Performance Evaluation of IMS-based Core Networks in Presence of Failures

Maurizio Guida, Maurizio Longo, Fabio Postiglione
Dipartimento di Ingegneria dell’Informazione ed Ingegneria Elettrica (D.I.I.E.)
Università degli Studi di Salerno
Via Ponte Don Melillo, I-84084 Fisciano (SA), Italy

Abstract—In order to evaluate performance of mobile networks, it is necessary to consider the occurrence of random failures causing performance degradation and the consequent repair actions. This approach is especially suitable for next generation networks based on the Third Generation Partnership Project (3GPP) IP Multimedia Subsystem (IMS), as a consequence of the very high Quality of Service levels required by telecommunication operators’ subscribers. IMS core network signalling servers can be modeled as multi-state elements, where server states correspond to different performance levels. The number of sessions handled by a single server per time unit is one of the performance figure that can be considered. Given a demand profile, some redundancy techniques must be adopted to meet the typical requirements for a telecommunication network in terms of service availability and, in this paper, a redundancy optimization problem is solved by using a Universal Generating Function approach.

I. INTRODUCTION

Next Generation Networks (NGNs) [1] are convergent networks aiming to offer services to users connected to both wired and wireless infrastructures. They are typically based on the Internet Protocol (IP) and expected to provide seamless global access and integration between different services with a high level of end-to-end Quality of Service (QoS) [2]. Scenarios of this kind are also referred to as Beyond-3G (B3G) or 4G networks [3]. In order to guarantee a given QoS level to its customers, telecommunication industry must cope with network and service failures affecting network performance. Comprehensive approaches for telecommunication networks are needed to assess performance by accounting failures occurring in the network elements [4], [5]: These methodologies are often referred to as performability analysis, firstly proposed in [6], [7]. In this paper, we address some performance analysis techniques in presence of random failures in telecommunication networks based on the IP Multimedia Subsystem (IMS) [8]. In particular, we model core network signaling servers, namely the Call Session Control Functions (CSCFs), as multi-state elements (each state corresponds to a different performance level), whose performance figures can be described by Markov models [9].

For every telecommunication system, a typical performance metrics related to the signaling server functionalities is the number of packets in the time unit forwarded by a single CSCF server to the next node [10]. Given a constant demand profile, we discuss the problem of meeting typical requirements for an IMS-based NGN in terms of signaling service availability, that is defined as the probability that the multi-state system is in one of the states where performance is not less than demand (acceptable states). The said problem is faced by adopting some redundancy technique for some signaling network elements. Since the costs of IMS signaling nodes can be different, some reliability optimization approaches are necessary in network scenarios of practical interest.

The paper is organized as follows. In Section II, we introduce the IMS system proposed by 3GPP and its main features and performability issues. In Section III, we describe a multi-state performance model suitable for the IMS core network signaling servers. In Section IV, we address how to meet an availability requirement for the system by adopting a parallel redundancy for signaling functionalities, given a service demand profile. A redundancy optimization problem is solved by means of the Universal Generating Function. In Section V, the proposed methodology is applied to a realistic example of IMS-based network scenario. In Section VI, we conclude the paper.

II. THE IP MULTIMEDIA SUBSYSTEM

Next generation mobile networks rely on the IP Multimedia Subsystem (IMS) [8], as stated by the Third Generation Partnership Project (3GPP), whose mission is standardizing next generation mobile systems. The IMS architecture is based on IP version 6 (IPv6) and provides both real-time audio/video communications and data transport services. It is possible to connect IMS core network to the Public Switched Telephone Network (PSTN) and Internet by means of specific gateways. The wireless connections between a User Equipment (UE), representing the mobile terminal, and the core network are provided by a Radio Access Network (RAN). RANs can be based on different radio technologies and coexist in the same telecommunication network. The 3GPP adopted Session Initiation Protocol (SIP) [11] as the signaling protocol for handling real-time audio/video sessions. To this aim, some new network nodes have been introduced into the General Packet Radio Service (GPRS) core network: Signaling flows are managed by the Call Session Control Function (CSCF) nodes, whereas multimedia flows are transported by the Real-time Transport Protocol (RTP) [12] and traverse generally different paths in the network. The CSCF functionalities are handled by three different servers: (i) the Proxy CSCF (P-CSCF), usually located in the Visited Network, is the first contact point to the
IMS network for the UE, and forwards the SIP signaling to the subscribers’ Home Network (HN), typically the operator to which the user is subscribed; (ii) the Serving CSCF (S-CSCF) is located in the HN and manages multimedia sessions; (iii) the Interrogating CSCF (I-CSCF), located in the HN, is responsible for selecting the appropriate S-CSCF for all the requests entering the HN. Another important signaling server in the Home Network is the Home Subscriber Server (HSS), a database containing users’ details retrieved by S-CSCFs, during the user registration phase, or I-CSCFs, during the call set-up procedures. The HSS databases are interrogated by means of the Diameter protocol [13]. A typical network scenario is shown in Fig. 1: it describes the call set-up of a real-time session between two roaming UEs and includes some signaling messages exchanged among UEs and IMS nodes, two Home Networks reproducing network domains of telecom operators to which users are subscribed, two Visited networks that represent service providers networks with two different RANs the roaming terminals are attached to. In this paper, it is assumed that UEs have already notified their actual location to their home networks, during the registration phase.

A real-time session can be established between two registered UEs if and only if all signaling functionalities are available during call set-up procedure, because the signaling network can be modeled as a series system [14]. Since all servers involved in the call set-up session are failure-prone, some reliability methodologies are necessary to evaluate the overall IMS-based system availability, which, in line of principle, should be comparable to the PSTN one for voice services. In addition, IMS nodes have typically different performance levels and several failure and repair modes with various effects on the entire signaling network performance. Hence, a reliability/performance model should consider this distributed nature of call control functionalities, influenced by every single node behavior and the variety of failures, ranging from accidental failures [15] of software modules and/or hardware equipments to man-made faults due to malicious users activity [16], [17], [18].

### III. A PERFORMANCE AND RELIABILITY MODEL OF THE IMS SERVERS

In order to analyze the performance and reliability of an IMS-based signaling core network, a suitable model for a generic node must be introduced. We assume that all links of the RANs, GPRS and IP networks and all layer-2 and layer-3 network equipments are always reliable and they do not limit the overall performance of the signaling network: only IMS servers behavior can thus influence performability analysis of the telecommunication network. One of the performance metrics that can be studied is the number of call set-up sessions that an IMS core network is able to manage for its customers/users. Obviously, this metric is a function of the performance levels of the nodes involved in the signaling path between two generic UEs.

A typical CSCF server can be seen as composed by a service logic software part and a core part [10]. The former consists of a given number of concurrent instances of the same CSCF server, each of them able to serve a certain amount of call set-up requests; the latter represents the hardware setting (power plane and supply, processors, memories, operations/maintenance blades, switch cards, etc.) and operating system, peripheral drivers, etc. We suppose, for sake of simplicity, that the proposed representation is still valid for HSS, by assuming that different replicas of the user profile database are available on the same hardware. In real settings, different hardware parts and some load sharing equipments may be necessary. Moreover, a suitable model for an IMS signaling network is strongly influenced by the specific solution design of the service architecture and the software/hardware implementation of every IMS node. However, the methodological guidelines we are going to propose are still valid.

#### A. The proposed multi-state performance model

For the generic element $i$ in the signaling path, let be $n(i)$ the number of call set-up sessions serving capacity of each service module instance, whose number is $k(i)$, where the type of instances are denoted by the superscripts $(P_1), (S1), (I_2), (H2), (S2), (P2)$, corresponding to P-CSCF, S-CSCF, I-CSCF, S-CSCF, and P-CSCF servers, respectively.

A multi-state performance model of the IMS network node $i$ is based on the following assumptions: (i) each server instance and the core part of an IMS node can be either up or down; (ii) failures and repairs are statistically independent; (iii) server modules and core failures are statistically independent Homogeneous Poisson Processes (HPPs), resulting in independent and exponential inter-failure arrivals and constant hazard rates $\lambda^{(i)}_s$ and $\lambda^{(i)}_c$, respectively; (iv) server instances and core repair actions are statistically independent HPPs, resulting in independent and exponential repair times with rates $\mu^{(i)}_s$ and $\mu^{(i)}_c$, respectively. Those software and core fault and repair rates might be derived by inferring real operation data.

The resulting model is the Continuous-Time Markov Chain (CTMC) reported in Fig. 2, where the states $j = 0, ..., k^{(i)}$. 

---

**Fig. 1.** A call set-up scenario involving two roaming User Equipments, attached to different Radio Access Networks. Users are subscribers of different telecom operators.
indicate that \( j \)-out-of \( k^{(i)} \) server instances of the node are operational, given that the core part of the element is working; the state \( j = -1 \) is related to the condition that the core part is not working; when the CTMC is in one of the states \( j = 0, \ldots, k^{(i)} \), transitions happen between states differing by one active module (HPPs are regular processes) or toward the state \( j = -1 \) (core failure); when CTMC is in the state \( j = -1 \), one transition is assumed possible, because a repair action on the core requires typically a costly activity of an expert technician that is generally concluded by a complete hardware/software restoration of the damaged node.

Since some parallel redundancy will be introduced, we represent the CTMC states, corresponding to the node performance levels, by the set

\[
g^{(i)}_1 = \{g^{(i)}_{l,-1}, g^{(i)}_{l,0}, \ldots, g^{(i)}_{l,k^{(i)}}\},
\]

where \( g^{(i)}_{l,j} \), the performance level of \( l \)-th parallel element of the network subsystem \( i \) in the state \( j \in \{-1, \ldots, k^{(i)}\} \), is \( g^{(i)}_{l,j} = \max\{j, n^{(i)}\} \).

The stochastic process \( G^{(i)}_i(t) \in g^{(i)}_1 \) represents the performance level of the \( l \)-th parallel element of the subsystem \( i \) at any instant \( t \geq 0 \). The probabilities associated with the different states (performance levels) of such a system at any instant \( t \) can be represented by the set \( p^{(i)}_{l,j}(t) = \{p^{(i)}_{l,-1}(t), p^{(i)}_{l,0}(t), \ldots, p^{(i)}_{l,k^{(i)}}(t)\} \), where \( p^{(i)}_{l,j}(t) = \Pr\{G^{(i)}_i(t) = g^{(i)}_{l,j}\} \), and \( \sum_{j=-1}^{k^{(i)}} p^{(i)}_{l,j}(t) = 1 \), for any \( t \).

The state probabilities of the CTMC in Fig. 2 can be derived from the system of differential equations

\[
\begin{aligned}
\frac{dp^{(i)}_{l,-1}(t)}{dt} &= -\mu^{(i)}_{l,c} p^{(i)}_{l,-1}(t) + \lambda^{(i)}_c \sum_{h=0}^{k^{(i)}} p^{(i)}_{l,h}(t) \\
\frac{dp^{(i)}_{l,0}(t)}{dt} &= -\left(\lambda^{(i)}_c + k^{(i)} \mu^{(i)}_s\right) p^{(i)}_{l,0}(t) + \lambda^{(i)}_s p^{(i)}_{l,1}(t) \\
\frac{dp^{(i)}_{l,j}(t)}{dt} &= \mu^{(i)}_{l,c} p^{(i)}_{l,-1}(t) + \lambda^{(i)}_c \sum_{h=0}^{k^{(i)}} p^{(i)}_{l,h}(t) \\
\frac{dp^{(i)}_{l,k^{(i)}}(t)}{dt} &= \mu^{(i)}_{l,c} p^{(i)}_{l,-1}(t) + \mu^{(i)}_{l,k^{(i)}-1}(t) + \lambda^{(i)}_c p^{(i)}_{l,k^{(i)}-1}(t) - (\lambda^{(i)}_s k^{(i)} + \lambda^{(i)}_c) p^{(i)}_{l,k^{(i)}}(t) \\
\end{aligned}
\]

with the initial conditions \( p^{(i)}_{l,-1}(0) = p^{(i)}_{l,0}(0) = \ldots = p^{(i)}_{l,k^{(i)}-1}(0) = 0, p^{(i)}_{l,k^{(i)}}(0) = 1 \).

Since we are interested on a performance evaluation in long runs\(^1\), the steady-state probabilities \( p^{(i)}_{l,j} = \lim_{t \to \infty} \Pr\{G^{(i)}_i(t) = g^{(i)}_{l,j}\} \) are computed. They can be derived from (2), where all derivatives are posed equal to 0 and \( p^{(i)}_{l,j}(t) \) are substituted by \( p^{(i)}_{l,j} \), together with \( \sum_{j=-1}^{k^{(i)}} p^{(i)}_{l,j} = 1 \).

\(^1\)However, numerical experiments show that the state probabilities in (2) converge quickly to their limiting values and can be considered indistinguishable to them within a time interval of few days, if parameters values of typical network scenarios are adopted (see Sect. V).

\[\text{Fig. 2. A multi-state performance model of an IMS signaling node.}\]

\[\text{Fig. 3. A series-parallel RBD of the signaling system with CSCF server parallel redundancy.}\]

### IV. SIGNALING SYSTEM AVAILABILITY

The signaling path (henceforth the *signaling system*) is a Multi-State System (MSS), being an interconnection of MSSs of different kinds. It is considered successful when it is able to meet a required performance level, also called *demand*. As pointed out in Section II, two terminals can establish a communication if *all* the network elements managing the signaling messages are actually able to handle call set-up requests and answers, namely the signaling path is available to UEs perspective. In most practical cases, some redundancy may be necessary to meet a demand \( W(t) \). Here, we consider parallel redundancy of each signaling network element, as depicted in Fig. 3 by a Reliability Block Description (RBD).

Since the signaling system can be considered a series-parallel one with *flow dispersion* [4], the stochastic process describing its performance levels is

\[
G(t) = \min_{i} \sum_{l=1}^{h^{(i)}} G^{(i)}_l(t),
\]

where \( h^{(i)} \) is the number of parallel elements implementing redundancy for the network element \( i \), and \( i \in \{P1, S1, I2, H2, S2, P2\} \). Let \( h \) be the vector containing all these numbers. The process \( G(t) \) and hence the random
variable $G$ representing its steady-state values ($t \to \infty$) have values in the set
\[ g = \{g_1, \ldots, g_K\}, \]  
where $K = \prod_i (k^{(i)})^{(i)}$ is the number of different states of the overall signaling system, and the associated (steady-state) probability set is
\[ p = \{p_1, \ldots, p_K\}, \]  
where $p_k = \lim_{t \to \infty} \Pr\{G(t) = g_k\} = \Pr\{G = g_k\}$. The collection of pairs $g_k, p_k$, $k \in \{1, \ldots, K\}$, completely determines the steady-state output performance distribution.

According to \cite{4}, the signaling system instantaneous availability $A(t)$ is the probability that the MSS at time $t > 0$ is in one of the states where performance is not less than demand (acceptable states), namely $A(t) = \Pr\{G(t) - W(t) \geq 0\}$. For large $t$, the MSS initial state has no practical influence on its availability. Therefore, given a constant demand level $W(t) = w$, the stationary signaling system availability $A_{IMS}(w)$ can be determined on the base of steady-state output performance distribution, viz.
\[
A_{IMS}(w) = \sum_{k=1}^{K} p_k 1(g_k \geq w) = \sum_{g_k \geq w} p_k, \tag{6}
\]
where $1(\text{True}) = 1$, and $1(\text{False}) = 0$.

A convenient approach to evaluate system availability is based on the Universal Generating Function (UGF), originally introduced by Ushakov \cite{19}. The UGF of the stationary output performance of the IMS signaling system $G$, a discrete random variable whose values are in (4), is the polynomial-shape $u$-function
\[
u_{IMS}(z) = \sum_{k=1}^{K} p_k z^{g_k}, \tag{7}
\]
where the probabilities $p_k$ are those reported in (5). By introducing two fundamental operators, it is possible to compute the $u$-function $\nu_{IMS}(z)$, i.e. all $p_k$ and $g_k$ in (7), on the basis of the steady-state behavior of all elements of the system.

Since the signaling system can be modeled as a series-parallel one composed by flow transmission multi-state components with flow dispersion \cite{4}, \cite{20}, the overall output performance of $n$ elements connected in parallel is given by the sum of the $n$ capacities of the elements. The corresponding $u$-function is provided by the following $\pi$ operator:
\[
u_p(z) = \pi(u_1(z), u_2(z), \ldots, u_n(z)) = \frac{n}{\prod_{l=1}^{n} \left( \sum_{j=1}^{k_l}(\prod_{i=1}^{l} p_{l,j_i} z^{s_{l,j_i}}) \right)} . \tag{8}
\]
Thus, the $u$-function $u_{IP}^{(i)}(z)$ of the subsystem $i$ of the IMS signaling system in Fig. 3, where $i \in \{P1, S1, I2, H2, S2, P2\}$, can be rearranged as
\[
u_{IP}^{(i)}(z) = \pi(u_1(z), u_2(z), \ldots, u_{h^{(i)}}(z)) = \sum_{j_1=-1}^{k^{(i)}} \sum_{j_2=-1}^{k^{(i)}} \ldots \sum_{j_{h^{(i)}}=-1}^{k^{(i)}} \left( \prod_{l=1}^{h^{(i)}} p_{l,j_l} z^{h_{l,j_l}} \right), \tag{9}
\]
where $g_{l,j_l}^{(i)}$ is the performance level of $l$-th parallel element of the network subsystem $i$ in the state $j_l$ as in (1), and $p_{l,j_l}^{(i)}$ is the corresponding steady-state probability derived from (2).

On the other hand, when the components are connected in series, the overall output performance is simply the minimum of individual components performance levels. To compute the $u$-function for $m$ series components, the following $\sigma$ operator is adopted:
\[
u_s(z) = \sigma(u_1(z), u_2(z), \ldots, u_m(z)), \tag{10}
\]
where for a pair of generic components connected in series it is $\sigma(u_1(z), u_2(z)) = \sigma(\sum_{i=1}^{m} \varepsilon_i z^{e_i}, \sum_{j=1}^{n} \varphi_j z^{f_j}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \varepsilon_i \varphi_j z^{\min(e_i, f_j)}$, where the parameters $e_i$ and $f_j$ are the performance levels of the two components, while $\varepsilon_i$ and $\varphi_j$ are the corresponding steady-state probabilities, $\forall i = 1, \ldots, I$ and $\forall j = 1, \ldots, J$.

Finally, the UGF of the IMS signaling system in (7) can be obtained by substituting (9) $\forall i \in \{P1, S1, I2, H2, S2, P2\}$ into (10), viz.
\[
u_{IMS}(z) = \sum_{k=1}^{K} p_k z^{g_k} = \sigma(u_{(P1)}(z), u_{(S1)}(z), u_{(I2)}(z), u_{(H2)}(z), u_{(S2)}(z), u_{(P2)}(z)). \tag{11}
\]
This polynomial-shape $u$-function provides immediately all the signaling system performance levels (the exponents of $z$) and their corresponding steady-state probabilities (the corresponding coefficients), that can be used to compute the stationary IMS system availability $A_{IMS}$ using (6).

Let $C^{(i)}$ be the cost of a network element $i$, the overall cost of the signaling system is $C_{IMS} = \sum_i h^{(i)} C^{(i)}$. The problem is to find the system configuration $h^*$ providing the minimal cost by respecting the constraint of a required level of stationary availability $A_0$, viz.
\[
h^* = \arg\{C_{IMS}(h^*) \to \min |A_{IMS}(w, h^*) \geq A_0\} . \tag{12}
\]
This optimization problem is typically referred to as redundancy optimization problem \cite{21}, typically solved by means of some heuristic techniques.

V. A NUMERICAL EXAMPLE

In order to illustrate the methodology proposed in Sects. III and IV, an illustrative numerical example is presented. We assume that failure and repair rates are the same for every signaling node, i.e. $\lambda^{(i)} = 7.716 \times 10^{-7}$ sec$^{-1}$ (corresponding to 2 server faults per month) and $\mu^{(i)} = 1.667 \times 10^{-3}$ sec$^{-1}$ (corresponding to a mean repair time of 10 min), respectively, while node core failure and repair rates are $\lambda^{(c)} = 6.342 \times 10^{-8}$ sec$^{-1}$ (corresponding to 2 core faults per year) and $\mu^{(c)} = 3.472 \times 10^{-6}$ sec$^{-1}$ (corresponding to a mean repair time of 8 hours, i.e one working day of a technician on the node site), respectively. Furthermore, we consider CSCF nodes with 3 concurrent server instances, $k^{(i)} = 3$, $i \in \{P1, S1, I2, S2, P2\}$, whose serving capacities are $n^{(i)} = 1400$ sessions per time unit. On the other hand, the HSS node has only 2 concurrent database instances $k^{(H2)} = 2$, with serving capacities $n^{(H2)} = 2500$ sessions.
The solution we obtain is $h$ for the HSS telecommunication market, we can consider an unitary cost been provided by adopting an UGF-based computation of the number of call set-up requests handled by the IMS signaling system and the corresponding steady-state probabilities. In presence of random failures causing overall performance degradation, where network nodes has been modeled as an MSS with different performance levels. We focused the number of call set-up requests handled by the IMS signaling system and a parallel HSS. It represents a signaling system composed by 3 parallel elements of all the CSCF server traversed by signaling flow and 2 parallel HSS. The output performance levels of this configuration for the signaling system and the corresponding steady-state probabilities are reported in Table I. The availability of this IMS signaling system is $A_{IMS}(w) = A_{IMS}(w, h^*) = 1 - 6.791 \times 10^{-6}$, while its cost is $C_{IMS}(h^*) = 15.5$.

In real networks, sojourn times in the model states of the IMS system are typically not exponentially distributed, mainly due to non memoryless repair distribution times. However, preliminary results obtained by the adoption of Semi-Markov processes [9] show that the steady-state behavior of the signaling system seems to be a function of mean times to failure and repair mean times (typically, the former is much greater than the latter) and does not seem to depend markedly on the specific distribution functions. Then, CTMCs can be adopted for a first performance assessment of an IMS-based signaling system in presence of random failures, that can be subsequently refined by using Semi-Markov models.

VI. CONCLUSIONS

In this paper, a performance evaluation of a signaling IMS-based next generation mobile network has been carried out in presence of random failures causing overall performance degradation, where network nodes has been modeled as an MSS with different performance levels. We focused the number of call set-up requests handled by the IMS signaling system and a parallel redundancy optimization solution has been provided by adopting an UGF-based computation of the system steady-state performance behavior, given a required availability level and a constant service demand.

In future research activities, some alternative performance metrics, such as the call set-up delay, can be analyzed. Furthermore, similar approaches to those presented in this paper can be useful in presence of human-made failures (attacks), even if in this case, it is not simple to model the behavior of malicious users. Some promising approaches try to adopt traditional methodologies, derived from system dependability evaluation, to assess some security attributes such as availability, integrity and confidentiality [16], [17].

REFERENCES