Performance Modelling and Evaluation of a Mobility Management Mechanism in IMS-Based Networks*

Is-Haka M. Mkwawa[1], Demetres D. Kouvatsos[1], Wolfgang Brandstätter[2], Gerhard Horak[3], Alfons Geier[4], and Christoforos Kavadias[5]

[1] NetPEn - Networks and Performance Engineering Research Unit, University of Bradford, UK
{I.M.Mkwawa1,D.Kouvatsos}@bradford.ac.uk
[2] Telekom Austria TA AG, Vienna, Austria
wolfgang.brandstaetter@telekom.at
[3] Alcatel-Lucent Austria AG
Gerhard.Horak@alcatel-lucent.com
[4] Nokia Siemens Networks, Germany
alfons.geier@nsn.com
[5] Teletel AS, Greece
C.Kavadias@teletel.eu

Abstract. The 3^{rd} Generation Partnership Project (3GPP) for an IP multimedia subsystem (IMS) architecture defined a number of functional units, which exchange session initiation protocol (SIP) messages with register users and set up or terminate multimedia sessions. The processing of SIP messages, however, requires a significant amount of processing, queueing and transmission times with adverse implications on the overall performance of the IMS core architecture. This tutorial introduces a mobility management mechanism for an IMS testbed implemented by Nokia-Siemens as part of the EU IST VITAL project. The mechanism employs an open queueing network model (QNM) with priorities to represent the functional units and application servers of an IMS architecture during a handover process of SIP messages between different access networks. The QNM is analysed via the principle of maximum entropy and numerical experiments are employed to assess the performance impact of bursty traffic flows of SIP messages on the IMS architecture.

Keywords: IP multimedia subsystem (IMS), 3^{rd} Generation Partnership Project (3GPP), Session Initiation Protocol (SIP), quality-of-service (QoS), performance modelling, performance evaluation, mobility, GSM, WLAN, Wi-Fi, serving call session control function (S-CSCF), handover, maximum entropy (ME), generalized exponential (GE), queueing network model (QNM).

* This work was supported in part by the EC NoE Euro-FGI (NoE 028022) and in part by the EC IST project VITAL (IST-034284 STREP).

D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 594–611, 2011.

1 Introduction

The co-existence of several radio access technologies in today's mobile devices has provided end users with a wider choice, which is mainly driven by low cost and quality of service (QoS)[1] guarantees. Wireless Local Area Networks (WLANs) can provide high speed connections to access networks at a very cheap rate as compared to those of cellular access networks. These are the two major factors that motivate during handover an ongoing voice call session of a cellular access network to discover a WLAN access point. With WLAN coverage restricted to a relatively small area compared to that of a cellular network, the handover from WLAN to cellular network for an ongoing session is crucial for seamless service continuity whenever there is a weakening of WLAN signal strength.

The interoperability between WLAN and cellular network is made possible with the use of the mobility management mechanism associated with the IP Multimedia Subsystem (IMS) architecture, based on the recommendations of the 3rd Generation Partnership Project [1]. The IMS is an overlay system that serves the convergence of mobile, wireless and fixed broadband data networks into a common network architecture, where all types of data communications will be hosted in all IP environments using the infrastructure of the Session Initiation Protocol (SIP) to exchange messages to register users and set up or terminate multimedia sessions. As the processing of SIP messages requires the creation of states, starting timers and execution of filtering criteria, these procedures consume a significant amount of processing, queueing and transmission times with adverse implications on the performance of the overall IMS core architecture.

Mobility management is one of the main challenges facing IMS which is designated to deploy its services over a mixture of network access technologies such as wireless, mobile and fixed accesses. IMS should be able to deal sufficiently with the important issue of terminals mobility management, that nowadays is tackled at physical interface level resulting into the weakening of services robustness and increased probability of disrupted communication.

IMS is logically divided into two main communication domains, namely i) The data traffic domain (i.e., real time protocol packets consisting of audio, video and data) and ii) The SIP signalling traffic domain. IMS has entry and exit functional units or proxies, such as proxy P-CSCF, serving S-CSCF and interrogating I-CSCF call session control functions. These proxies exchange SIP messages to register users and setup/terminate multimedia sessions.

The VITAL IMS testbed [2] has a mobility management application server (MMAS) for voice continuity call with dual mode handset (DMH) roaming between global system for mobile communications (GSM) and WLAN-IMS. The active call is handed over from GSM to WLAN and vice versa. Handover in IMS specifically refers to voice call continuity from a SIP and a Real Time Protocol (RTP) based call moving from WLAN-IMS to cellular, or vice-versa. GSM-WLAN-IMS handover denotes the ability to change the access network from GSM to WLAN-IMS and vice versa, while a voice call is ongoing. This

[1] A list of all the acronyms used in this tutorial paper can be seen in Appendix I.

voice continuity call (VCC) feature applies to users with a GSM-IMS subscription and DMH. It is performed by the MMAS and the client on the DMH.

In this VITAL IMS architecture, the MMAS acts as back to back user agent (B2BUA). Whenever a handover is initiated for an existing call, the DMH first performs a new call that is routed to the MMAS. Then the MMAS operating as B2BUA connects the new call leg with the existing call by using a SIP Re-Invite [3]. Finally, the MMAS deletes the old call leg that is no longer used. The details of the handover concept are described below based on four specific handover scenarios:

- An IMS-IMS call is ongoing. Then the A-Party performs an IMS originating handover: IMS → GSM. The result is a GSM-IMS call.
- An IMS-IMS call is ongoing. Then the B-Party performs an IMS terminating handover: IMS→GSM. The result is an IMS-GSM call.
- A GSM-IMS call is ongoing. Then the A-Party performs an IMS originating handover: GSM→IMS. The result is an IMS-IMS call.
- An IMS-GSM call is ongoing. Then the B-Party performs an IMS terminating handover: GSM'IMS. The result is an IMS-IMS call.

In all handover scenarios, the following basic rules apply:

- On the IMS terminating side (for UE-B), the MMAS is the last IMS application server in the path.
- On the IMS originating side (for UE-A), the MMAS is the first IMS application server in the path.
- The voice application server (VAS) is invoked between originating MMAS and terminating MMAS.
- In the case of UE-A handover, the MMAS on the originating side performs the handover control (e.g. breaks the call).
- In the case of UE-B handover, the MMAS on the terminating side performs the handover control (e.g., breaks the call).
- The MMAS and the VAS need to be included into the SIP signalling path at the call setup such that a handover can be performed during the call. The MMAS cannot be included into the SIP signalling path after the call setup.

The handover switch time is significantly felt when a cellular/WLAN signal either is substantially weakened or abruptly disappears. In this context, the handover switching time is mainly due to the control messages traversing through a control path in order to setup a new call leg without disconnecting an ongoing call session.

The VITAL mobility specification [2] shows that the control functions in the path process several SIP messages per single handover request. With SIP messages often exceeding more than 1000 bytes, they are likely to have an adverse impact on the overall performance of the IMS architecture. Therefore, performance modelling and analysis may guarantee efficient and smooth IMS architecture operations whilst providing QoS guarantees. Within this framework, the design and implementation of IMS architecture will be well planned if performance metrics, such as server utilization, throughput and response time, are predicted quickly and efficiently.

Most of the ongoing research in the IMS field focuses on improving the development, engineering and performance impact of SIP messages. Cortes et al [4] have shown that SIP messages face significant challenges in SIP servers and, hence, special attention should be devoted on the design handling and parsing of SIP messages functionalities. Batterman et al [5] proposed SIP messages prioritization method in SIP servers. It was shown that the prioritized SIP messages can be processed without a major delay. A SIP offload engine was introduced in [6], whereby the scheme transforms SIP messages in binary format at the front end in order to parse the binary in the SIP stack. Rajagopal and Devetsikiotis [7] have proposed a modelling methodology that uses real life workload characterization, queueing analysis and optimization. The proposed methodology gives a systematic way for the selection of system design parameters in order to guarantee network performance whilst maximizing the overall system utility. Moreover, studies relating to the registration and session setup for accessing application servers in IMS [8] and the performance modelling and evaluation of handover mechanisms in IMS [9], respectively, were shown to specifically comply with the operational characteristics of an open central server queueing network model (QNM) [10]. Finally, Forte and Schulzrinne [11] introduced a new compression mechanism, based on the concept of templates that may be used in wireless networks by cellular operators. Such mechanism makes it possible to achieve the delay requirements of most time-critical applications, such as push to talk over cellular (PoC) in IMS.

This tutorial paper has its roots in the IMS performance evaluation studies reported in [8, 9] relating to the VITAL network architecture [2] and the analytic works on the entropy maximisation and open GE-type QNMs with arbitrary configuration (c.f., [12, 13]. More specifically, it introduces a mobility management mechanism for an IMS testbed, which was implemented by Nokia-Siemens as part of the deliverables of the VITAL project [2]. It also studies and evaluates the bursty traffic flows of SIP messages during a handover process involving WLAN and GSM access networks. Moreover, it models the functional units and application servers of an IMS architecture with SIP messages as an open (QNM) with finite capacity, generalised exponential (GE) interarrival times, head-of-line priorities and complete buffer sharing management scheme under a repetitive service blocking with random destination (RS-RD) (c.f., [13]). Finally, typical numerical experiments are included to validate the credibility of the analytic ME results against those devised by discrete event simulation and assess the performance impact of traffic flows of SIP messages on the functional IMS units during handover between different IMS access networks.

Note that this work assumes that all user equipments (UEs) are subscribers of IMS and associated cellular networks whilst the terms user, client, DMH and UE are used interchangeably. Moreover, it deals with the first scenario only, i.e., an IMS-IMS call is ongoing. Then the A-Party performs an IMS originating handover IMS→GSM. The result is a GSM-IMS call.

The rest of the paper is organised as follows: Section 2 introduces GE-type distribution whilst Section 3 reviews the maximum entropy (ME) methodology

as applied to the analysis of arbitrary open queueing network models (QNMs) with RS-RD blocking. Section 4 describes the message flows during the handover process. Section 5 devises a custom made open QNM of the functional units and application servers of an IMS network architecture during a handover session. Sections 6 explains the simulation setup used for the validation of the QNM and presents some typical numerical experiments, which validate the credibility of the ME performance metrics against simulation and also assess the performance impact of bursty GE-type traffic flows of SIP messages on the IMS architecture. Conclusions follow Section 7 and a list of acronyms used throughout the manuscript is placed in Appendix I.

2 The GE-Type Distribution

The GE-type distribution is a mixed interevent-time distribution of the form (c.f., Fig. 1)

$$F(t) = P(A \leq t) = 1 - \tau e^{-\sigma t}, \quad t \geq 0, \tag{1}$$

$$\tau = \frac{2}{C^2 + 1} \tag{2}$$

$$\sigma = \tau\nu \tag{3}$$

where A is a mixed-time random variable (rv) of the interevent-time and $(1/\nu, C^2)$ are the mean and squared coefficient of variation (SCV) of rv A (c.f., [12]- [13]). The GE-type distribution is versatile, possessing pseudo-memoryless properties which make the solution of many GE-type queueing systems and networks analytically tractable.

For $C^2 > 1$, the GE is an extremal case of the family of Hyperexponential-2 (H_2) distributions with the same (ν, C^2) having a corresponding counting process equivalent to a compound Poisson process (CPP) with parameter $2\nu/(C^2 + 1)$

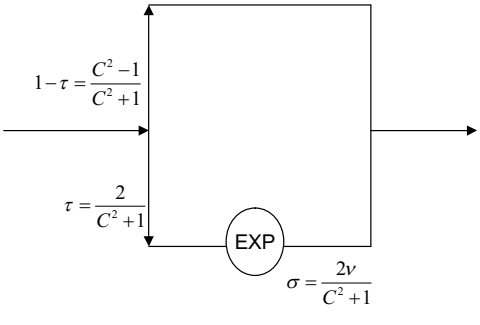

Fig. 1. The GE distribution with parameters τ and $\sigma (0 \leq \tau \leq 1)$

and geometrically distributed bulk sizes with mean, $(1+C^2)/2$ and SCV, $(C^2-1)/(C^2+1)$. The CPP is expressed by

$$P(N_{cp}=n)=\begin{cases}\sum_{i=1}^{n}\frac{\sigma^i}{i!}e^{-\sigma}\binom{n-1}{i-1}\tau^i(1-\tau)^{n-i}, & n\geq 1\\ e^{-\sigma}, & n=0\end{cases} \tag{4}$$

where N_{cp} is a CPP rv of the number of events per unit time corresponding to a stationary GE-type interevent rv.

The choice of the GE distribution is further motivated by the fact that measurements of actual interarrival or service times may be generally limited and so only few parameters can be computed reliably. Typically, only the mean and variance may be relied upon, and thus, a choice of a distribution which implies least bias (i.e., introduction of arbitrary and therefore, false assumptions) is that of GE-type distribution.

3 Entropy Maximization and QNMs RS-RD Blocking: A Review

This section presents a review of an extended product-form approximation and a related queue-by-queue decomposition algorithm, based on the principle of ME, for open QNMs with arbitrary topology and RS-RD blocking [13]).

3.1 A Product Form Approximation

Consider an arbitrary open QNM at equilibrium with M single GE-type server queues, $R\,(R>1)$ distinct HOL priority classes (indexed from 1 to R in descending order of priority), GE-type external inter-arrival times, random routing with class switching, CBS buffer management scheme and RS-RD blocking. Each queueing station $k\,(k=1,2,\ldots,M)$ is assumed to be modelled by a building block GE/GE/1/N$_k$/HOL/CBS queue k with finite capacity N_k $(N_k\geq 1)$.

Notation

For each queue k $(k=1,2,\ldots,M)$ and job class i $(i=1,2,\ldots,R)$, let

λ_{ki}, C^2_{aki} be the mean rate and SCV of the overall actual inter-arrival process of class i jobs at queue k, respectively,

μ_{ki}, C^2_{ski} be the mean rate and SCV of the actual service process of class i jobs at queue k, respectively,

$n_{ki}(0\leq n_{ki}\leq N_k)$ be the number of jobs of class i at queue k waiting and receiving service,

$\mathbf{n}_k=(n_{k1},n_{k2},\ldots n_{km})$ be the aggregate joint state the network,

π_{ki} be the blocking probability that a completer from any queue m, $m\neq k$ of class i is blocked by queue k,

π_{cki} be the blocking probability that a completer of class i will be blocked by a downstream queue,

$\{a_{kio}, a_{kimj}\}$ be the transition probabilities (first order Markov chain) that a class i job transmitted from queue k leaves the network or attempts to join queue m as a class j job, respectively,

$\{\lambda_{0ki}, C^2_{a\,0ki}\}$ be the mean arrival rate and SCV of the actual external inter-arrival process of class i jobs at queue k, respectively,

$\{\lambda_{kimj}, \hat{\lambda}_{kimj}\}$ *and* $\{\hat{C}^2_{a\,kimj}\, C^2_{a\,kimj}\}$ be the mean arrival rates and SCVs of the actual and effective, respectively, inter-arrival processes of class i jobs transmitted from queue k to queue m as class j jobs,

$\{\pi_{kimj}\}$ be the blocking probabilities that a job of class i upon its service completion from queue k will be blocked by queue m, as class j,

$\{\pi_{0ki}\}$ be the blocking probabilities that an external arrival of class i is blocked by queue k.

A credible universal product-form ME approximation of the joint state probability $\{P(\mathbf{n}), \mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2, \ldots \mathbf{n}_M)\}$, subject to normalization and marginal (per queue and class) constraints of server utilization, busy state probability, mean queue length and full buffer state probability (when a class i job is in service at queue k) can be established (c.f., [13]), namely

$$P(\mathbf{n}) = \prod_{k=1}^{M} P_k(\mathbf{n}_k), \tag{5}$$

where $\{P_k(\mathbf{n}_k),\ k = 1, 2, \ldots, M\}$ are the marginal (class) ME state probabilities at queue k.

3.2 ME Queue-by-Queue Decomposition Algorithm

A ME queue-by-queue decomposition algorithm for the approximate analysis of the aforementioned arbitrary open QNMs is summarized in Algorithm 1.

A pictorial presentation of the flow streams and queue-by-queue decomposition can be seen in Fig. 2.

Remarks

- The GE-type distribution is used to approximate the effective inter-arrival and inter-departure time processes for each class $i (i = 1, 2, \ldots, R)$ at each queue $k (k = 1, 2, \ldots, M)$ of the network. The algorithm incorporates a feedback correction of the original service parameters $\{\mu_{ki}, C^2_{ski},\ \forall\ k, i\}$ in order to mitigate the strong underlying assumption that arrival streams per class within the network can be modelled via renewal CPPs. Note that, under the RS-RD blocking mechanism, the ME algorithm utilizes stochastic closed-form expressions for the calculation of the effective service time parameters $\{\hat{\mu}_{ki}, \hat{C}^2_{ski}\}$ (c.f., [13]). Since RS-RD blocking imposes a dependence

Algorithm 1. The ME Decomposition Algorithm

Input Data

- M, R,
- $\{N_k, \lambda_{0ki}, C^2_{a\,0ki}, \mu_{ki}, C^2_{s\,ki}, a_{kimj}\}$ $k = 1, 2, \ldots, M, m = 0, 1, \ldots, M$, $i, j = 1, 2, \ldots, R$.

Begin

Step 1 Feedback correction

Step 2 Initialize π_{0ki} & π_{kimj} to any value in (0,1), $\forall\, k, m = 1, 2, \ldots, M$ and $\forall\, i, j = 1, 2, \ldots, R$; Set $C^2_{dki} = 1, \forall k, i$;

Step 3 Solve the system of the non-linear equations of blocking probabilities $\{\pi_{0ki}, \pi_{kimj}, \forall k, m, i, j\}$;

Step 3.1 For each censored GE/GE/1/N/HOL/CBS queueing station $k, k = 1, \ldots, M$ under RS blocking, calculate the effective flow transition probabilities $\{\hat{a}_{kimj}, \forall k, m, i, j\}$;

Step 3.2 Calculate effective inter-arrival time message flow balance equations for $\{\hat{\lambda}_{0ki}, \hat{\lambda}_{ki}, \forall k, i\}$;

Step 3.3 Calculate the effective service-time parameters, $\{\hat{\mu}_{ki}, \hat{C}^2_{s\,ki}, \forall k, i\}$ under RS-RD blocking mechanism;

Step 3.4 Calculate the overall GE-type inter-arrival-time parameters, $\{\lambda_{ki}, C^2_{a\,ki}, \forall k, i\}$;

Step 3.5 Obtain new values for $\{\pi_{0ki}\, \pi_{mjki}, \forall k, i\}$, by applying Newton Raphson method;

Step 4 Calculate GE-Type inter-departure parameters $\{\lambda_{dki}, C^2_{d\,ki}, \forall k, i\}$;

Step 5 Obtain a new value for the overall inter-arrival-time SCVs, $\{C^2_{a\,ki} \forall\, , k, i\}$;

Step 6 Return to Step 3 until convergence of $\{C^2_{a\,ki},\ \forall\, k, i\}$;

Step 7 Obtain GE-type performance metrics of interest.

End

relationship on the actual routing of jobs from one queueing station to another, it is necessary to create and adopt an effective transition probability matrix $\hat{A} = (\hat{\alpha}_{ki0}, \hat{\alpha}_{kimj})$ in the solution process.

- The ME algorithm describes the computational process of solving the non-linear equations for job loss, $\{\pi_{0ki}\}$ and blocking, $\{\pi_{kimj}\}$, probabilities under GE-type flow formulae [12, 13] for the determination of the first two moments of merging, splitting and departing streams. The main computational cost of the algorithm is of $\mathrm{O}\{\mathrm{kR^2M^2}\}$, where k is the number of iterations in step 3 and $\mathrm{R^2M^2}$ is the number of operations for inverting the associated Jacobian matrix of the system of non-linear equations $\{\pi_{0ki},\ \pi_{kimj},\ \forall\, k, m, i, j\}$ via a quasi-Newton numerical method.

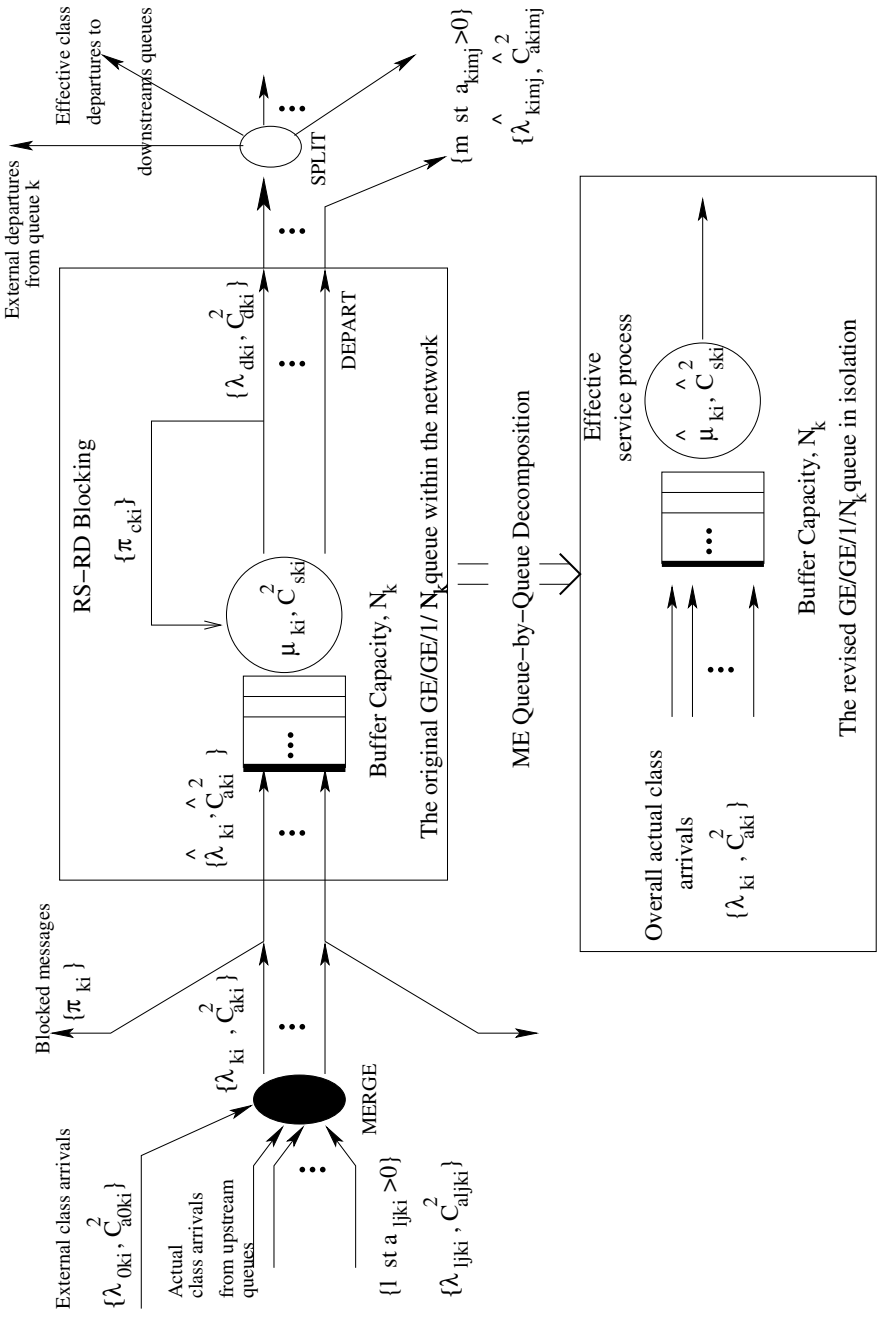

Fig. 2. Flow streams and ME queue-by-queue decomposition

4 Message Flows

The message flows considered in this paper consist of voice calls with the following profile: Whilst an IMS-IMS voice call is ongoing between GSM/IMS subscribers with DMH, the UE-A performs an IMS originating handover from WLAN/IMS to GSM/IMS (c.f., [2]). The outcome of these operations is a GSM-IMS call.

Before the handover occurs, a call is established from UE-A to UE-B. This is illustrated in numbers 1 to 6 of Fig. 3 The handover is initiated by the DMH roaming (c.f., number 7 in Fig. 3). This may be based on measurements of the WLAN signal strength in the DMH that triggers the handover procedure when the signal becomes weaker or abruptly disappears. The handover procedure may also be triggered manually (e.g., when pressing a button on the DMH).

The handover is performed by the following steps: The UE-A initiates a voice call via the GSM network. The call setup request is addressed to UE-A (the mobile basically makes a call to itself). The originating mobile switching centre (MSC) of the visited GSM network forwards the call Prefix-Routing based on the Calling-Party (A-Party) address to the media gateway control function (MGCF). When the MMAS receives the call setup request (i.e., the SIP Invite with Roaming Number Originating (RN-O) in the SIP Request-URI) in number 12 (c.f., Fig. 3), then a handover is required. This is because the MMAS already has a SIP call in IMS ongoing from UE-A and thus, it is aware that UE-A wants to roam from WLAN to GSM (while the call is ongoing). To actually perform the handover, the MMAS (acting as B2BUA) sends a SIP Re-Invite message to UE-B (n.b. this Re-Invite message is not shown in Fig. 3). The Re-Invite traverses along the SIP signalling path of the existing call and carries the session description protocol (SDP) with the IP address and port information used for the voice bearer of the call in the media gateway (MGW). The MMAS obtained this SDP from the MGCF with the SIP Invite message (c.f., numbers 10-12 of Fig. 3). When the SIP Re-Invite arrives at the UE-B, then the UE-B redirects the voice bearer traffic from UE-B to the MGW and answers the request with a 200-OK, which carries the SDP of UE-B. At the end of the procedure, the voice bearer is exchanged between the MGW and the UE-B. Finally, the MMAS closes the WLAN/IMS based call leg between the UE-A and the MMAS by sending a SIP Bye message via the S-CSCF to the UE-A. This completes the handover procedure. The new call now follows the signalling path indicated in Fig. 3.

The MMAS of the UE-A, the Voice AS and the MMAS of UE-B remain within the SIP signalling path after the handover. This is in order to i) control another handover of UE-A ii) control a handover of UE-B and iii) provide Supplementary Services (SS) based on VAS.

Note that the aforementioned handover mechanism is only applicable to a single session. Therefore, the DMH cannot have several voice calls active at the same time, which may then get successfully handed over from WLAN/IMS to

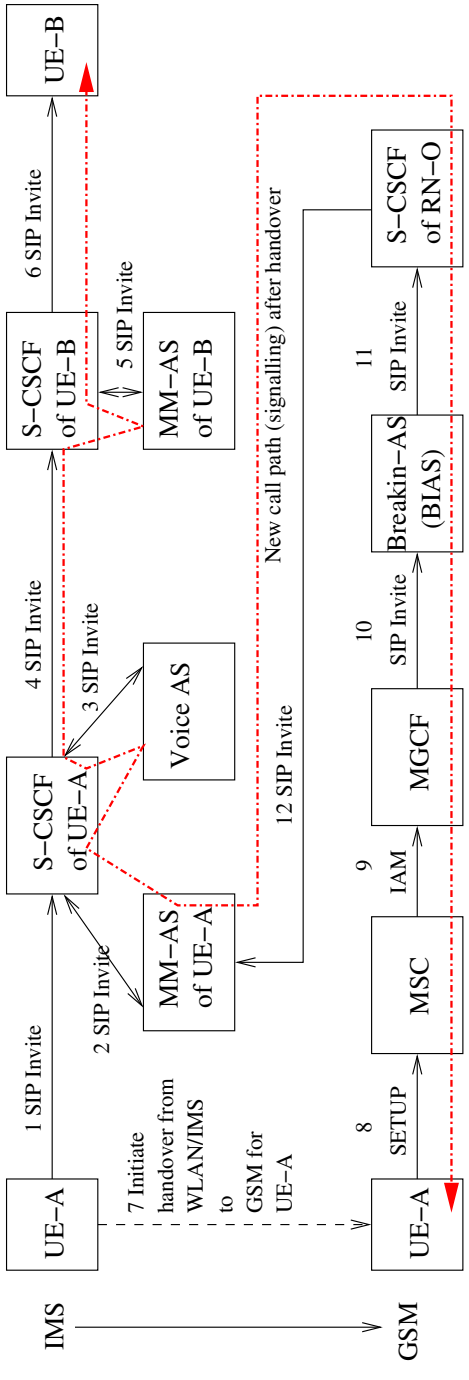

Fig. 3. An IMS originating handover from WLAN/IMS to GSM/IMS

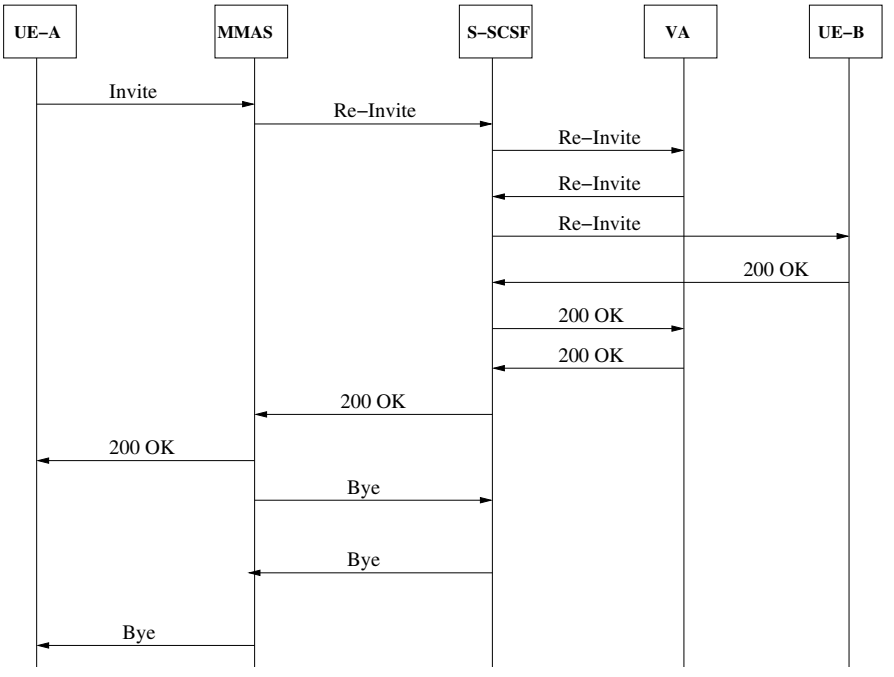

Fig. 4. Flows of SIP messages during WLAN/IMS to GSM/IMS handover mechanism

GSM/GSM and vice versa. This limitation was made to simplify the logic in the MMAS and DMH.

Fig. 4 depicts the flow chart of SIP messages during a handover mechanism from WLAN/IMS to GSM/IMS.

5 An Open QNM of the IMS Functional Units and Application Servers

A diagrammatic illustration of an open QNM of the IMS functional units (P-CSCF,S-CSCF and I-CSCF) and application servers (VAS and MMAS) during a handover process is displayed in Fig.5. There is an additional application server for future use that represents an IP Television (IPTV).

Note that $\hat{\mu}_{ki}$ and $\hat{C}^2_{ski}, \forall ki$ are the effective service time parameters, λ_{ki} and C^2_{aki} are overall actual class arrivals. In the context of this paper, $R = 2$, i.e., two distinct HOL priority classes are considered. The high priority is given to SIP Re-invite messages that trigger the handover process. The rest of the SIP messages and other traffic flows are given the low priority. Note that Real-Time Transport Protocol (RTP) traffic does not flow and pass through the IMS core network.

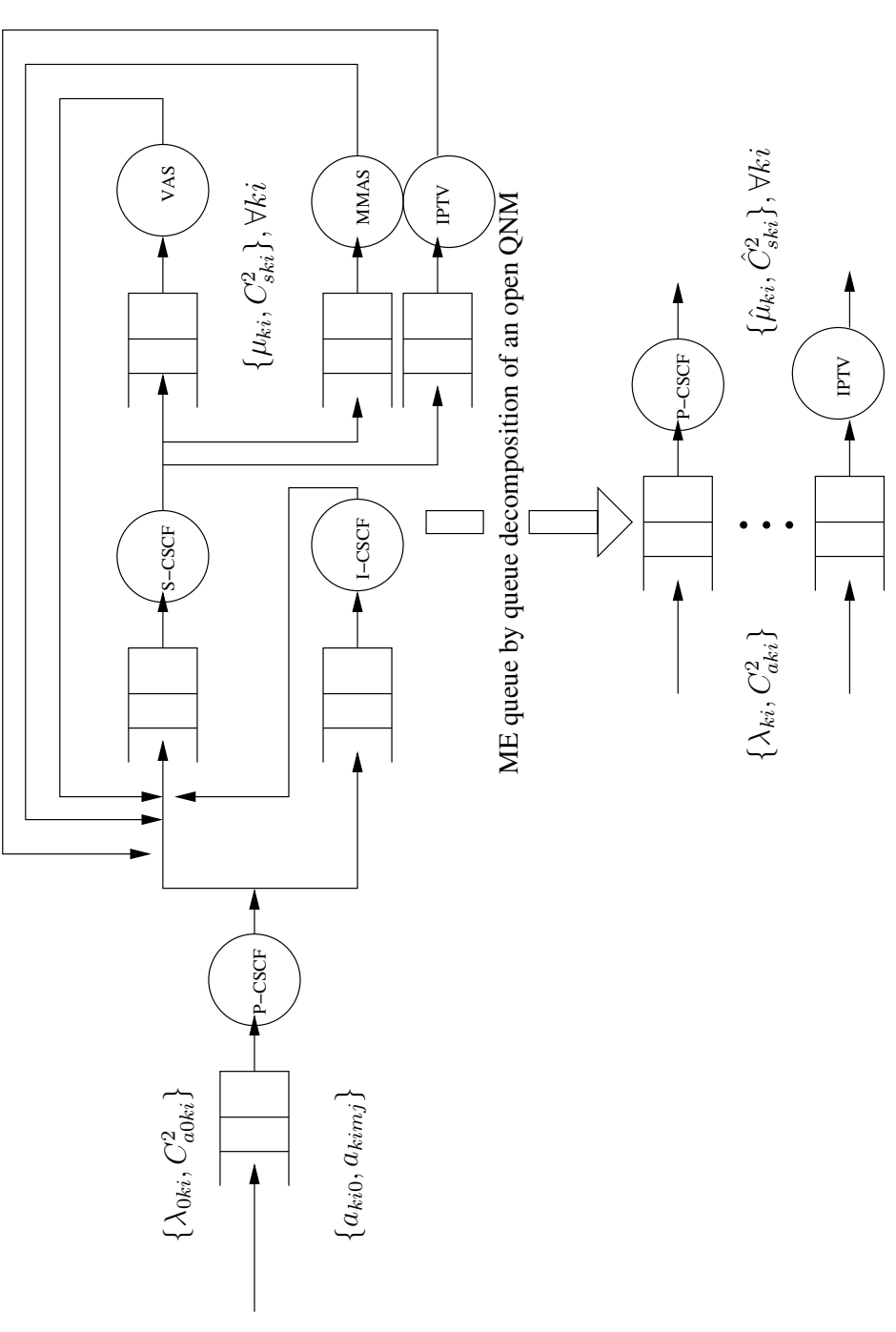

Fig. 5. An open QNM for IMS Functional Units and Application Servers

6 Simulation Setup and Numerical Results

This section presents numerical validation results based on the ME analysis and simulation of the proposed open QNM in Fig. 5. The input data used in the simulation setup in order to validate the open QNM for the IMS handover management mechanism against simulation are presented in Table 1 at 95% confidence intervals.

Table 1. Numerical Inputs and Results for the Open QNM of Fig. 5

Number of Classes=2					
$\lambda_{011} = 0.7$	$\lambda_{012} = 0.5$	$C^2_{a011} = 5.0$	$C^2_{a012} = 5.0$		
Buffer capacity at each link=5					
$\mu_{11} = 2$	$C^2_{s11} = 3$	$\mu_{12} = 3$	$C^2_{s12} = 5$	$\mu_{21} = 1$	$C^2_{s21} = 2$
$\mu_{22} = 3$	$C^2_{s22} = 4$	$\mu_{31} = 2$	$C^2_{s31} = 5$	$\mu_{32} = 3$	$C^2_{s32} = 4$
$\mu_{51} = 1$	$C^2_{s51} = 4$	$\mu_{52} = 2$	$C^2_{s52} = 5$	$\mu_{41} = 2$	$C^2_{s41} = 4$
$\mu_{42} = 1$	$C^2_{s42} = 2$	$\mu_{61} = 3$	$C^2_{s61} = 4$	$\mu_{62} = 1$	$C^2_{s62} = 5$
Transition Matrix					
$a_{2151} = 0.5$	$a_{2252} = 0.5$	$a_{2141} = 0.5$	$a_{2242} = 0.5$	$a_{1121} = 0.5$	$a_{1222} = 0.5$
$a_{3222} = 0.5$	$a_{3121} = 0.5$	$a_{4222} = 0.5$	$a_{4121} = 0.5$	$a_{1131} = 0.5$	$a_{1232} = 0.5$
$a_{2151} = 0.5$	$a_{2252} = 0.5$	$a_{5222} = 0.5$	$a_{5121} = 0.5$	$a_{6212} = 0.5$	$a_{6111} = 0.5$

The simulation was implemented using for communication purposes the Java Remote Method Invocation (RMI) package. The main program was developed using Java SDK 6 whilst the S-CSCF, P-CSCF, I-CSCF, MMAS, VAS and an extra application server for IPTV were simulated using Linux 2.6.9-42.0.2.ELsmpi686 Athlon i386 machine with 4GB of RAM. Other background traffic flows were introduced in the network consisting of file transfer protocols (FTPs) and hyper text transfer protocol (HTTP). It was assumed that 4 UE-Bs and 10 UE-As were available.

Without loss of generality, the comparative study is based on marginal performance metrics of mean response time and aggregate mean queue length of SIP Re-Invite handover messages by varying the SCV of the inter-arrival times. It can be seen from Fig. 6 that the mean response time of SIP messages during a handover process begins to deteriorate for increasing values of inter-arrival time SCV.

Fig. 7 illustrates the effect of varying inter-arrival SCV on the aggregate mean queue length of SIP messages. In Fig. 8, each functional unit of the IMS is compared to each other in terms of the corresponding mean response time of SIP Re-Invite messages against the the SCV of the inter-arrival traffic. From the analytical and simulation results, it can be seen that S-CSCF provides the most pessimistic mean response time.

It can be observed in Figs. 6-8 that the analytic ME results compare favourably with those of the simulation. Moreover, the variability of the SCV of the inter-arrival times has an adverse impact on the performance of the IMS functional units. Typically, the S-CSCF is the bottleneck in the IMS core network because it processes a large number of SIP messages.

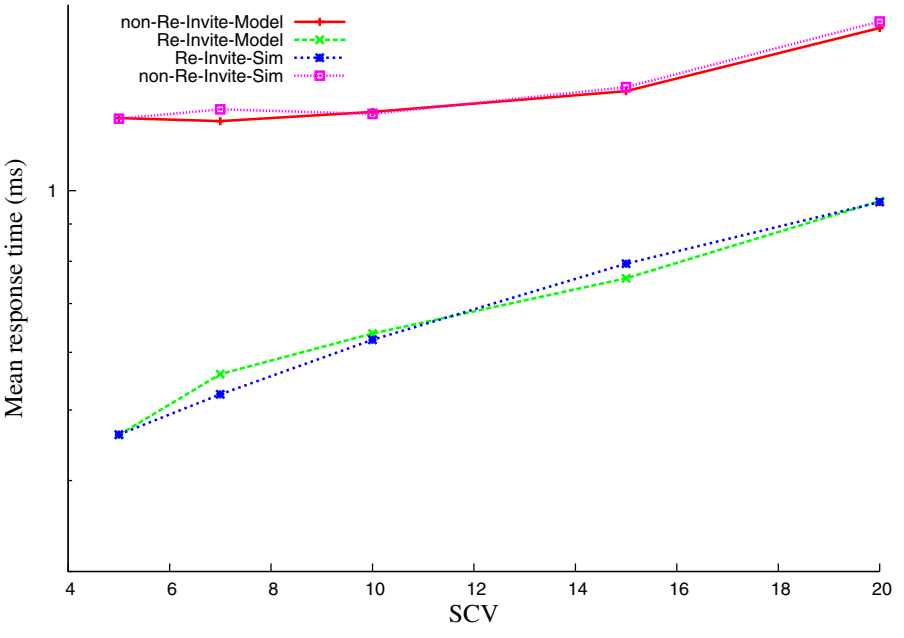

Fig. 6. The effect of traffic variability on the mean response time of SIP messages

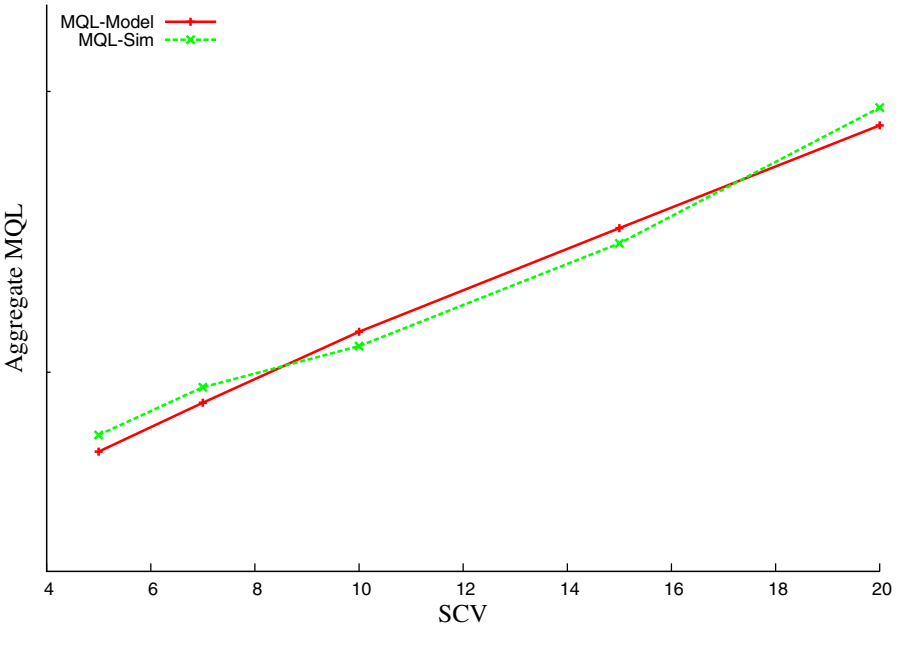

Fig. 7. The effect of traffic variability on the aggregate mean queue length

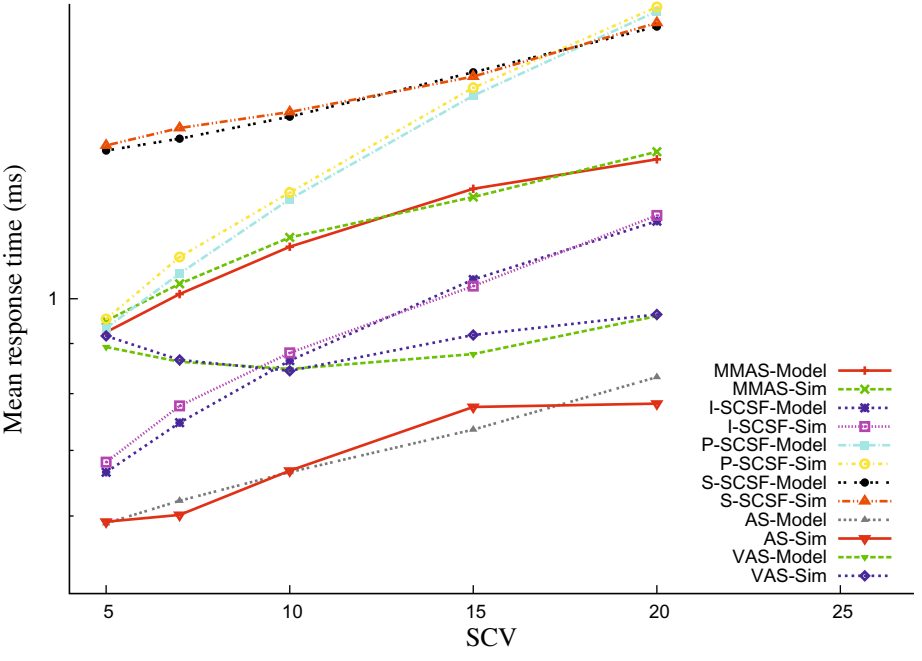

Fig. 8. The mean response time at each IMS functional unit

7 Conclusions

A mobility management mechanism was introduced for an IP multimedia subsystem (IMS) testbed implemented by Nokia-Siemens as part of the EU IST VITAL project [2]. The GE-type distribution [12] was used to characterise the bursty traffic flows of SIP messages during the handover process between the WLAN and GSM access networks. The functional units and application servers of the IMS architecture were presented in the context of an open QNM with HoL priority classes for the SIP messages and complete buffer sharing under RS-RD [13].

The quantitative analysis of the open QNM was based on the universal ME product-form approximation and related queue-by-queue decomposition algorithm as well as discrete event simulation. Typical numerical experiments showed that the analytic ME solutions were very comparable to the corresponding simulation results. Moreover, the performance prediction study revealed that the variability of bursty traffic described by the interarrival time SCV had an adverse impact on the performance of the IMS functional units. In particular, the S-CSCF proxy was identified as the bottleneck device of the IMS core network architecture processing a large number of SIP messages during handover.

References

1. 3GPP: Technical Specification Group Services and System Aspects, IP Multimedia Subsystem (IMS), Stage 2, TS 23.228, 3rd Generation Partnership Project (2006), Website, http://www.3gpp.org/specs/specs.htm
2. VITAL: Enabling Convergence of IP Multimedia Services Over Next Generation Networks Technology (2006), Website, http://www.ist-vital.eu
3. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E.: SIP: Session Initiation Protocol. RFC 3261 (2002)
4. Cortes, M., Ensor, R., Esteban, J.: On SIP Performance. Bell Labs Technical Journal 9(3), 155–172 (2004)
5. Batterman, H., Meeuwissen, E., Bemmel, J.: SIP Message Prioritization and its Application. Bell Labs Technical Journal 11(1), 21–36 (2006)
6. Zou, J., Xue, W., Liang, Z., Zhao, Y., Yang, B., Shao, L.: SIP Parsing Offload. In: Global Telecomminications Conference (Globecom), pp. 2774–2779 (2007)
7. Rajagopal, N., Devetsikiotis, M.: Modeling and Optimization for the Design of IMS Networks. In: 39th Annual Simulation Symposium, pp. 34–41 (2006)
8. Mkwawa, I.M., Kouvatsos, D.D.: Performance Modelling and Evaluation of IP Mmultimedia Subsystems. In: HET-NETs 2008 International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks, pp. B07.1–B07.7 (2008)
9. Mkwawa, I.M., Kouvatsos, D.D.: Performance Modelling and Evaluation of Handover Mechanism in IP Multimedia Subsystems. In: ICSNC 2008: Proceedings of the 2008 Third International Conference on Systems and Networks Communications, pp. 223–228. IEEE Computer Society Press, Washington, DC, USA (2008)
10. Baskett, F.: Mathematical Models of Multiprogrammed Computer Systems. TSN-17, Computer Centre, The University of Texas, Austin, Texas (1971)
11. Forte, G., Schulzrinne, H.: Template-Based Signaling Compression for Push-To-Talk over Cellular (PoC). In: Schulzrinne, H., State, R., Niccolini, S. (eds.) IPTComm 2008. LNCS, vol. 5310, pp. 296–321. Springer, Heidelberg (2008)
12. Kouvatsos, D.D.: Entropy Maximization and Queueing Network Models. Annals of Operation Research 48, 63–126 (1994)
13. Kouvatsos, D.D., Awan, I.: Entropy Maximisation and Open Queueing Networks with Priorities and Blocking. Perform. Eval. 51(2-4), 191–227 (2003)

Appendix I Table of Acronyms

3GPP	3rd Generation Partnership Project (3GPP)
AS	Application Server
B2BUA	Back to Back User Agent
DMH	Dual Mode Handset
GE	Generalised Exponential
GSM	Global System for Mobile communications
HoL	Head of Line
I-CSCF	Interrogating Call Session Control Function
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPTV	IP Television
ME	Maximum Entropy
MGW	Media Gateway
MMAS	Mobility Management Application Server
MGCF	Media Gateway Control Function
MSC	Mobile Switching Center
P-CSCF	Proxy Call Session Control Function
PoC	Push to talk over Cellular
QNM	Queueing Network Model
RN-O	Roaming Number-Originating
RN-T	Roaming Number-Terminating
RS-RD	Repetitive Service blocking with Random Destination
RTP	Real-time Transport Protocol
S-CSCF	Serving Call Session Control Function
SCV	Squared Coefficient of Variation
SDP	Session Description Protocol
SIP	Session Initiation Protocol
SS	Supplementary Services
UE	User Equipment
URI	Uniform Resource Identifier
VAS	Voice Application Server
VCC	Voice Continuity Call
WLAN	Wireless Local Area Network