Analysis of User Perceived QoS in Ubiquitous UMTS Environments Subject to Faults

Andrea Bondavalli, Paolo Lollini, and Leonardo Montecchi

Università degli Studi di Firenze,
Dipartimento di Sistemi e Informatica,
viale Morgagni 65, I-50134, Firenze, Italy
http://rcl.dsi.unifi.it/
{bondavalli,lollini,lmontecchi}@unifi.it

Abstract. This paper provides a QoS analysis of a dynamic, ubiquitous UMTS network scenario in the automotive context identified in the ongoing EC HIDENETS project. The scenario comprises different types of mobile users, applications, traffic conditions, and outage events reducing the available network resources. Adopting a compositional modeling approach based on Stochastic Activity Networks formalism, we analyze the Quality of Service (QoS) both from the users' perspective and from the mobile operator's one. The classical QoS analysis is enhanced by taking into account the congestion both caused by the outage events and by the varying traffic conditions.

Keywords: QoS analysis, UMTS networks, partial outages, compositional modeling, stochastic activity networks, simulation.

1 Introduction

Ubiquitous infrastructures are typically composed by a high number of mobile devices that move within some physical areas, while being connected to networks by means of wireless links. The supported mobile-based applications should be capable to provide the expected services in a dependable way, and maintaining the required Quality of Service (QoS) levels.

In this paper we adopt a compositional modeling approach based on Stochastic Activity Networks to assess the QoS provided over complex, ubiquitous and dynamic infrastructures, taking as motivating example a use-case scenario defined in the ongoing EC HIDENETS project [1]. The analyzed system is characterized by a UMTS communication network composed by several partially overlapping cells, and by a set of users (i.e., cars and emergency vehicles, equipped with UMTS network devices) moving through the network and requiring different UMTS-based applications (e.g., voice call and entertainment). The user perceived QoS level should always be higher than a minimum level, and this aspect becomes particularly critical when emergency situations are considered, for example in the case of an ambulance that is using a streaming application to transmit the ECG traces of an injured person while moving to the hospital. Since

U. Brinkschulte, T. Givargis, and S. Russo (Eds.): SEUS 2008, LNCS 5287, pp. 186–197, 2008.

the user perceived QoS level depends on the availability of network resources, base stations' faults are also considered. More in detail we allow the presence of partial outages that may affect the availability of the UMTS resources.

The rest of this paper is organized as follows. Section 2 provides the description of the analyzed system and it outlines the corresponding QoS measures of interest. The main UMTS aspects influencing the QoS analysis are discussed in Section 3. The modeling approach is then sketched in Section 4, while Section 5 presents and discusses some of the obtained results. Finally, the conclusions are drawn in Section 6.

2 The System Context and the QoS Indicators

HIDENETS [1] is an ongoing EC project addressing the provision of available and resilient distributed applications and mobile services with critical requirements on highly dynamic and possibly unreliable open communication infrastructures. A set of representative use-case scenarios has been identified, each one composed by different applications (mostly selected from the field of car-to-car and car-to-infrastructure communications), different network domains (ad-hoc/wireless multi-hop domains, infrastructure network domains), different actors (end users, servers, routers, gateways), and characterized by different failure modes and challenges. In the following we give a brief description of the "car accident" scenario, which is analyzed in this paper and used as motivating example to describe the modeling process.

2.1 Definition of the "Car Accident" Use-Case Scenario

The "car accident" use-case scenario evolves around a scene with an accident on a road, involving cars. The use-case covers mainly what happens after the accident but also involves some issues directly before and during the accident. The analyzed network scenario is composed by a set of overlapping UMTS cells covering a high-way, and a set of mobile network devices (embedded or inside cars and emergency vehicles) moving in the high-way and requiring different UMTS class of services (e.g., conversational, interactive, and background).

Directly before the accident, several applications are used by the different mobile users, like entertainment and voice call. Right after the accident, many people may try to call the emergency services, call home, and send text and multimedia messages. Some time after the accident, an ambulance is approaching. Arriving at the place of the accident, and heading back to the hospital with the injured, there will be a need to transmit information on the positioning of the ambulance to communicate that it is approaching the hospital and at the same time maintain a multimedia connection with the medical expertise by use of voice, video and data transmission ("access to medical expertise" application).

The concrete UMTS scenario under analysis is depicted in Figure 1. Four base stations are considered: A, B, C and D. The base stations are subject to faults, which may reduce their available network resources. The users are moving in

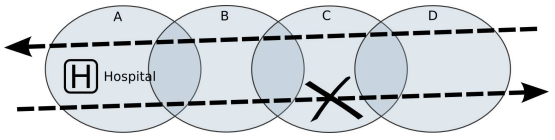

Fig. 1. The analyzed scenario

two different road lanes: part in the left to right lane (from A to D) and part in the right to left one (from D to A). We assume that the accident occurs in the C zone, in the left to right lane, forcing other users approaching that area to stop until the ambulance arrives, the crash site is cleaned and the normal traffic flow restored. The emergency vehicle heads back to the hospital towards the A zone where we suppose the hospital is located.

Concerning the available UMTS services, we suppose that a generic user can use three different services (Telephony, Web Browsing and File Transfer), while the ambulance uses the "access to medical expertise" application that consists of two simultaneously running services (Emergency Streaming to transmit the ECG traces, and Emergency Video-conference to fully interact with the hospital), having higher requirements in term of signal to interference ratio with respect to the non-emergency services. The services mainly differ for the activity factor, the uplink and downlink throughput and the required signal-to-interference ratio. Using well-known UMTS formulas (e.g., see [2]), these parameters are summarized into a single value that represents the workload increment they produce on the network (δ_{ul} and δ_{dl} parameters of Equation (1), Section 3).

2.2 QoS Indicators

The measures of interest concern the QoS levels both from the users' perspective and from a mobile operator's point of view. The QoS level perceived by users (both normal cars and emergency vehicles) depends on their capability to successfully use the network services when required and for the time required. The users involved in the traffic-jam should be capable to call home, while the ambulance should be capable to maintain the multimedia connections while moving towards the hospital. Typical user-oriented QoS indicators are: the following:

- The probability that a service request is successfully completed (P_{succ}),
- The probability that a service request is blocked (P_{block}) or dropped (P_{drop}).

The network workload is another system aspect that deserves special attention. Right after the accident, the behavior of the users involved in the consequent traffic-jam changes from normal to emergency, for example intensifying the service requests and trying to call the emergency services and call home, and this may cause congestion in the radio access network. In this context, typical mobile operator-oriented indicators are the following:

– The load factor, both in uplink (η_{ul}) and downlink (η_{dl}),
– The number of allocated traffic channels, which corresponds to the average number of served users.

3 Communication-Level Aspects Influencing the QoS Analysis

In this section we focus on the communication level aspects related to the "car accident" use-case, and in particular on three UMTS characteristics having important effects on the QoS: the *random-access procedure*, the *admission control* strategy and the *soft handover* mechanism. These characteristics mainly influence the so called "connection-level" QoS, which are the quality indicators related to the connectivity properties of the network, like the call blocking or dropping probability.

When a user needs a service from the UMTS network, its User Equipment (UE) sends a channel request to the network through the Physical Random Access CHannel (PRACH), a specific channel dedicated to the uplink transmission of channel request. The access method, based on a *random-access procedure* (RACH), may cause collisions among requests by different UEs, thus worsening the expected QoS (e.g., see [3] for more details on this aspect).

The *admission control* strategy is needed to decide whether a new service request can start based on the available network "capacity". Once the network receives the channel request, it performs the admission control procedure to decide if a traffic channel can be allocated to this new request. The goal is, in general, to ensure that the interference created after adding a new call does not exceed a pre-specified threshold, thus preventing the QoS to degrade below a certain level. There are several types of admission control algorithms studied in the literature, each one having different properties and aiming at optimizing different network parameters (e.g., [4]). Here we consider an admission control algorithm based on the workload of the UMTS cell: a new call is accepted if the workload level reached after adding the call does not exceed a pre-specified threshold, both in uplink and in downlink. Equivalently:

$$\eta_{ul} + \delta_{ul} \leq \eta_{ul\_threshold}, \qquad \eta_{dl} + \delta_{dl} \leq \eta_{dl\_threshold}, \tag{1}$$

where η_{ul}, δ_{ul} and $\eta_{ul\_threshold}$ (or η_{dl}, δ_{dl} and $\eta_{dl\_threshold}$) are, respectively, the cell workload before the admission of the new call, the workload increment due to the admission of the new call and the pre-specified threshold level in uplink (or downlink).

Another key aspect to be addressed is *soft handover*, a feature of the 3rd generation mobile networks, where a User Equipment can have two or more simultaneous connections with different cells (or cell sectors) and receive from them the same information signal. The signal received from different sources is then combined using rake receivers and under certain conditions this results in a amplified signal and better link quality. Beside providing better link quality, soft handover is also a key point in maintaining an ongoing service call, since it provides seamless switching between base stations.

4 Modelling Process

In such ubiquitous landscape, system complexity comes out to be a paramount challenge to cope with from a number of different points of view, including dependability and QoS evaluation. In order to master complexity, a modelling methodology is needed so that only the relevant system aspects need to be detailed, allowing numerical results to be effectively computable. The complexity of models depends on the dependability measures to be evaluated, the modelling level of detail, and the stochastic dependencies among the components. Several works have been presented in the literature trying to cope with the complexity problem (see [5] for a nice survey), and some of them try to tackle the complexity problem building models in a modular way through a *composition* of its submodels (e.g., [6,7]), which are then solved as a whole. Most of the works belonging to this class define the rules to be used to construct and interconnect the sub-models, and they provide an easy way to describe the behavior of systems having a high degree of dependency between subcomponents.

In this paper we adopt a compositional modeling approach based on Stochastic Activity Networks (SAN) [8], that are stochastic extensions to Petri Nets. The composition operators available for SAN are the *join* and *replicate* operators [9]: the first is used to compose different system models possibly sharing some places, while the second is used to combine multiple identical copies of a submodel, which are called replicates. Another key point of the modeling approach is the "model parametrization". Following the object oriented philosophy, we develop some "template" SAN models describing the general behavior of the main system components. The overall model results from the composition of some "instances" of such classes, where an instance is a specification of a template model with a proper parameters' setting. Using this approach we avoid duplicating the code and the structure of similar models, which is a very time-consuming and error-prone process; as a consequence, the overall model is easier to be modified and it can be more easily adapted to represent different scenarios.

In Figure 2 we have depicted the main basic SAN models (called "atomic" models in the SAN language) with their dependency relations (the arrows). An arrow from model X to Y means that model X can influence the stochastic behavior of model Y or, equivalently, that the Y state can depend on the X state.

For the sake of brevity the actual implementation of the atomic models is not described here. An exhaustive and detailed description of such models can be found as a technical report in [10]. In the following we outline the main system aspects captured by the different models.

- **Phases** atomic model. It represents the sequence of periods (phases) composing the system lifetime, each one characterized by diverse applications running, diverse types of users's behavior (normal behavior, before the car accident, or emergency behavior, right after the car accident) and different dependability properties to be ensured.
- **User** atomic model. It describes the user's behavior mainly in terms of services requested, duration of the services and idle periods.

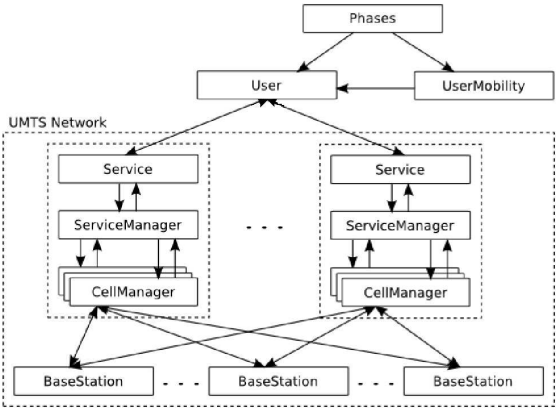

Fig. 2. Atomic models and their interactions

- **UserMobility** atomic model. It represents the user movement across the UMTS network scenario.
- The "UMTS Network" model consists of several instances of the **BaseStation** atomic model and a number of models representing the available services. A network service is represented using three kinds of atomic models: **Service**, **ServiceManager** and **CellManager**. The Service atomic model represents the upper network layers and it is directly connected with the User model. When the user requests a network service, the User model interacts with the respective Service model which serves as interface between the user and the network. The Service model then asks for the needed resources to the ServiceManager atomic model, which handles the soft handover mechanisms allowing user to be served by multiple base stations. This is achieved using several instances of the CellManager atomic model, which serve as interfaces between the ServiceManager atomic model and each BaseStation model. Finally the BaseStation model represent a UMTS base station, with failure and repair activities, and holds the current base station state, like its current workload and the number of allocated channels. In case of outage events this model also implements the congestion control algorithm, which drops (interrupt) a certain number of connections if the current workload exceeds the remaining available system resources.

Once such basic template models have been developed, several different scenarios can be easily obtained resembling different network topologies, users' behaviors, users' mobility patterns and available applications. Therefore, the modularity of the modeling framework improves both the readability and the maintenance of the models, as well as their reusability. In the following Section 4.1 we detail the overall model for the "car accident" use-case scenario defined in Section 2.1.

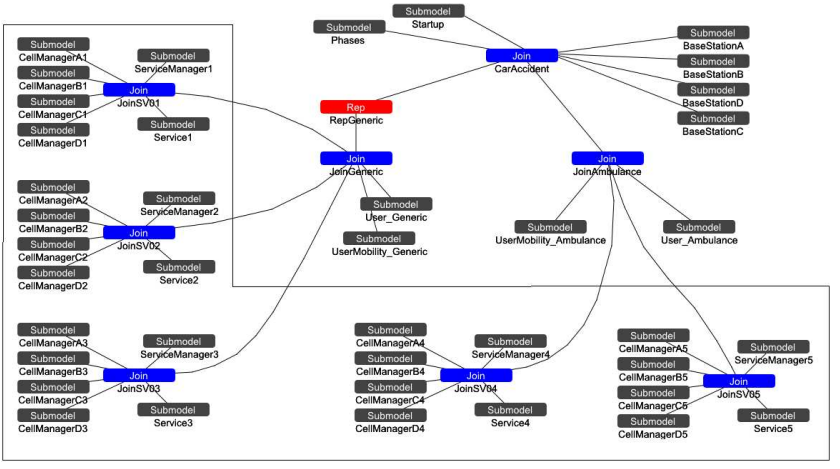

Fig. 3. The overall composed model corresponding to the analyzed use-case scenario

4.1 The Overall UMTS Network Model

As described in Section 2.1, the analyzed scenario consists of four partially overlapping base stations and five services: services 1 (Telephony), 2 (Web Browsing) and 3 (File Transfer) are assumed to be services for the "normal" users, while 4 (Emergency Streaming) and 5 (Emergency Video-conference) are assumed to be services used by the emergency vehicle and will have higher requirements in term of signal to interference ratio.

Figure 3 depicts the corresponding composed model. The composition involves three join levels. Starting from the lower level (the boxed part of the figure), joins relative to different services are shown, each one formed by four CellManager models (one for each base station), a Service and a ServiceManager model, as sketched in Figure 2. In the second level the services are joined with the respective user models, so services 1-3 are composed with User_Generic and UserMobility_Generic (on left part of the figure), while services 4-5 are composed with User_Ambulance and UserMobility_Ambulance (on right part). The generic user is then replicated as needed and added to the top-level join, which also includes the ambulance join, the four BaseStation models, the Phases model and the Startup model (used to initialize the multiple instances of the other atomic models with the proper values).

The advantage of using model parametrization is evident considering the effort required for the model construction process. To build the composed model shown in Figure 3, 40 atomic models are needed (exactly 20xCellManager, 5xServiceManager, 5xService, 4xBaseStation, 1xStartup, 1xPhases, 1xUser_Generic, 1xUserMobility_Generic, 1xUser_Ambulance and 1xUserMobility_Ambulance). Using model parametrization we need to create the basic template atomic models only, one for each type. For this scenario only 10 atomic models have been built, and those depicted in Figure 3 are just instances of these basic 10 models. Once

the basic template models have been defined, we can easily build and analyze different scenarios with a very small effort. For example, deleting the JoinAmbulance composed model in Figure 3 we can limit the analysis to normal (not emergency) services, while adding another base station (thus obtaining a different network topology) would simply consist in adding another CellManager atomic model to each JoinSV composed model, and another BaseStation atomic model (BaseStationE) to the CarAccident composed model.

5 Numerical Evaluations

In this section we sketch some of the results that we obtain through the solution of the models previously described. A transient analysis has been performed, using the simulator provided by the Möbius tool [11]. Each point of the graphs has been computed as a mean of at least 1000 simulation batches, converging within 95% probability in a 0.1 relative interval, and a AMD Athlon XP 2500+ PC (2Gb RAM) has been used for the computations.

Table 1. Workload increment per accepted service, in uplink (δ_{ul}) and downlink (δ_{dl}), with or without soft-handover

		Service1	Service2	Service3	Service4	Service5
Workload increment (no SHO)	Downlink (δ_{dl})	0.01357	0.08774	0.2125	0	0.09917
	Uplink (δ_{ul})	0.01632	0.06675	0.0620	0.11923	0.10814
Workload increment (with SHO)	Downlink (δ_{dl})	0.00995	0.03781	0.1275	0	0.07083
	Uplink (δ_{ul})	0.00984	0.03411	0.03781	0.08543	0.0620

The setting of the model's parameters has been mainly derived from [2] and adapted to the analyzed scenario. With reference to Equation (1), the values assigned to δ_{ul} and δ_{dl} are shown in Table 1, while the maximum load factor in uplink ($\eta_{ul\_threshold}$) and downlink ($\eta_{dl\_threshold}$) has been set, respectively, to 0.65 and 0.8. Each base-station has a coverage area of 2 Km, and 25% of the cell radius is overlapping with the adjacent cell. Whenever not differently specified, we consider a total of 50 cars moving in the scenario (and 1 ambulance), with an average speed of 90 Km/h (120 Km/h for the ambulance) when not involved in the traffic jam caused by the car accident. Moreover, we suppose that the car accident happens in the C cell at time t=10500 sec., and that the ambulance stays 600 sec. at the crash site before heading back to the hospital. The complete set of model's parameters and their setting can be found in [10], and it is not reported here for the sake of brevity.

5.1 Results

Figure 4 shows the load factor of the base station C, considering no outage events (i.e., 100% of the network resources are always available). The vertical line represents the instant of time when the accident occurs, while the horizontal ones represent the maximum allowed load factor in downlink ($\eta_{dl\_threshold}$) and

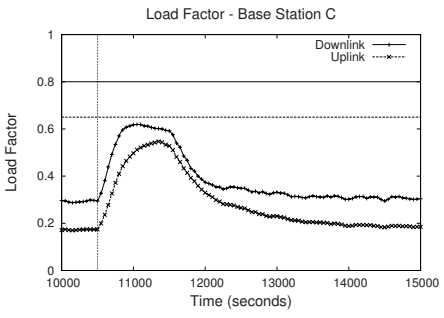

Fig. 4. Load Factor of the base station C

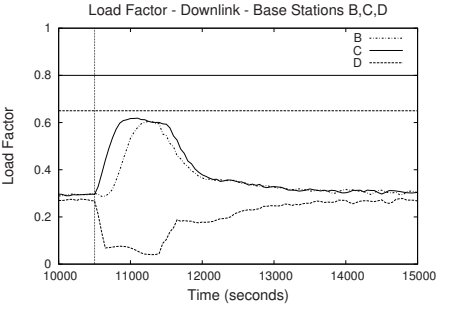

Fig. 5. Load Factor of different base stations (downlink)

in uplink ($\eta_{ul\_threshold}$). As shown in Figure 4, right after the accident the C cell becomes rapidly congested due to the users that are stopped in that area (due to the consequent traffic-jam). The congestion phenomenon is also exacerbated considering that the users' behavior changes during emergency conditions, in particular reducing the idle time between two consecutive service requests. After a while the load factor starts to decrease, because the normal traffic flow is restored after the ambulance heads back to the hospital, and because the users behavior becomes again normal. The load factors of base stations near the accident zone are summarized in Figures 5 (downlink) and 6 (uplink). After a certain delay a congestion is also produced on the base station B and this is due to the traffic-jam reaching its coverage area. On the contrary, the load factor of D rapidly decreases right after the accident, since the cars are blocked in the preceding cells. When the crash site is cleared and user (cars) are capable to move again all the load factors slowly return to the level they had before the car-accident.

Figure 7 shows the impact of an outage affecting the base station C at time t=11000 sec. on the "access to medical expertise" application used by the ambulance (plot '4+5 Combined'). The probability of service interruption rapidly increases considering higher percentage of resources unavailability, reaching its maximum for values greater than 60%. Analyzing the single services forming the application (plots 'Emergency Streaming' and 'Emergency Video-conference'), initially the probability is lower for 'Emergency Streaming', but when limited resources are available 'Emergency Video-conference' has a lower probability of interruption. This happens because we have assumed that 'Emergency Streaming' requires more uplink resources than 'Emergency Video-conference' (see Table 1), and after the outage the available uplink resources are lower than the downlink ones. For a better understanding, in Figure 8 we depict the uplink and downlink load factor of the base station C considering an outage equal to 70% (i.e., 70% of the cell resources becomes unavailable), at varying of time. The load factor (both in uplink and downlink) increases after the car accident (at time t=10500 sec.), and then rapidly decreases after the outage event

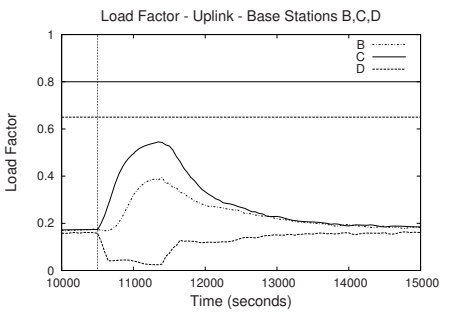

Fig. 6. Load Factor of different base stations (uplink)

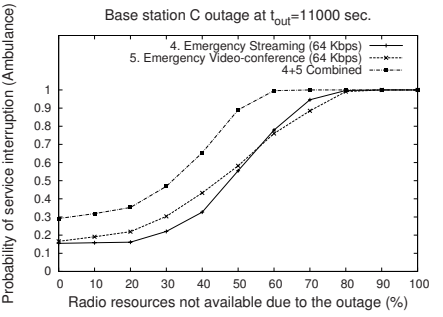

Fig. 7. Probability of interruption of emergency services, with C cell outage

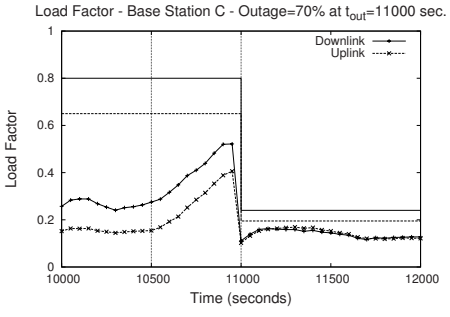

Fig. 8. Load Factor of the base station C, with outage=70%

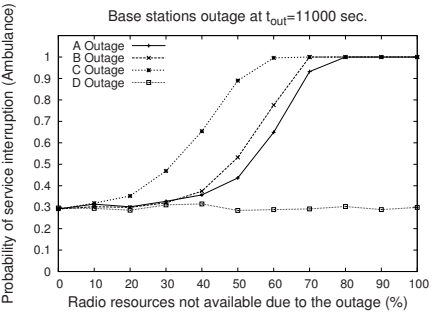

Fig. 9. Probability of interruption of ambulance services (4+5 combined) when a single base station fails

at time t=11000 sec. (due to the dropped services). After the outage the uplink load factor is near to its maximum allowed value and then the services requiring higher uplink resources will be probably not satisfied (due to the selected admission control algorithm, see (1)).

Figure 9 shows the impact of the outage on the probability that the "access to medical expertise" application is interrupted, at varying of the outage severity (percentage of unavailable cell resources) and at varying of the base station affected by the outage. As expected, base station C is the most critical one, since it is the cell where the car accident occurs and the traffic is blocked, thus determining a high network congestion. Cell D does not influence the ambulance connection at all, since the ambulance doesn't even enter the D zone.

Figure 10 shows the probability that the multimedia connections between the ambulance and the hospital are interrupted while the ambulance is going back to the hospital, varying the vehicle's average speed (no outages considered). Individual probabilities for the emergency services 'Emergency Streaming' and

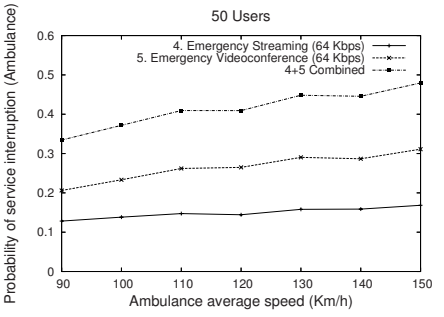

Fig. 10. "Access to medical expertise" interruption probability

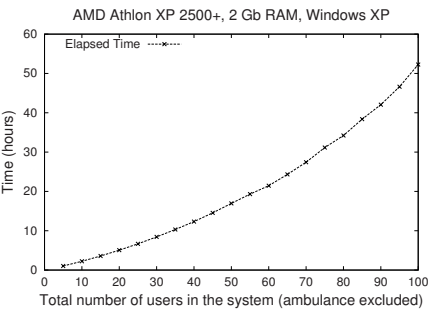

Fig. 11. Average time to produce Figure 10, at varying of the total number of users in the system

'Emergency Video-conference' are shown, as well as the overall probability that corresponds to the "access to medical expertise" application. The probability of interruption is lower for service 'Emergency Streaming' because we assumed that it only uses uplink bandwidth and then it requires less network resources than service 'Emergency Video-conference'. Results also show that the probability increases when vehicle speed increases and this effect is in part caused by the RACH procedure delay.

In Section 4.1 we have shown the effectiveness of the modeling approach in facilitating the construction of the overall model, which can be obtained as composition of 40 models derived from a set of 10 basic template models only. Anyway, the modeling approach is really effective only if the computational cost required to solve the overall model is still manageable. In Figure 11 we present the average time (in hours) needed to produce Figure 10, at varying of the total number of users in the system. The values in Figure 10 have been obtained performing 7 simulations, one for each considered 'ambulance average speed' value. As we can see, the computational time increases almost linearly for a low number of users, and the rate of grow slightly increases considering more than 60 users. Nevertheless, in the worst case (i.e., for number of users equal to 100) the whole set of simulations completed in less than 56 hours (therefore, less then 8 hours for each simulation).

6 Conclusions

In this paper we have proposed a QoS analysis of a dynamic, ubiquitous UMTS network scenario identified in the ongoing EC HIDENETS project, including different types of mobile users, applications, traffic conditions, and outage events affecting the availability of the network resources. The final goal was to quantitatively evaluate some QoS measures regarding both the users (the probability that an ongoing service request is interrupted) and the mobile operators (the load factor of the UMTS cells). To do this, we have adopted a modular,

hierarchical modeling approach based on composition, replication and parametrization, which facilitates the model construction process as well as the model reusability. The produced numerical results provides a useful insight in the relationships between the selected QoS measures, the users' behavior and the users' mobility. In addition, they show the effectiveness of the modeling approach considering the computational time required to solve the overall model by simulation.

Acknowledgment

This work has been partially supported by the EC IST Project HIDENETS [1].

References

1. European Project HIDENETS: contract n. 26979, http://www.hidenets.aau.dk
2. Laiho, J., Wacker, A., Novosad, T.: Radio Network Planning and Optimisation for UMTS, 2nd edn. Wiley, Chichester (2006)
3. Lollini, P., Bondavalli, A., Di Giandomenico, F.: QoS Analysis of a UMTS cell with different Service Classes. In: CSN-2005 The Fourth IASTED International Conference on Communication Systems and Networks, September 12-14 (2005)
4. Andersin, M., Rosberg, Z., Zander, J.: Soft and safe admission control in cellular networks. IEEE/ACM Transaction on Networking 5(2) (1997)
5. Nicol, D.M., Sanders, W.H., Trivedi, K.S.: Model-based evaluation: From dependability to security. IEEE Transactions on Dependable and Secure Computing 1(1), 48–65 (2004)
6. Rojas, I.: Compositional construction of SWN models. The Computer Journal 38(7), 612–621 (1995)
7. Bernardi, S., Donatelli, S.: Stochastic petri nets and inheritance for dependability modelling. In: Proceedings of the 10th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2004), March 2004, pp. 363–372 (2004)
8. Sanders, W.H., Meyer, J.F.: Stochastic activity networks: formal definitions and concepts, pp. 315–343 (2002)
9. Sanders, W.H., Meyer, J.F.: Reduced base model construction methods for stochastic activity networks. IEEE Journal on Selected Areas in Communications 9(1), 25–36 (1991)
10. Lollini, P., Montecchi, L., Bondavalli, A.: On the evaluation of hidenets use-cases having phased behavior. Technical Report rcl071201, University of Florence, Dip. Sistemi Informatica, RCL group (December 2007), http://dcl.isti.cnr.it/Documentation/Papers/Techreports.html
11. Daly, D., Deavours, D.D., Doyle, J.M., Webster, P.G., Sanders, W.H.: Möbius: An extensible tool for performance and dependability modeling. In: Schaumnurg, I.L., Haverkort, B.R., Bohnenkamp, H.C., Smith, C.U. (eds.) TOOLS 2000. LNCS, vol. 1786, pp. 332–336. Springer, Heidelberg (2000)