK_down* and puts them back into place *GbLINK_up* through the output gate *GbLINK_repair_all*.

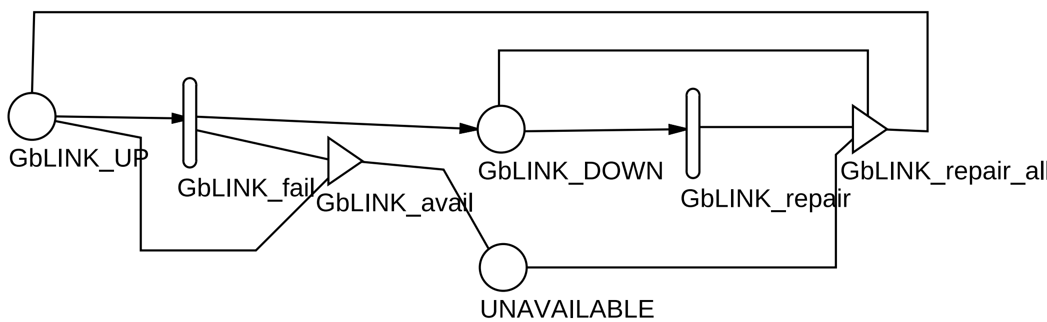

Figure 6: SAN model of the redundant Gb link

The output gate *GbLINK_avail* puts a token in place UNAVAILABLE when no more Frame Relay connections are available. The token is removed when the subnet has two tokens in place *GbLINK_down*, and the repair gets completed.

The subnet distributions and their parameters are listed in Table 2. The definition of the function executed in the output gate of the model is given in Table 3.

Activity	*Distribution*	*Parameter values*
GbLINK_fail	Exponential	rate → *1/MTTF_GbLINK*
GbLINK_repair	Deterministic	value → *MTTR_GbLINK*

Table 2: Activity distributions for the GbLINK model

Gate	*Type*	*Definition*
GbLINK_avail	Output	If (MARK(GbLINK_up)==0) MARK(UNAVAILABLE)++;
GbLINK_repair_all	output	MARK(GbLINK_up)+=MARK(GbLINK_down); if (MARK(GbLINK_down)==2) MARK(UNAVAILABLE)- -; MARK(GbLINK_down)=0;

Table 3: Output gate definition for the GbLINK subnet

3.2.3. SAN subnet of the SGSN and GGSN redundant network elements

The SAN subnet in Figure 7 models the SGSN component, according to the redundancy management description given in the previous subsection. The SAN model for the GGSN is the same (apart from a renaming of the subnet elements) as that of the SGSN in Figure 7.

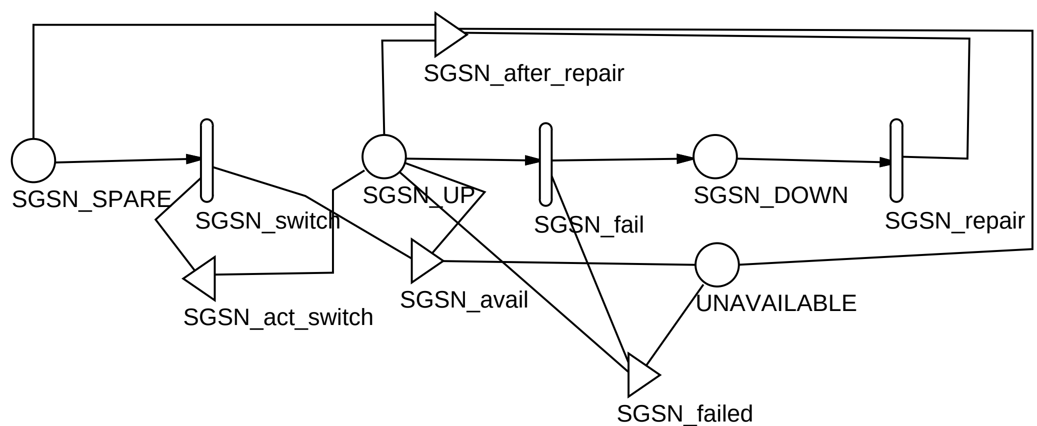

Figure 7: SAN Model of SGSN

Exactly N and M tokens are put in place *SGSN_up* and *SGSN_spare* to initialize the subnet. Activity *SGSN_fail* models the failure of one active SGSN component. At firing time, one token is removed from place *SGSN_up* and put into place *SGSN_down*. As soon as the number of tokens in place *SGSN_up* becomes less than N, activity *SGSN_switch* gets enabled through input gate *SGSN_act_switch*. Firing of activity *SGSN_switch* moves one token from *SGSN_spare* to *SGSN_up*, representing the activation of a spare component. The firing of *SGSN_repair* models the completion of a repair activity. At firing time, output gate *SGSN_after_repair* puts a token in place *SGSN_up* or *SGSN_spare* depending on the current SGSN model marking. The token is put into place *SGSN_up* if and only if its marking is less than N (i.e. more active SGSN are needed), or in place *SGSN_down* otherwise (i.e. the newly repaired component becomes a spare one).

A token is put in place UNAVAILABLE whenever one active SGSN fails. Tokens are removed from UNAVAILABLE by the output gates *SGSN_avail* and *SGSN_after_repair*, according to the rules specified in Table 5. The rationale behind this token game is that the token remains in place UNAVAILABLE only during the switch-on time of spare components, or when no more spares can take over and the SGSN has to wait for the repair.

The full definition of distributions and gates for the subnet is given in Table 4 and Table 5, respectively.

Activity	*Distribution*	*Parameter values*
SGSN_fail	Exponential	Rate→ *MARK(SGSN_up)/MTTF_SGSN*
SGSN_repair	Deterministic	Value → *MTTR_SGSN*
SGSN_switch	Exponential	Rate→*(N-MARK(SGSN_up))/MTTS_SGSN*

Table 4: Activity distributions for SAN model of SGSN

Gate	Type	Definition
SGSN_act_switch	input	enable iff MARK(SGSN_up)<N
SGSN_avail	output	if (MARK(SGSN_up)==(N-1)) MARK(UNAVAILABLE) - -;
SGSN_failed	output	if (MARK(SGSN_up)==(N-1)) MARK(UNAVAILABLE) ++;

SGSN_after_repair	output	if (MARK(SGSN_up)==(N-1))
		MARK(UNAVAILABLE) - -;
		if (MARK(SGSN_up)<N)
		MARK$(SGSN_up)++;
		else MARK(SGSN_spare)++;

Table 5: Gate definitions for the SGSN subnet

3.2.4. UltraSAN composed model

The overall network availability model is obtained from the subnets defined above by joining them with the SAN Join operator. The UltraSAN composed model is shown in Figure 8. All the subnets are joined over the common place UNAVAILABLE. This implies that a single UNAVAILABLE place will exist in the model, with all subnets using that same place. As a result, the number of tokens in place UNAVAILABLE will represent the number of unavailable network elements of the GPRS system.

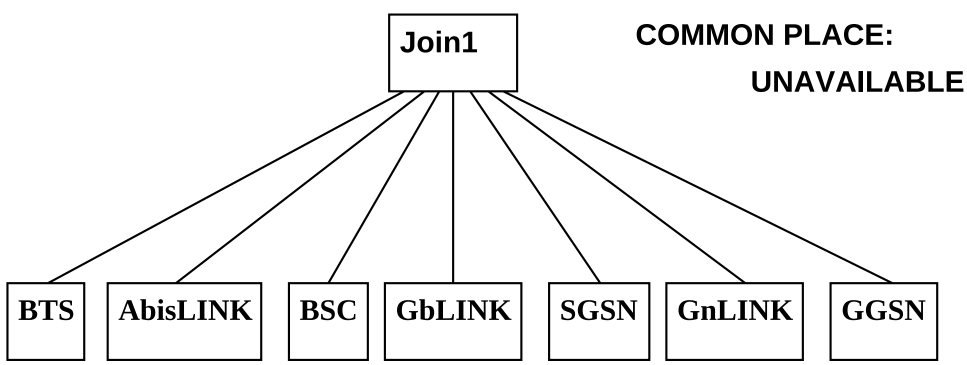

Figure 8: Composed model for the network availability model

The advantage of this definition of the common place is the easiness in the definition of the availability measures. Indeed, it is enough to check the marking of place UNAVAILABLE to verify whether the GPRS infrastructure is available or not. This same place will be exported to the upper-level model of service availability.

3.3. Evaluation of the Network Availability Model

We conduct now a numerical evaluation of the availability of the GPRS infrastructure, by solving the composed network availability model of Figure 8. A simulation approach is adopted, using the simulator offered by the UltraSAN tool [8]; simulation is in fact appropriate for our evaluation purposes, given the nature of the measures and the order of magnitude of the searched results. At the same time, it allowed to represent real system conditions better than analytical approaches do (e.g., by choosing distribution functions resembling the occurrence of specific phenomena, and not be forced to the exponential distribution).

In this experiment, as well as in subsequent numerical evaluations in Sections 4 and 5, results have been determined with 95% confidence interval and a relative confidence interval lower than 1%.

Table 6 summarizes the main parameters involved in the analysis, together with the numerical values assigned to them in our experiments.

Parameter	**Value**	**Parameter**	**Value**
N	5	MTTR_AbisLINK	4 hours
M	1	MTTR_BSC	1 hour
MTTF_AbisLINK	1 year	MTTR_BTS	2 hours
MTTF_BSC	6 months	MTTR_GGSN	1 hour
MTTF_BTS	4 months	MTTR_GbLINK	4 hours
MTTF_GGSN	6 months	MTTR_GnLINK	1 hour
MTTF_GbLINK	1 year	MTTR_SGSN	1 hour
MTTF_GnLINK	1 year	MTTS_GGSN	5 seconds
MTTF_SGSN	6 months	MTTS_SGSN	5 seconds

Table 6: Settings for numerical evaluation

Figure 9(a) shows the distribution of the outage duration. Since the repair times of faulty components are deterministic, there are only four possible values of outage duration spanning from seconds to hours (see Table 6). From the figure, the most probable outage has a duration of 5 seconds, which corresponds to the Mean Time to Spare (MTTS) of SGSN and GGSN components. Figure 9(b) shows the availability at varying values of the Mean Time to Failure (MTTF) and Mean Time to Repair (MTTR) of hardware components listed in Figure 8. Let K_F and K_R be two constants representing a scaling factor of MTTF and MTTR, respectively. The availability implied by $K_F = K_R = 1$ is that in case Figure 9(a). Setting $K_F > 1$ implies reducing the probability of hardware failure while $K_R < 1$ leads to faster components repair. Then, not surprisingly, the case $K_F = 2$ and $K_R = 0.8$ results in the highest availability in the example depicted in Figure 9(b).

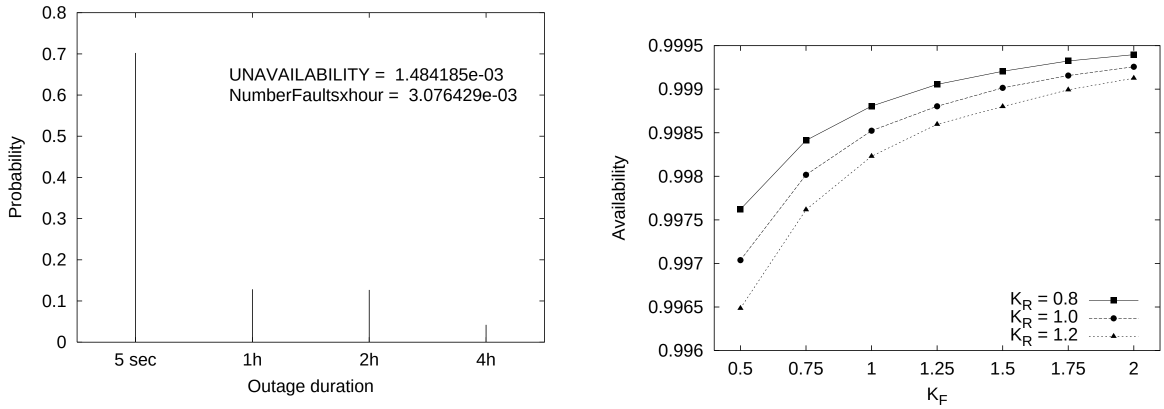

Figure 9: Outage distribution (a) and availability at varying of K_F and K_R(b)

4. Modelling the effects of outages on the GPRS behaviour

This section introduces the GPRS *service dependability* model, elaborated from the study conducted in [11], to analyze the effects of outages on the GPRS services. We focused on the GPRS behavior during the contention phase performed by users when making channel requests; this is actually a critical part, being a potential bottleneck for 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 the contribution given by this study in a more complete framework, are interesting directions for next studies.

Before developing the model, interesting indicators appropriate for the evaluation of outages have been identified. They are:

- The time necessary for the system to get close enough to its expected, steady-state behavior, following the end of an outage (referred to as recovery time in the following);
- A measure 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. Moreover we modeled the users in such a way that they are forced to wait for the satisfaction of the requests they ask for. This choice is not realistic but conservative. It forces a higher load on the system, inducing a less favorable 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.

Then, the model has been built so as to allow the evaluation of both indicators; in fact, both of them are measurable by observing the marking of the place representing the users which had their last request satisfied and are not yet trying to get another service. By comparing the number of tokens in the place *active* (see section 4.2) with the mean number of tokens contained in it when the system is in its steady state, one can perceive the degradation of the service. The recovery time is measured instead as the time necessary to get close enough to the steady-state marking.

4.1. Assumptions

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

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 make requests with the same rate; 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, thus no retransmissions 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 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;

7 In case traffic channels are completely busy, accepted requests are queued through an Access Grant Reservation.

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

4.2. The Service Dependability Model

As for the previous model, also the dependability model has been derived using the Stochastic Activity Networks formalism, as illustrated in Figure 10. The meaning of the main model elements is explained in the following.

- The number of tokens in the place *active* represents the number of users that have sent successfully their up-link data;
- The timed transition *to_req* represents the issue of a request by an active user (represented by a token in the place *active*);
- Tokens in place *new_request* denote those users that send a new access request to transfer data;
- The instantaneous transition *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;

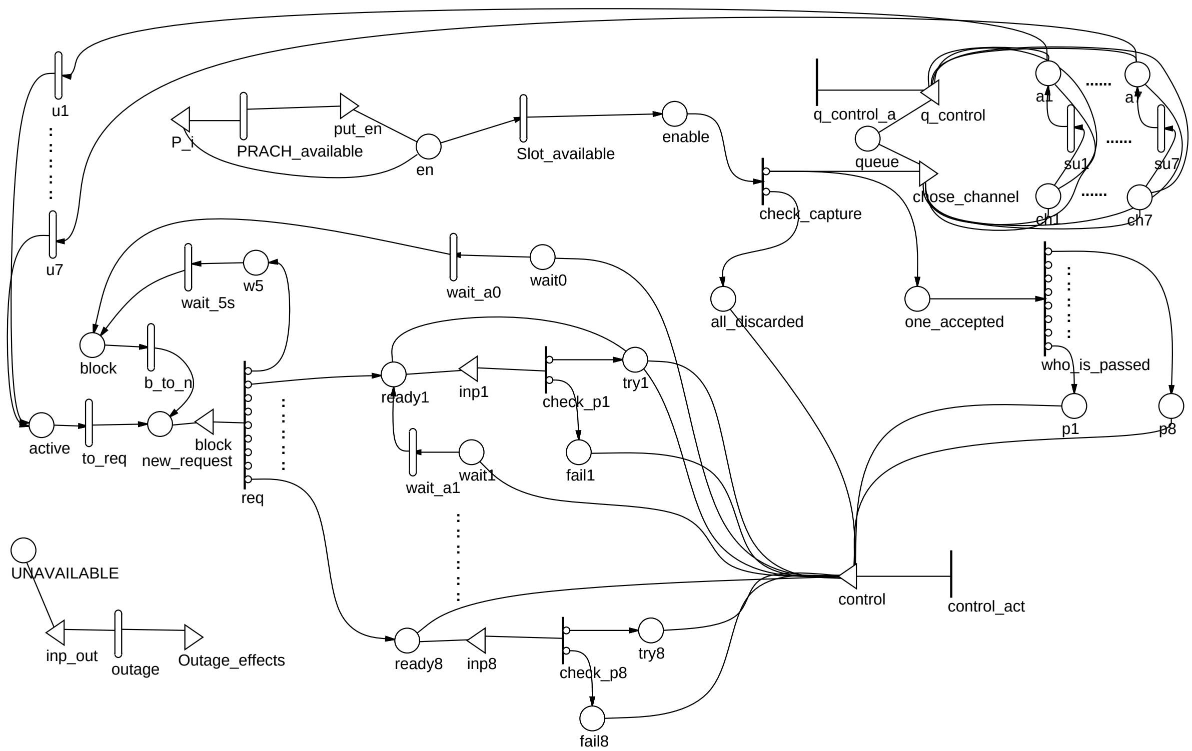

Figure 10: SAN model of the Random Access Procedure of the GPRS system

- 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. A user that is ready to send a request extracts a random value R with a uniform distribution in the set {0,1,...,15}; if R is equal to, or bigger than, P he can send an Access Burst (and move on the place *try*), otherwise he moves on the correspondent place *fail* . Therefore, tokens in places *try1,..., try8* denote the number of users that will send an Access Burst in correspondence of the next PRACH slot. Tokens in places *fail1,..., fail8* represent the number of users that have not passed the persistence control, but they can still do 1,...,8 attempts, respectively;
- The input gates *inp1,...,inp8* and *block* manage the correct running of the model, avoiding multiple instantaneous transitions to be enabled simultaneously and simulating faults, respectively;
- The number of tokens in places *wait0,..., wait7* represents the number of users having sent an Access Burst and waiting some time (modelled by the timed transitions *wait_a0,...,wait_a7*), during which the reception of the Packet Up-link Assignment message can occur; should this not happen, the contention is tried again, if there are residuals attempts;
- Place *w5* and activity *wait_5s* take into account those users that haven't been assigned any attempt. According to the standard specification, they have to wait 5 seconds before moving in the place *block*;
- The place *block* contains tokens representing those users which failed to get the right to access the network. 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 the Automatic Retransmission Time (ART);

- The activity *slot_available* simulates the 4 time slots of the PRACH radio block. The input gate *P_i* enables the transition *PRACH_available* only when the last PRACH period is already ended. The activity *PRACH_available* fires when a PRACH radio block is available; in this case, the output gate *put_en* puts 4 tokens (the 4 time slots in a PRACH) in the place *en*. A token in the place *enable* simulates the arrival of one burst period of a PRACH;

- 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*, unless the queue is full and there is no available traffic channel, in which case a token is put in *all_discarded*. The instantaneous transition *who_is_passed* fires when there is a successful receipt of one Access Burst (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; this probability is then proportional to the number of channel requests for each level). The input gate *control* and the activity *control_act* update the number of tokens in the places *ready1,...,ready8, try1,...,try8, fail1,...,fail8, wait0,...,wait7* and *p1,...,p8*;

- When there is a successful receipt of one Access Burst, the output gate *choose_channel* puts a token in one of the places *ch1,...,ch7* if there is a free channel (that is at least a free pair between *ch1-a1,....,ch7-a7*), otherwise it puts a token in the place *queue*;

- A token in the place *queue* represents a pending request waiting for up-link channel reservation;

- 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;

- The timed activities *su1,...,su7* simulate the set-up time of a radio link to send user data;

- The timed activities *u1,...,u7* simulate the data send time.

The subnet including Petri net elements *UNAVAILABLE, inp_out, outage,* and *Outage_effects* is used to represent the occurrence of outages, and the consequent repair of the system.

- A token in the place *UNAVAILABLE* represents the unavailability of the network. Note that this place is in common with the underlying network availability model;

- The input gate *inp_out* enables the transition *outage* without removing the token from the place UNAVAILABLE;

- The firing of the timed (but very fast) activity *outage* triggers the effects of an outage by means of the output gate *Outage_effects.* This gate simulates the effects of the fault through the inhibition of the immediate activity *req* and the gradual moving of the tokens of the whole net in the place *new_req*. For the sake of simplicity, the others Petri net elements which assure that the *outage* transition fire only once when there is a token in place *UNAVAILABLE* without removing it have been omitted.

4.3. Evaluating the effects of outages in GPRS systems

A detailed evaluation campaign on the effects of outages has been performed in [11]. There, the focus has been on a single outage of varying duration; its impact on the identified indicators (degradation during the outage and recovery time) has been deeply analyzed under varying users population, users characteristics and system workload.

In order to summarize the results shown in [11], we present here a couple of figures picturing interesting results on the effects of outages; this will help the reader in better understanding the following evaluations. Table 7 reports the values assigned to the main parameters in the numerical evaluation. The varying parameter in the experiments is the duration of outages.

Figure 11(a) shows the effects of outage on the number of served users, both during the outage and during the recovery time. Several curves have been plotted for different values of the outage duration. At growing values of the outage duration, more and more tokens exit the place *active*, meaning that more users are making new service requests. This is traced through the descendant line in the Figure. Then, the ascendant lines show the time necessary for obtaining the "normal" number of tokens in the place *active* (as determined by a steady state analysis under normal system conditions, i.e. in absence of outages, and indicated by the upper horizontal line).

Symbol	Description	Value
S-PDCH	number of Slave Packet Data Channel	3
U_A	number of users in the cell	150
PRACH	number of Packet Random Access CHannel per multiframe	2
Q_L	length of the queue (per data channel)	2
T_req	user inter-request time, following an exponential distribution	60 sec. (average)
T_out	outage duration, following a deterministic distribution	20-300 sec
R_size	request size, following a uniform distribution	1300 bytes (average)

Table 7: Relevant parameters and their values

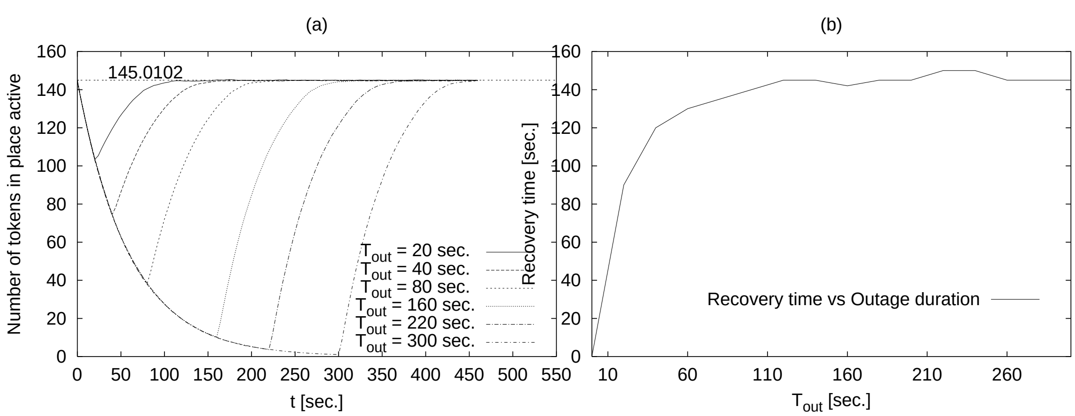

Figure 11: Behavior during outages and recovery (a) and time to steady state vs. outage duration (b)

The recovery time depends on the duration of outages as shown in Figure 11(b). For low values of the outage duration the recovery time varies significantly, becoming almost independent from it when outage duration gets high values (in the figure, greater than 160 seconds).

5. The GPRS Composed Model and its Evaluation

In this section the network model (Section 3) and service dependability model (Section 4) are combined and solved together. Evaluations on the join model are carried out to gain some insight on the effects of system availability and outages on the service provided by the GPRS system. At last, a sensitivity analysis shows how the frequency and duration of outages, given a fixed availability, affects the performance of GPRS.

5.1. The GPRS Composed Model

We now illustrate how the network availability model and the service dependability model have been composed in order to provide an accurate availability estimation of the GPRS system. The composed model allows taking into account both the availability of the components of the GPRS infrastructure, and the effects of outages on the user's perceived quality of service.

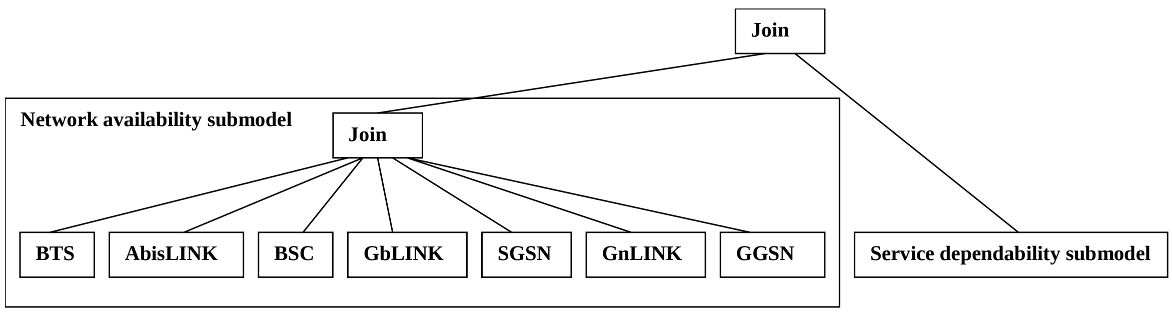

Figure 13: Network availability submodel composed with the service dependability submodel

The composition of the previously described models is actually quite straightforward, as shown in Figure 12, since it consists in joining the overall network availability (shown in Figure 8) with the service dependability subnet using the SAN Join operator. All the subnets are joined together having the place *UNAVAILABLE* as the only common place through which all the interactions take place.

5.2. Accurate Availability Evaluation

Figure 13(a) shows the reduction (percentage) of users served per hour, at varying inter request time, while Figure 13(b) shows the change of availability when the service dependability model is considered with respect to the network availability as estimated in Section 3.3. The settings of the others parameters of the service dependability model are the same shown in Table 7. The second x-axe shows the corresponding network traffic load as percentage of the network capacity. The dashed line reports the values obtained from the network model as estimated in the subsection 3.3. This is a constant value, because it does not depend on the transient effect of outages on the service model. The solid line represents the "real" availability as perceived by the class of users modeled.

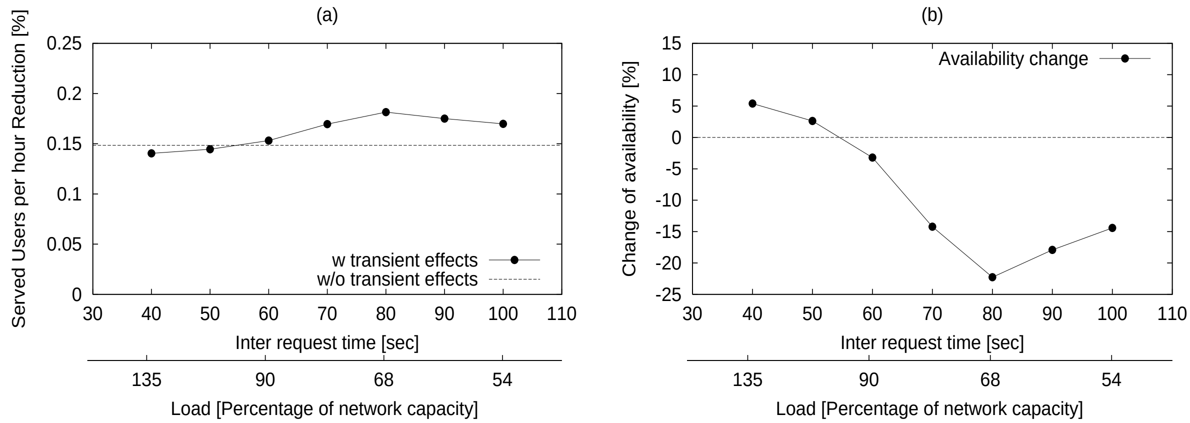

Figure 13: Percentage reduction of served users per hour with and without considering the transient effects (a) and availability reduction (b)

Looking at the Figures one may observe that for high load the user perceives a better QoS than just considering the network model. An explanation for such behavior is that the smooth degradation of the user perceived QoS when outages start out-weights the slow recovery after the system resumes operations. Actually the system is "always" overloaded so the overload at restarting does not cause any particular harm. Very different is the situation at lower loads. In these cases the effect of the slow restart determines the users to perceive a loss of QoS due to the congestion at system restart. In the present parameters setting, for inter request times of 80 seconds the degradation is the highest leading to a worsening of availability of about 20%.

We can conclude by saying that the bare estimation of QoS derived by the sole network model is an underestimation of what really users perceive, and that accounting of the transient effect brought about outages is necessary.

5.3. On the Impact of Outage Frequency and Duration

In the previous subsection we performed the evaluation of a specific GPRS architecture with its network availability, on which we obtained a more accurate estimation of the service availability provided to GPRS users. Additionally, we are interested in performing some sensitivity analysis at varying the frequency and duration of outages periods and for different values of network availability. Thus, in this subsection we consider a model that allows to perform such detailed analysis of service unavailability. We replace all the architecture submodels with a very simple submodel with two places (*Available* and *Unavailable*) and two timed transitions to model failure and repair of the architecture. The firing times of these transitions are derived once we fix the network availability and the Mean Time To Outage (MTTO) or the Mean Time To Repair (MTTR). The settings of the GPRS parameters are the same shown in Table 7.

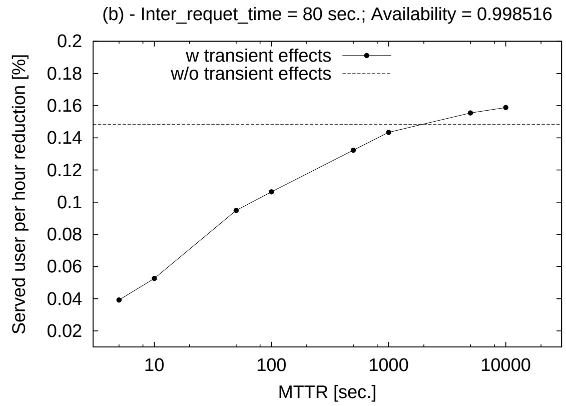

Figure 14: Number of served users per hour at varying of the MTTR

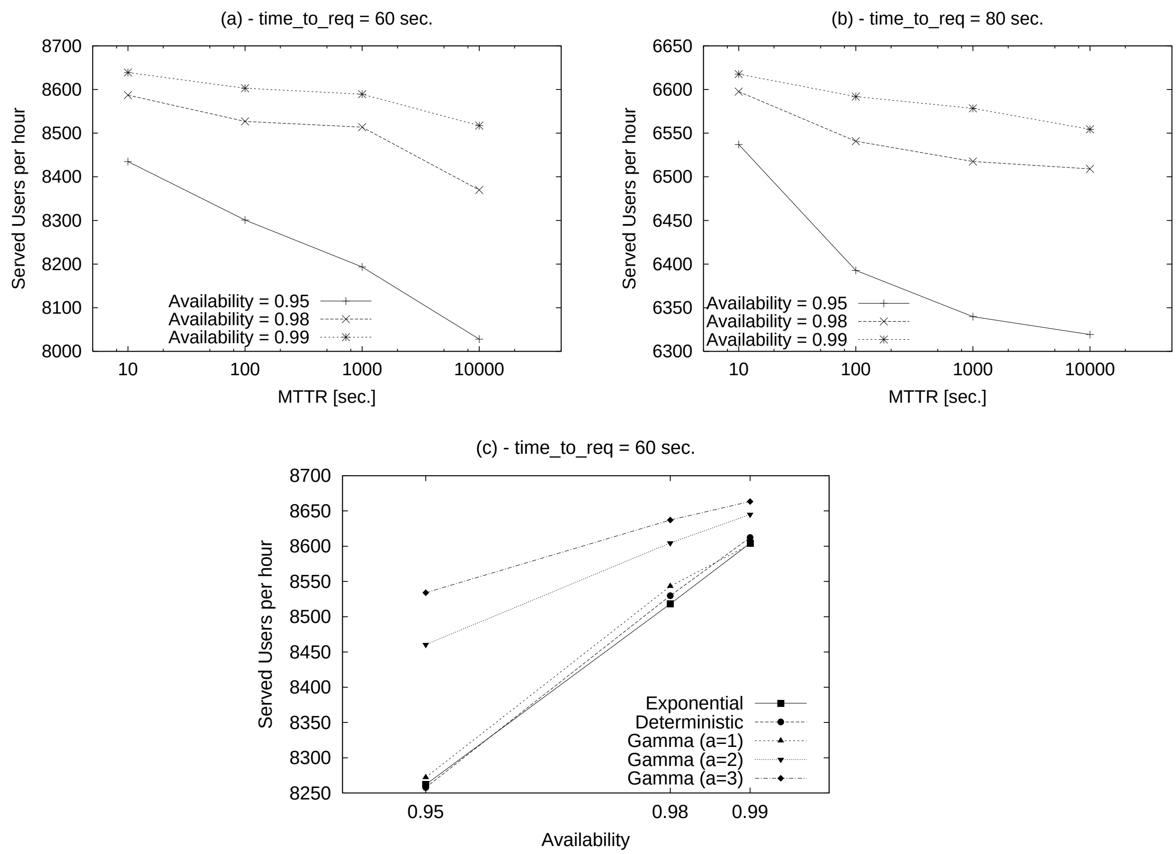

Figure 15: Served users per hour as a function of MTTR for inter_request_time = 60 sec (a), 80 sec (b), and at varying of availability for different MTTF distributions (c)

Figure 14 shows the reduction (percentage) of users served per hour at varying MTTR for inter request time of 80 sec., and fixing the network availability to .998516 (which is the availability of the system studied throughout the paper). We can observe that if the MTTR is short we have actually a gain in the QoS provided to users. The shorter (i.e. more frequent) the outages the better the user perceived QoS for the same “basic” system availability.

Figures 15(a) and 15(b) plot the served users per hour at varying the MTTR and the availability, for inter request times fixed to 60 and 80 seconds, respectively. They show how QoS improves as availability improves (as expected) but also as MTTR becomes shorter. Last, Figure 15(c) shows the served users per hour as a function of availability for different distributions of the MTTF, just to point out how not only the MTTF is important in determining the QoS, but also the distribution of the time to failure.

6. Conclusions

The work presented in this paper contributes to the analysis of GPRS by providing a modeling approach to better understand the effects of outage periods on the service provision. The goal was to improve the behavior analysis of such systems, to gain insights on the user perception of the QoS. A modeling approach has been followed, adopting the powerful modeling capabilities of Stochastic Activity Networks. Starting from a quite classical availability study of GPRS, based on the dependability of the various components of the GPRS infrastructure, the network dependability figures are combined into a very detailed service model that is used to analyze the overload effect that GPRS has to face after outages. The result of this modeling is an accurate availability analysis, which goes beyond the classical network perspective that is commonly taken when studying the availability of telecommunications systems, by including in the analysis a user and application perspective of the dependability of GPRS services. A typical GPRS configuration has been deeply analyzed and evaluated, in terms of a few identified QoS indicators, namely the number of served users per hour (and related measures) and the variation in system availability when including outages effects with respect to the bare network availability analysis. Interesting results have been observed, which can be fruitfully exploited in devising GPRS configurations adequate to maintain an acceptable QoS also in critical, overload conditions. Additionally, some sensitivity analysis has been performed at varying the frequency and duration of outages periods and for different values of network availability. Again, the objective of this study has been to enrich the knowledge of the impact of the outage phenomenon, in order to better cope with it and improve user satisfaction.

7. References

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

[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, 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, MA, 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.), 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, "Analysis of the Effects of Outages on the Quality of Service of GPRS Network Systems," *Proc. of the 2001 Int. Conference on Dependable Systems and Networks*, Göteborg, Sweden, July 2001, pp 235-244.