Reliability and Survivability Analysis for UMTS Networks: An Analytical Approach

S. Dharmaraja, Senior Member, IEEE, Vaneeta Jindal, Student Member, IEEE, and Upkar Varshney, Member, IEEE

Abstract—Reliability and survivability are the two important attributes of cellular networks. In the existing literature, these measures were studied through the simulation. In this paper, we construct an analytical model to determine reliability and survivability attributes of third generation and beyond Universal Mobile Telecommunication Systems (UMTS) networks. Hierarchical architecture of UMTS networks is modeled using stochastic models such as Markov chains, semi-Markov process, reliability block diagrams and Markov reward models to obtain these attributes. The model can be tailored to evaluate the reliability and survivability attributes of other beyond third generation cellular networks such as All-IP UMTS networks and CDMA2000. Numerical results illustrate the applicability of the proposed analytical model. It is observed that incorporating fault tolerance increases the network reliability and survivability. The results are useful for reliable topological design of UMTS networks. In addition, it can help the guarantee of network connectivity after any failure, without over dimensioning the networks. Moreover, it might have some impact from the point of view of the design and evaluation of UMTS infrastructures.

Index Terms—Reliability, Survivability, Markov chain, Reliability block diagram, Hierarchical modeling, UMTS networks

I. INTRODUCTION

UNIVERSAL Mobile Telecommunication System (UMTS) network is a third generation cellular network that provides broadband services to the world of wireless and mobile communications. It preserves the global roaming capability of second generation Global System Mobile (GSM)/General Packet Radio Service (GPRS) networks and provides new enhanced capabilities [1]. The UMTS network is designed to deliver pictures, graphics and video communication, in addition to voice and data services, to the mobile users. As a result, the UMTS network carry a steadily increasing amount of network traffic and mobile users demand more reliable services [2]. The network operators can guarantee this reliability only if the network is equipped with network survivability mechanisms that are able to minimize the impact of network failures. These failures include hardware failures, software errors, loss of wireless/wired links and power outage in the network elements. Reliability and survivability are of paramount importance since they will help the network designers to get a better insight into the various reliability and survivability considerations for the UMTS network.

In recent past, several studies have focussed on the survivability of the networked systems [3], [4], [5]. A multi-layer survivability framework that includes survivability metrics, for instance, call blocking probabilities, and the fault tolerance strategies such as SONET rings, overlay networks, restoration techniques and redundancy, was presented in [6], [7], [8]. A discrete event simulation model based on Wireless Infrastructure Building (WIB) block approach, where each WIB block consists of a Mobile Switching Center (MSC), several Base Station Controllers (BSC) managed by a MSC and multiple Base Stations (BS) controlled by a BSC, was proposed in [9]. The analytical approaches available in the literature for reliability analysis include model based reliability analysis [10] and continuous time Bayesian network reliability approach [11]. However, these approaches were not applied for the reliability analysis of wireless networks. Simulation modeling technique has been used for reliability analysis of wireless access network topologies such as ring and mesh topology [12].

However, simulation models can become untractable, computationally intensive, takes longer time, very specific to parameters chosen, and difficult to be used by researchers working on different platforms and facilities. This facilitates the need to develop an analytical model which is mathematically tractable, less time consuming and can be re-used by other researchers and this is the main contribution of this paper. In this paper, we develop an analytical model for the reliability and survivability analysis of UMTS network. The proposed model will be useful to the designers as well as to the researchers in the area of network domain. The results are also useful for reliable topological design of UMTS networks. The numerical results given in this paper show that incorporating fault tolerance enhances the network reliability substantially. It is observed that increasing the number of elements at any level beyond a maximum number will not enhance the network reliability. In addition, it can help the guarantee of network connectivity after any failure, without over dimensioning the network. The proposed analytical model can be reproduced with minor modifications for computing the reliability of other cellular networks. This updated model is solvable by using the basic knowledge of probability. Therefore, the proposed analytical model is in the interests of network designers and
researchers in the area of wireless networks.

The rest of this paper is organized as follows: Section II describes the general reference architecture of UMTS network and several fault tolerance strategies to mitigate the impact of failures. In Section III and IV, we develop analytical models for evaluating the network reliability and survivability respectively. Section V discusses the compatibility of the proposed model with other cellular networks such as All-IP UMTS networks and CDMA2000. Finally, Section VI presents the practical insights of the proposed model. Pointers to conclusions and future work are given in Section VII.

II. UMTS NETWORK ARCHITECTURE

In order to compute the reliability and survivability for UMTS network, we first present the general reference hierarchical architecture for this network. The level abstraction facilitates the presentation of UMTS network elements such as Node B, Radio Network Controller (RNC) and Core Network (CN) [13] and [14], and encourages for step-by-step evaluation of reliability and survivability attributes.

The general architecture of UMTS network is shown in Fig. 1. The mobile users in a network coverage area are associated to a cell, where each cell is served by a Node B. The main function of Node B is to provide wireless channels to the mobile users. It monitors the signal quality of these channels and transmits the information to the RNC at the higher level to which it is connected. Each RNC at level 2 controls a group of Node Bs. The functions of RNC include operation and maintenance of network elements, radio resource control, handover of the ongoing calls, broadcast signalling etc. [14]. Each RNC is controlled by a CN at level 3 of the UMTS network architecture. The CN provides switching, routing, transport and database functions for user traffic. It contains Circuit Switch (CS) elements such as MSC and interworking managers. It also contain Packet Switch (PS) elements such as Service Gateway Support Node (SGSN) and GPRS Gateway Support Node (GGSN). Further, database (for example, Home Location Register and Visitor Location Register (VLR)) stores the user location information [15]. Communication links between Node B, RNC and CN are wired/microwave links.

Several failures such as hardware failure, software failure, power outage, faulty links occur at the network elements. These failures are caused by incorrect specifications, flaws in network designing, poor testing and operator errors [5]. Hence, fault tolerant network design becomes an important issue. Redundancy in hardware components, back-up power supply, automatic restoration and fault tolerant interconnection architectures, for example, ring, SONET and mesh, are among the many fault tolerance techniques that are incorporated in the UMTS network [8], [16] to mitigate the impact of network failures. In the next section, we present an analytical model for computing the reliability of UMTS network.

III. RELIABILITY MODEL

In real-world problems, most of the failure and repair times follow general distributions like Weibull, Pareto and lognormal. However, in general, analytical models with general (non-exponential) distributions are not mathematically tractable with closed form or numerical solution. Therefore, phase-type distribution, which is convolution of many exponential phases [17] is used for approximating many general distributions and then construct the mathematically solvable analytical models. Since exponential distribution is a particular case of phase-type distribution, we consider exponential distribution to model all the random variables and construct the analytically tractable model for reliability and survivability analysis.

A. Reliability at level 1

Node B is at the lowest level, that is, level 1 of UMTS networks. Reliability of Node B depends on its four principle components: hardware, software, power supplies; and wired/microwave links to RNC. Hardware of Node B comprises of channel unit processor (CUP), antennas, transmitters and receivers. CUP is the most critical component since all the channel processing functions are accomplished here. Its failure will result in complete outage of Node B. To mitigate the impact of CUP failure, network vendors provide a redundant CUP [16]. On failure of the active CUP, Node B switches over to the redundant CUP. Damages to other hardware components such as antennas, transmitters and receivers of Node B are caused by malicious attacks and natural disasters which are rare events. Therefore, in this paper, we suppose that the probability of failures for antennas, transmitters and receivers is negligible.

Second, software module of Node B, implemented in CUP, includes the channel processing functions such as access channel detection, encoding/decoding, multi-path detection, spreading and scrambling [18]. Software faults in any of these functions result in failure (non-availability) of the wireless channels. If these software faults are not recovered, then wireless channels are not made available to mobile subscribers(MS). In this case, Node B is in “down” state. N-version programming, error-control coding and recovery block techniques are some of the fault tolerance strategies that are incorporated for recovering the software failures [5].

Third, during irregular power supplies, back-up batteries are brought into immediate operation to sustain the communication [19]. If backup batteries also fail or are exhausted, Node B is unable to continue with its services. Finally, the most important component of Node B is its link to RNC which
fails due to interference, noise and fading. On its failure, the ongoing communication at Node B is terminated at once. Therefore, highly reliable wired/wireless microwave links are used for connecting Node B to RNC.

The reliability model of a Node B is presented as a semi-Markov process with state space \( S = \{ U, U_H, U_S, U_P, U_{HS}, U_{SP}, U_{HP}, U_{PH}, U_{PS}, D \} \). The state transition diagram is shown in Fig. 2. Initially, all components of Node B are working and the system is in state \( U \). The active CUP fails with rate \( \lambda_h \). On its failure, Node B restores services by switching over to the redundant CUP. The intermediate steps to recovery such as switch over, synchronization etc. were considered in Fig. 7. As a result, the time to HW failure and its recovery is generally distributed with mean \( 1/\lambda_{hr} \). The faults in the software module are recovered through the fault tolerance strategies with coverage probability \( c \). With probability \((1 - c)\), these faults are not recovered and the system moves to the down state \( D \). If there is a power outage, Node B switches over to the backup power with rate \( \lambda_p \). Node B enters the down state \( D \) when backup batteries are exhausted or failed. The states belonging to \( S - \{ D \} \) are up states since Node B is working in these states.

By following the methodology in [20], the semi-Markov process is solved to obtain the time-dependent state probabilities. In our study, we assumed that \( M \) node Bs are managed by a RNC. Then, reliability of a \( m^{th} \) (\( m \in \{ 1, 2, \ldots, M \} \)) Node B, labeled as \( R_{NB-m}(t) \), is obtained by summing the time-dependent probabilities of the up states and is given as:

\[
R_{NB-m}(t) = \sum_{k \in S- \{ D \}} \pi_k(t)
\]  

(1)

Based on the statistics of population type and information such as income, distribution of wealth for this population, provided by the government sources, the required number of Node Bs in a network area is estimated [21], [22]. Note that the number of Node Bs in a network as well as the number of Node Bs controlled by a RNC depends on the capacity of a RNC. Each Node B has different parameters for failure and recovery rates. However, for simplicity, we suppose that the Node Bs controlled by one RNC are identical, i.e., \( R_{NB-m}(t) = R_{NB}(t) \ \forall \ m \). At least one of the \( M \) Node Bs should be available, for if, all the Node Bs are in down state, the mobile users in the coverage of the RNC could not access the networks. Therefore, we compute the reliability of a group of Node Bs, denoted by \( R_{GNB}(t) \), and is given as

\[
R_{GNB}(t) = \sum_{m=1}^{M} \left( \frac{M}{m} \right) [R_{NB}(t)]^m[1 - R_{NB}(t)]^{M-m}
\]  

(2)

where \( m \) out of \( M \) Node Bs are available and \( R_{NB}(t) \) is substituted from equation (1).

B. Reliability at Level 2

RNC constitutes the level 2 of UMTS networks and consists of components namely, ethernet chassis, telecom adaptor, single board computer, software unit, links to CN and power supplies. After identifying the subcomponents of a RNC, we construct a RBD to depict the structural dependency of reliability of RNC on its components. Fig. 3 shows the RBD for evaluating reliability of RNC.

Reliability for a component \( R(t) \), is given by \( R(t) = P(T > t) \), where the random variable \( T \) denotes the time to failure of a component [23]. As mentioned earlier, we consider that \( T \) follows be exponentially distributed. Let us assume that \( N \) RNCs are connected to a CN. After determining the reliability of the individual components, reliability of a \( m^{th} \) (\( m \in \{ 1, 2, \ldots, N \} \)) RNC is obtained by following the methodology in [23], and is given as:

\[
R_{RNC-m}(t) = R_{GNB}(t) \times \left( 2e^{-\lambda_{eth}t} - e^{-2\lambda_{eth}t} \right) \times e^{-\left( \lambda_{comp} + \lambda_{adp} \right)t} \times \left( 2e^{-\lambda_{pow}t} - e^{-2\lambda_{pow}t} \right) \times e^{-\left( \lambda_{link} + \lambda_{f1} \right)t}
\]  

(3)

where \( R_{GNB}(t) \) given by equation (2) is the reliability of a group of Node Bs controlled by this RNC. Symbols and their failure rates are as follows: Ethernet (\( \lambda_{eth} \)), single board computer (\( \lambda_{comp} \)), adaptor (\( \lambda_{adp} \)), power (\( \lambda_{pow} \)), wired links to CN (\( \lambda_{link} \)) and software (\( \lambda_{f1} \)).
Let at least one RNC should be in operation in order to provide connectivity to the mobile users [21]. It is to be noted that these N RNCs are not identical. Then, the reliability of a group of RNC, denoted by $R_{GRCN}(t)$, is obtained as follows: Define $S = \{1, 2, \ldots, N\}$ and $\mathcal{A}(S) = P(S) - \{\Phi\}$, where $P(S)$ is the power set of $S$. Then, $R_{GRCN}(t)$ is given as:

$$\sum_{S_j \in \mathcal{A}(S)} \prod_{m \in S_j} R_{RNC_m}(t) \prod_{k \in S_j} [1 - R_{RNC_k}(t)] \quad (4)$$

where $R_{RNC_m}(t)$ is given by equation (3). $R_{GRCN}(t)$ will be substituted while deriving the reliability of CN. We next derive the expression for reliability of CN.

C. Reliability at level 3

At level 3 of UMTS network, assume that there are, say, $K$ CNs. Each CN comprises of CS and PS domain entities that support voice and data traffic respectively. Associated with CN is the database (HLR/VLR) that stores the user location information, authentication and routing information. Fig. 4 shows the RBD for evaluating the reliability of a $m^{th}$ ($m \in \{1, 2, \ldots, K\}$) CN. Since CS and PS domain comprises of many components, reliability of CS and PS domain, denoted by $R_{CS}(t)$ and $R_{PS}(t)$ respectively, are obtained in sequel. We assume exponential distribution for the time to database failure with rate $\lambda_{DB}$ and is denoted as $Exp(\lambda_{DB})$. Reliability of a $m^{th}$ ($m \in \{1, 2, \ldots, K\}$) CN is given as:

$$R_{CN_m}(t) = R_{GRCN}(t) \times [1 - \{1 - R_{PS}(t)\} \times [1 - R_{CS}(t)] \times [e^{-\lambda_{DB}t}]] \quad (5)$$

where $R_{GRCN}(t)$ is substituted from equation (4).

1) Reliability of CS domain: The main component of CS domain is MSC which constitutes of two classes of servers: Interworking managers (IMs) that act as gateways to external network elements and core servers that perform call processing functions. The other components are high speed LAN that connects IMs and core servers, and dual power system.

Fig. 5 shows the RBD for CS domain. Let the time to failures of core server, IMs, LAN and power be $Exp(\lambda_{ser}), Exp(\lambda_{IM}), Exp(\lambda_{LAN})$ and $Exp(\lambda_{pow})$ respectively. Then, reliability of CS domain, denoted by $R_{CS}(t)$, is given as:

$$R_{CS}(t) = e^{-(\lambda_{ser} + \lambda_{IM} + \lambda_{LAN})t} \times (2e^{-\lambda_{pow}t} - e^{-2\lambda_{pow}t}) \quad (6)$$

2) Reliability of PS domain: The two main components of PS domain are SGSN and GGSN. Fig. 6 shows the RBD of PS domain.

Let the time to failure of SGSN and GGSN be $Exp(\lambda_{SGSN})$ and $Exp(\lambda_{GGSN})$ respectively. Reliability of PS domain, denoted by $R_{PS}(t)$, is given as:

$$R_{PS}(t) = e^{-(\lambda_{SGSN} + \lambda_{GGSN})t} \quad (7)$$

Substituting $R_{CS}(t)$ and $R_{PS}(t)$ respectively, from equations (6) and (7) in equation (5), reliability of CN is obtained.

Finally, we compute the reliability of entire UMTS networks (denoted by $R_{net}(t)$). It is based on the CNs and their interconnection pattern and is given as:

$$R_{net}(t) = R_{arch}(t) \times R_{GCN}(t) \quad (8)$$

where $R_{GCN}(t)$ and $R_{arch}(t)$ denote the reliability of a group of CNs and inter-connection architecture (ring, star and SONET) respectively. $R_{GCN}(t)$ is obtained as follows: Out of a group of $K$ CNs, at least one should be in operation in order to provide connectivity to the mobile users. These $K$ CNs need not be identical. Then, $R_{GCN}(t)$ is given as below:

Define $S = \{1, 2, \ldots, K\}$ and $\mathcal{A}(S) = P(S) - \{\Phi\}$, where $P(S)$ is the power set of $S$. Then, $R_{GCN}(t)$ is given as:

$$\sum_{S_j \in \mathcal{A}(S)} \prod_{m \in S_j} R_{CN_m}(t) \prod_{k \in S_j} [1 - R_{CN_k}(t)] \quad (9)$$

where $R_{CN_m}(t)$ is given by equation (5).

In a ring architecture, the number of links is same as the number of CNs. The failure of any of the links result in outage of entire network. In a star topology, number of links connecting CN is one less than the number of CNs as all CNs are connected through a central CN. In SONET dual rings, the number of links is same as that in a ring, however, there is a spare link that acts as a backup link. The reliability for three different inter-connection architectures will be given by:

$$R_{arch}(t) = \begin{cases} (R_{link}(t))^K, & \text{Ring} \\ (R_{link}(t))^{K-1}, & \text{Star} \\ (1 - (1 - R_{link}(t))^2)^K, & \text{SONET} \end{cases} \quad (10)$$

where $K$ is the number of CNs and $R_{link}(t)$ denotes reliability of a single link connecting two CNs. With the assumption of exponential distribution for the time to failure of a single link, $R_{link}(t)$ is given as $e^{-\lambda_{link}t}$. 

---

**Fig. 4.** RBD for CN Reliability.

**Fig. 5.** RBD for CS Domain Reliability.

**Fig. 6.** RBD for PS Domain Reliability.
IV. Survivability Metrics Evaluation

In [6], percentage of MS affected was identified as the survivability metrics for level 1 and level 2 respectively. In [24], average number of incoming calls lost at a mobile due to database access delay was derived, with the assumption that the BS and BSC at the lower levels are in working conditions. In accordance with the existing survivability metrics, we derive these measures for UMTS networks. The percentage of MS affected due to the failure of Node B and RNC in subsections IV-A and IV-B respectively. In subsection IV-C, we evaluate the average loss of incoming calls at a mobile on the failure of HLR on releasing the assumption that Node B and RNC are in working conditions.

A. Impact of Node B failures

Different types of failures for Node B at level 1 are discussed in Section III-A. Recall that Node B is fully operational in states $U$ and $U_P$, partially up in states $U_H$, $U_S$, $U_{SP}$ and $U_{PHS}$ and completely down in state $D$. The impact of failures in an $i^{th}$ Node B is quantified by the percentage of MS turned away when it is in partially up states and down state and is given as:

$$I_{nodeB_i}(t) = \left\{ (P^H_i(t) \times (\pi_{U_H}(t) + \pi_{U_{PH}}(t))) \right\} + \left\{ (\pi_D(t)) \right\} + \left\{ (P^S_i(t) \times (\pi_{U_S}(t) + \pi_{U_{SP}}(t))) + ((P^S_i(t)P^H_i(t)) \times (\pi_{U_{HS}}(t) + \pi_{U_{PHS}}(t))) \times 100\% \right\}$$

(11)

where $P^H_i(t)$ and $P^S_i(t)$ are call blocking probabilities due to hardware and software failure respectively. $\pi_{U_H}(t)$, $\pi_{U_S}(t)$, $\pi_{U_{PH}}(t)$, $\pi_{U_{SP}}(t)$, $\pi_{U_{HS}}(t)$, $\pi_{U_{PHS}}(t)$ and $\pi_D(t)$ are the transient probabilities of the states $U_H$, $U_S$, $U_{PH}$, $U_{SP}$, $U_{HS}$, $U_{PHS}$ and $D$ respectively, and are obtained by solving the state transition diagram shown in Fig. 2. We now determine call blocking probabilities for hardware and software failures, denoted by, $P^H_i(t)$ and $P^S_i(t)$ respectively.

1) Computation of $P^H_i(t)$: $P^H_i(t)$ denotes the probability that MS are turned away due to hardware failures at any time $t$. To compute $P^H_i(t)$, we first construct the reliability model at top level that accounts for hardware failures at redundant CUPS. Then, a lower level performance model is constructed to compute call blocking probability for each state of the reliability model. The top level reliability model is then turned into a MRM [23], where the reward rates come from lower level pure performance model and are supplied to top level reliability model.

Reliability model for hardware failure: UMTS networks include a redundant CUP at Node B. On the failure of an active CUP, MS are unable to access the channels available at Node B. Operation and Maintenance Center of the networks monitor the performance, detect the failures of CUP and initiate switch over process to bring redundant CUP into operation. On successful switch over with probability $p$, synchronization of new CUP is started. In case of unsuccessful switch over with probability $(1 - p)$, switch over function is first recovered and then the switch over process begins. We assume the time to failure detection, switch over process, repair of switch over and synchronization be $Exp(\delta_t)$, $Exp(\delta_r)$, $Exp(\mu_r)$ and $Exp(\delta_s)$ respectively. During the switch over and synchronization, only $r_r\%$ (say) of total channels at Node B are available for non-real time calls, since these calls are delay-tolerant and do not have hard QoS bounds, while delay sensitive real time calls are lost. On the other hand, during failure detection and unsuccessful switch over, both the types of calls are rejected.

Failure and synchronization behavior of CUP is modeled as a CTMC shown in Fig. 7, where the state space is given by

$$\begin{align*}
I = \{ (u, u), & (d, u), (r, u), (s, d), (d, f), (u, d), (d, d) \}
\end{align*}$$

State $(u, u)$ is an initial state with both CUPS functioning. On the failure of an active CUP with an exponential rate $\lambda_f$, the system moves to state $(d, u)$. After detecting the failure, a switch over process starts to bring the redundant CUP as an active CUP (state $(r, u)$). If the switch-over fails, the system moves to state $(d, f)$. If the switch-over is successful, the synchronization process starts for the new active CUP (state $(s,d)$) with rate $\delta_s$. The system then moves to the state $(u, d)$. New active CUP may fail with exponential rate $\lambda_f$ and the system moves to down state $(d, d)$. In states $(u, u)$ and $(u, d)$, both real time and non-real time calls are supported, whereas in states $(d, u)$, $(d, f)$ and $(d, d)$ both the types of calls are completely blocked. In states $(r, u)$ and $(s, d)$, only non-real time calls are sustained with reduced capacity $r_r\%$ of total channels available at Node B. Thus, in this model, $(u, u)$ and $(u, d)$ are the up states, $(r, u)$ and $(s, d)$ are the partial up states, and $(d, u)$, $(d, f)$ and $(d, d)$ are the down states. The transient probability $\pi_i(t)$ for each of the state $i, i \in I$ is obtained by solving the CTMC shown in Fig. 7. Next, we construct the lower level performance model.

Performance Model

Suppose a Node B has $n$ channels in its channel pool. A real time and a non-real time call arriving at a Node B is accepted only if an idle channel is available at Node B. Otherwise, the call is lost. Let the inter-arrival time and channel holding time (CHT) for both types of calls be $Exp(\lambda)$ and $Exp(\mu)$ respectively. Let $X(t)$ be the number of busy channels at time $t$. Then, $\{X(t), t \geq 0\}$ is a CTMC and is modeled as a $M/M/n/n$ loss system[23].

Call blocking probability, labeled as $p_b(n)$, is the probability that all $n$ channels are busy, which is same as the probability that state of system is $n$. Thus, $p_b(n) = \pi_n$, where, $\pi_n$ is the steady state probability of being in state $n$. Using Erlang B formula, $p_b(n)$ is given as

$$p_b(n) = \pi_n = \frac{(\frac{\lambda}{\mu})^n}{\sum_{j=0}^{n} (\frac{\lambda}{\mu})^j} \frac{1}{n!}$$

(12)

Using this performance model, we now obtain blocking probability for each state of the top level reliability model.
shown in Fig. 7. Let \( P_b(i) \), denote the steady state call blocking probability of state \( i \), \( i \in I \) and \( I \) is the state space of the top level reliability model. In states \( i = (u,u) \), \( (u,d) \), both real time and non-real time calls can access the network only if an idle channel is available and are blocked otherwise. Call blocking probability in these states is obtained from the lower level performance model and is given by \( P_b(u,u) = P_b(u,d) = p_b(n) \) where \( p_b(n) \) is given by equation (12). In states \( i = (r,u) \), \( (s,d) \), \( x = r \% \) of \( n \) channels are available and the system can support only non-real time calls with these channels. These calls are blocked if all the \( x \) channels are busy. Further, we assume that, out of the total traffic, \( r_b \% \) is non-real time calls and \( (1 - r_b) \% \) is real time calls. The value of \( r_b \) depends on the call arrival rates for non-real time and real time calls. The call blocking probabilities for non-real time calls is \( p_b(x) \) with \( r_b \% \) of non-real time calls and the call blocking probabilities of real time calls is \( 1 \) with \( (1 - r_b) \% \) of real time calls. Hence, call blocking probabilities for the states \( (r,u) \), \( (s,d) \), \( x = r \% \) of \( n \) channels is given by \( P_b(r,u) = P_b(s,d) = r_b \times x + (1 - r_b) \times 1 \). Here, \( p_b(x) \) denotes the blocking probability of non-real time calls and is obtained from the lower level performance model with \( n = x \). In states \( i = (d,u) \), \( (d,f) \) and \( (d,d) \), both real time and non-real time calls are blocked and therefore call blocking probability in each of these states are \( 1 \).

For computing the total call blocking probability \( P_{b}^{\text{t}}(t) \), we utilize MRM which assigns a reward rate to each state of the top level reliability model. Given that the state of the system is \( i \), \( i \in I \), its call blocking probability, \( P_b(i) \), is assigned as a reward rate to the state \( i \). This follows from the fact that in each of the state \( i \), the calls (real time or non-real time) are blocked with probability \( P_b(i) \). The total call blocking probability is now given by the expected reward rate which can be written as

\[
P_{b}^{\text{t}}(t) = \sum_{i \in I} P_b(i) \pi_i(t) \quad (13)
\]

where \( \pi_i(t) \), \( i \in I \), is the probability that the state of the system is \( i \) at time \( t \). These state probabilities are obtained by solving the CTMC in Fig. 7. After computing the call blocking probability due to hardware failures at CUP in Node B, we next determine the call blocking probability due to software failures at Node B.

2) Computation of \( P_{b}^{S}(t) \): \( P_{b}^{S}(t) \) represents the transient probability that MSs are turned away due to the software failures which include failures in channel encoding/decoding, segmentation, spreading and scrambling [18]. This leads to the loss of channels and in turn, the arriving calls are blocked. To compute \( P_{b}^{S}(t) \), we first construct the top level reliability model that accounts for failure and recovery of channels. An automatic recovery mechanism at CUP accounts for the recovery of failed channels. Then, with \( i \) available channels, the call blocking probability is obtained from the performance model at the lower level. The top level reliability model is turned into a MRM, where the reward rates are taken from lower level pure performance model and are supplied to top level reliability model.

Reliability model for software failures: Each Node B provides a number of channels, say ‘\( n \)’ for mobile users to communicate with the system. Normally, one channel is dedicated to transmit the control signal information. Such a channel is called a control channel. The total number of available talking channels is \( (n - 1) \). With probability \( q (1 - q) \), the failure occurs in a control (traffic) channel and the time to channel failure is \( \text{Exp}(\alpha) \). The detailed failure and repair behavior is shown in Fig. 8.

Failure of a control channel will cause the entire Node B to fail since Node B cannot transmit the necessary information to the RNC. Automatic software recovery mechanisms either recover the control channel failure with probability \( c_1 \) or reconfigures an idle, non-failed traffic channel to an access channel with probability \( c_2 \). With probability \( 1 - (c_1 + c_2) \), failed access channel is neither recovered nor reconfigured and as a result, Node B moves to down state \( D \). On the other hand, the failed traffic channel is recovered (not recovered) with probability \( p(1 - p) \). Let the time to recovery for a failed control (traffic) channel be \( \text{Exp}(\delta) \) and the time to repair the failed channel is \( \text{Exp}(\tau) \).

With these assumptions, the underlying stochastic process is a reducible CTMC with state \( D \) as an absorbing state. The state diagram is depicted in Fig. 8.

From state \( i, 1 \leq i \leq n \), where \( i \) represents the number of non-failed channels, the system moves to state \( F_i \), \( (S_i) \) with an occurrence of a control (traffic) channel failure. We suppose that no other channel failure takes place until a traffic and control channel failure has been recovered, that is, no transitions from state \( S_1 \) to \( F_1 \) or vice-versa, occurs. In state \( S_i \) and \( F_i \), the recovery of failed channel is ongoing and therefore, the network traffic arriving to Node B is blocked. In this reliability model, state \( i, 1 \leq i \leq n \), is an up state, and remaining states \( D, F_1, 1 \leq i \leq n \), and \( S_i, 2 \leq i \leq n \) are the down states. The time-dependent state probabilities are obtained by solving this reducible continuous time Markov chain (CTMC).

Given that \( i \) non-failed channels are available, the call blocking probability is obtained from the lower level performance model discussed earlier in this section. Thus, when \( i(1 \leq i \leq n) \) channels are available at time \( t \), the call blocking probability \( P_{b}(i) \) is given as \( P_{b}(i) = p_{b}(i) \) where \( p_{b}(i) \) is the probability that all \( i \) channels are busy and it is substituted from Equation (12). The call blocking probabilities in other
states is 1.

To obtain the call blocking probability \( P^b_S(t) \), we assign call blocking probability \( P_b(i) \) obtained from the lower level performance model as the reward rate to each up state \( i \) in the top level reliability model. For each down state, \( D, F_i, (1 \leq i \leq n) \), and \( S_i, (2 \leq i \leq n) \), the corresponding call blocking probabilities \( P_b(D), P_b(F_i) \) and \( P_b(S_i) \) are assigned as the reward rates.

Then, \( P^b_S(t) \) is expressed as expected reward rate and is given by

\[
P^b_S(t) = \sum_{i=1}^{n} \pi_i(t)P_b(i) + \sum_{i=1}^{n} \pi_{F_i}(t)P_b(F_i) + \sum_{i=1}^{n} \pi_{S_i}(t)P_b(S_i) + \pi_D(t)P_b(D) \tag{14}
\]

where \( \pi_i(t) \) gives the time-dependent probability of the state \( i \) in the top level reliability model. Finally, substitute the values of \( P^b_H(t) \) and \( P^b_S(t) \), respectively, from equations (13) and (14), in equation (11) to obtain the impact of failures in Node B. In the coverage area of RNC, MSs are affected by failure of RNC, the percentage of affected MSs is obtained from the lower level UMTS network architecture.

B. Impact of RNC failures

We next derive an analytical expression for the percentage of MSs affected due to the failures in RNC at level 2 of the UMTS network architecture.

Combining the impact of failures at a group of Node Bs as well as failure of RNC, the percentage of affected MSs is given as

\[
I_{nodeB}(t) = \frac{1}{M} \sum_{i=1}^{M} I_{nodeB,i}(t) \tag{15}
\]

where \( I_{nodeB,i}(t) \) is the impact of failures at the \( i \)th Node B and is given by equation (11).

C. Impact of HLR failures

Due to the HLR failure, the incoming calls arriving at a mobile could not access the location information of the called mobile from the database and are eventually lost. Average number of incoming calls lost measures the impact of failure of HLR associated with CN at level 3. To ensure the minimum loss, HLR is made to function again through the back ups that are maintained regularly. However, the information stored in the back-ups is obsolete for the MS who changed their location during the failure period, therefore, incoming calls still could not be delivered to the called mobiles.

Location update procedures implemented at HLR update the HLR in any of the two cases: First, HLR receives a call delivery request for a MS whose information is obsolete. Note that an incoming call originates from a non-failed Node B in the coverage area of non-failed HLR. Second, MS changes its location and invokes a location update procedure at the HLR. A location update procedure is invoked only if MS moves to a non-failed Node B, for, if the MS moves to a failed Node B or a Node B controlled by a failed RNC, it could not trigger a location update. In case, HLR does not receive a call delivery or location update requests within the time interval \( s \), it sends automatic location update requests to MS at a period \( s \). Thus, the maximum time for recovery is \( s \). The efficiency of the location update procedures is inversely related to the average loss of incoming calls to a mobile. An inefficient location update procedure increases the database access delay, thereby, increasing the average number of incoming calls to a mobile that are lost.

Let the number of call arrivals at a mobile during HLR failure and recovery period be represented by \( X_F \) and \( X_R \) respectively. \( X_{tot} = X_F + X_R \) denotes the total number of call arrivals at a mobile during failure and recovery period. Average loss of incoming calls is the sum of average number of calls arriving at a mobile during the failure and recovery period, that is, \( E(X_{tot}) = E(X_F) + E(X_R) \). To evaluate \( E(X_{tot}) \), we make the following assumptions: Calls to a mobile arrive at the HLR according to Poisson process with parameter \( \lambda_u \). Let \( T_c \) represents the inter-arrival time for call delivery requests and is \( Exp(\lambda_u) \). The inter-arrival time between location update requests is \( Exp(\lambda_u) \). Recall that the location update request is sent when MS moves to a non-failed Node B controlled by a non-failed RNC. Let \( T_u \), inter-arrival time between location update requests whenever a location update procedure is triggered, is \( Exp(p\lambda_u) \). Here \( p \) gives the probability of moving to a non-failed Node B controlled by a non-failed RNC and is obtained as follows: We suppose that the MS can move to another Node B with uniform probability \( \frac{1}{2} \). Where, \( M \) is the number of Node Bs that are controlled by a RNC. Probability that a MS moves to a different RNC is also assumed to be uniformly distributed with probability \( \frac{1}{N} \), where \( N \) RNCs are managed by the MSC server. Then, \( p \) is given by \( p = \frac{1}{2} \cdot \frac{1}{N} \). Let \( t_c \) and \( t_u \) be the time interval of the instant at which HLR becomes functional and an incoming call (location update) arrives. Then, by the memoryless property of exponential distribution, \( t_c \) and \( t_u \) are exponentially distributed with parameters \( \lambda_c \) and \( p\lambda_u \) respectively. Time of HLR failure, denoted by random variable \( U \), with \( f_U(\cdot) \) as probability density function, is assumed to be exponentially distributed with parameter \( \lambda_r \). With these assumptions, we next compute the \( E(X_F) \) and \( E(X_R) \).

**Expected number of incoming calls at a mobile during the failure period**

Probability that \( k \) calls arrive at a mobile during a failure period of length \( U = u \), is

\[
P(X_F = k | U = u) = \frac{(\lambda_u u)^k}{k!} e^{-\lambda_u u}, \quad k = 0, 1, \ldots
\]
Fig. 9. Reliability of Node B, Node B group and RNC.

Consequently,

\[ P(X_F = k) = \int_0^{\infty} Pr(X_F = k| U = u) f_U(u) du = \int_0^{\infty} \frac{(\lambda_a u)^k}{k!} e^{-\lambda_a u} \lambda_r t, e^{-\lambda_r u} du = \frac{\lambda_a^k \lambda_r}{(\lambda_a + \lambda_r)^{k+1}}. \]

Expected number of incoming calls lost at a mobile during failure period is

\[ E(X_F) = \sum_{k=0}^{\infty} \frac{k \lambda_a^k \lambda_r}{(\lambda_a + \lambda_r)^{k+1}} = \frac{\lambda_a}{\lambda_r}, \quad \lambda_r > 0 \]

**Expected number of incoming calls at a mobile during the recovery period**

Average number of incoming calls arriving at a mobile during the recovery period is given by \( E(X_R) = \lambda_s E(L) \). \( L \) represents the length of recovery period which is \( \min(s, t_c, t_u) \) and \( s \) denotes the maximum recovery time. The probability density function \( f_l(t) \) of the random variable \( L \) is given as [24]

\[ f_l(t) = \begin{cases} (p \lambda_a + \lambda_s) e^{-(p \lambda_a + \lambda_s)}t, & 0 \leq t < s \\ e^{-(p \lambda_a + \lambda_s)} \delta(t-s), & t = s \\ 0, & \text{otherwise} \end{cases} \] (17)

\( \delta(\cdot) \) is dirac delta function. The average length of the recovery period, \( E(L) \), is

\[ E(L) = \frac{1 - e^{-(\lambda_c + p \lambda_a) s}}{\lambda_c + p \lambda_a}. \]

Therefore, the expected number of calls arriving at a mobile in the coverage region of failed database is given as

\[ E(X_F) + E(X_R) = \frac{\lambda_a}{\lambda_r} + \lambda_a E(L) \] (18)

This survivability metric is also important to evaluate the expected cost associated with HLR failure [24], [25].

**V. COMPATIBILITY WITH OTHER CELLULAR NETWORKS**

The proposed analytical model for studying the reliability and survivability attributes of UMTS network is based on step-by-step evaluation for the network elements at each level and finally for the entire network. This model makes use of CTMC, semi-Markov process and RBDs. The analytical model is equally useful for analyzing the reliability and survivability attributes of several other cellular networks such as All IP-UMTS network and CDMA2000 network.

The hierarchical architecture of All IP-UMTS network possesses is analogous to UMTS network. However, All IP-UMTS network consists of packet switched domain which comprises of SGSN, GGSN and network elements that support voice over internet protocol (VoIP) and other multimedia technology [14]. CDMA2000 wireless network also possesses hierarchical architecture with BS at level 1, BSC at level 2 which are controlled by MSC at level 3 [26]. After identifying the sub-components of network elements in All IP-UMTS and CDMA2000 networks, reliability and survivability model is constructed for computing these attributes at each level of the network architecture and finally for the entire networks. However, due to page limitations, we are not showing the broad application of the proposed model for computing the reliability and survivability attributes of these cellular networks.

**VI. PRACTICAL INSIGHTS**

In this section, we give numerical illustration to demonstrate how the proposed reliability and survivability models are useful in studying the reliability and survivability attributes for UMTS network. The vendors/manufacturers provide the information about the downtime and availabilities of middleware components of the network elements [4], [8]. Computing the failure and recovery rates from the information about the downtime and availability of the components is out of the scope of this paper. However, the readers interested in the computations for failure and recovery rates may refer to [27], [28]. For the purpose of numerical illustration, we set the parameter values of several components of network elements as follows: In this example, let the values of the parameters for Node B be set as: \( \lambda_p = 0.01 hr^{-1}, \lambda_{hr} = 0.008 hr^{-1}, \lambda_c = 0.10 hr^{-1}, \lambda_b = 0.001 hr^{-1} \) and \( \lambda_l = 0.01 hr^{-1} \). The values of parameters for RNC are assumed as: \( \lambda_{pow} = 0.015 hr^{-1}, \lambda_{eth} = 0.001 hr^{-1}, \lambda_{cmd} = 0.0001 hr^{-1}, \lambda_{link} = 0.0005 hr^{-1}, \lambda_{SFT} = 0.001 hr^{-1}, \lambda_{CM} = 0.0005 hr^{-1} \). Finally, the parameters of CN are set as: \( \lambda_{SER} = 0.0005 hr^{-1}, \lambda_{IM} = 0.0001 hr^{-1}, \lambda_{LAN} = 0.0001 hr^{-1}, \lambda_{pow} = 0.0005 hr^{-1}, \lambda_{GGSN} = 0.00002 hr^{-1}, \lambda_{GGSN} = 0.00003 hr^{-1}, \lambda_{DB} = 0.00001 hr^{-1} \) and \( \lambda_{link} = 0.002 hr^{-1} \).

**A. Impact on Network Reliability**

Substituting from the set of parameters given above in equation (1), we compute the reliability of Node B. Further, we assume that a group of \( M = 5 \) identical Node Bs is controlled by a RNC. Then, reliability of a group of Node Bs and the RNC controlling this group is computed from equations (2) and (3) respectively. Fig. 9 shows the reliability of a Node B, a group of Node Bs and a RNC. We observe that the reliability for a group of Node Bs is more than a Node B since MSs in
In its coverage area. Therefore, it is observed from Fig. 9 that RNC handling the control functions of a group of Node Bs is less than the group of Node Bs handled by it. The number of Node Bs that gives the maximum achievable reliability is determined and is helpful for deciding the parameters for the RNC. Second, reliability of RNC (number of Node Bs = 6), is depicted in Fig. 10 (b) for the case of with and without the parallel ethernet chassis. It is observed that RNC with a parallel ethernet switch is more reliable in comparison to RNC without a parallel ethernet switch.

We next present numerical results to show the effect of inter-connecting architecture on reliability of the entire UMTS networks. Substituting the parameters for CN in equations (8) and (9), network reliability for three inter connection architectures namely, ring, star and SONET dual ring architectures are shown in Fig. 11(a). It is observed that ring architecture gives the lowest reliability while fault tolerant SONET ring accounts for highest reliability. Also, we observe that reliability of star architecture lies in between that of ring and SONET inter connection cases. This is explained from the fact that in case of ring architecture, UMTS networks is vulnerable to outage on the failure of any link or CN, whereas, with star inter connection, only the central CN could affect the whole network. On the other hand, SONET dual rings allow self-healing, thereby increasing the system reliability.

In Fig. 11(b), the reliability of UMTS networks with number of CN =1, 2 and 4, where CNs are inter-connected through SONET dual ring architecture, is plotted using equations (8) and (9). It is observed that reliability of UMTS networks improves as the number of CNs increase from 1 to 2, but further increasing the number of CN to 4 decrease the network reliability. It can be concluded from these observations that on increasing the number of CN will not increase network reliability. Incorporating more CNs will only incur heavy maintenance cost instead of enhancing the network reliability. Therefore observations of Fig. 11 are beneficial in deciding the optimal number of CN while designing the UMTS networks.

B. Impact on Network Survivability Metrics

In this subsection, we present the numerical results to illustrate the impact of failures on the network performance in terms of survivability metrics-percentage of MS affected and average loss of incoming calls at a mobile. At a Node B, hardware and software failure impacts are obtained separately through MRM. We set the values of parameters for the top level hardware reliability model shown in Fig. 7 as: $\lambda_f = 0.003/hr^{-1}$, $\delta_d = 0.005/hr^{-1}$, $\delta_r = 0.002/hr^{-1}$, $\delta_s = 0.001/hr^{-1}$, $\mu_r = 0.001/hr^{-1}$, $p = 0.95$, $r_r = 80\%$, and $r_n = 25\%$. The values of the parameters for the software reliability model shown in Fig. 8 are set as: $\alpha = 0.004/hr^{-1}$, $\delta = 0.5/hr^{-1}$, $\beta = 0.002/hr^{-1}$, $\tau = 0.4/hr^{-1}$, $p = 0.90$, $q = 0.25$, $c_3 = 0.65$ and $c_2 = 0.22$. At the lower level performance model, we consider a Node B with $n = 15$ channels, call arrival rate $\lambda = .3$ calls/hr and exponential channel holding times with parameter $\mu = 0.4/hr^{-1}$.

Substituting for the parameters in equation (11), impact of failures at Node B are obtained in terms of the percentage of MS affected due to failures at Node B. Fig. 12 (a) shows the variation of impact of RNC and Node B failure on system survivability. Note that the impact of RNC failure has been computed by using equation (16). It is observed that percentage of MS affected due to RNC failures is always higher than that of percentage of MS affected due to failures at Node B. This is explained from the fact that failure of RNC results in loss of connectivity to MS in all Node Bs in its control, whereas, failures at Node B affect MS in its coverage area only.
Next, we study the impact of failure of the database for a network with \( N = 4 \) RNCs and \( M = 5 \) Node Bs. We set the parameters as \( p = 3.5, \lambda_0 = 0.05 hr^{-1}, \lambda_c = 0.5 hr^{-1}, \lambda_u = 0.2 hr^{-1}, s = 0.2 hr \). On substituting these parameters in equation (18), the expected loss of incoming calls to a mobile are obtained and plotted in Fig. 12(b) against \( \lambda_r \). Note that time of HLR failure is \( EXP(\lambda_r) \). It is observed that average number of incoming calls lost decrease as \( \lambda_r \) increases since with an increase in the parameter \( \lambda_r \), mean time of failure of HLR decreases and hence the average number of calls lost also decreases. This observation is beneficial for the network designers to trade-off for database failure and the effect on the mobile users.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we presented analytical models for computing the reliability and survivability attributes of UMTS networks. Assuming exponential distribution for failure and recovery, reliability and survivability models are developed using semi-Markov process, Markov chains, reliability block diagrams and Markov reward models. This analytical framework is compatible with All-IP UMTS networks and CDMA2000 networks. From the numerical results, we conclude that (a) substantial gains in network reliability can be achieved by incorporating fault tolerance in UMTS networks architecture (b) increasing the number of network elements at any level beyond a maximum number will not enhance the network reliability and (c) the proposed model can provide important guidelines related to reliability, survivability and fault tolerance to the network designers of cellular networks. The proposed model can be further extended with non-exponential distribution for failures and repairs of the individual components in the UMTS networks, which will be investigated in our future research.

ACKNOWLEDGEMENT

This research was supported, partially by MHRD, India under the grant number RP 01626 and partially by Department of Science and Technology, India under the grant number RP 1907. The work of V.J. was supported by the Ph.D scholarship from CSIR, India. The excellent comments of the anonymous reviewers are greatly acknowledged that have helped a lot in improving the quality and readability of the paper.

REFERENCES


Vaneeta Jindal received the Ph.D in Mathematics from Indian Institute of Technology Madras, India in 1999. She is currently an associate professor at Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India. Before joining this institute, she was a post-doctoral fellow at the department of electrical and computer engineering, Duke University, USA and then was a research associate at the TR Labs, Winnipeg, Canada. His work has been published in several international journals and conferences. His current research interests include queueing theory, Markov modeling, and performance issues of wireless networks and dependability analysis of communication systems.

Upkar Varshney is an associate professor of CIS at Georgia State University. His interests include wireless networks, pervasive healthcare, and mobile commerce. He has chaired international conferences and has authored over 120 papers including 60 in journals. He has also received grants from National Science Foundation. He has won several teaching awards and has been an editor/guest editor for journals including IEEE TRANSACTIONS ON IT IN BIOMEDICINE, ACM/KLUWER MOBILE NETWORKS, IEEE COMPUTER, and DECISION SUPPORT SYSTEMS.