# A Traffic Model for the IP Multimedia Subsystem (IMS)

V.S. Abhayawardhana*, R. Babbage†
*BT Mobility Research Unit, pp B28/2B, Adastral Park, Ipswich IP5 3RE, UK.
viraj.abhayawardhana@bt.com
†Network Performance Group, pp MLB4/1A, Adastral Park, Ipswich IP 3RE, UK.
ruth.babbage@bt.com

***Abstract*— The IP Multimedia Subsystem (IMS) could very well be the panacea for most telecom operators. Defined originally as the core network for 3G mobile systems by the 3rd Generation Partnership Project (3GPP), the more recent releases have included interfaces to fixed line networks and Wireless LANs. British Telecom is embarking on a 10 year long bold endeavor, called the 21st Century Network (21CN), to completely overhaul its core network to one that is based on 3GPP IMS. The ultimate goals are to reduce operational cost and provide converged services to its customers. At the heart of the IMS is the Home Subscriber Server (HSS), the master database that holds all customer profiles. The two main protocols used for session control procedures are the Session Initiation Protocol (SIP) and Diameter. Both are sent in clear text and very heavy weight. Although IMS promises an exciting world of converged services, the sheer amount of signaling traffic could prove to be too costly. Since there are no known large scale IMS networks, a representative signaling traffic model is still unavailable. We, at BT, have defined a signaling traffic model for IMS using the experience we gained through 21CN. The model quantifies the traffic and latency for various procedures defined in IMS, starting from the basic call flows. We present the model in this paper and also compare the IMS traffic with other traditional schemes and make conclusions on its efficiency.**

## I. Introduction

The IP multimedia Subsystem (IMS) was first defined by the 3rd Generation Partnership Project (3GPP) in Release 5 as the core network architecture for the 3G cellular system. It's an open-systems architecture that supports a range of IP-based services over both PS and CS networks. It enables peer-to-peer real time services, such as voice and video. It has a common session control layer based on Session Initiation Protocol (SIP) [1], which gives the ability to manage parallel user services and mix different multimedia in a single or parallel sessions. It is also access independent, hence subsequent releases of the 3GPP standards have seen it opened to Wireless LANs (R6) and Fixed networks (R7). This will pave the way for Fixed-Mobile Convergence (FMC).

British Telecom (BT) has identified the importance of IMS and has taken the radical step of embarking on a 10-year plan worth £10 billion to completely overhaul the core network to one based on the IMS model. Having modified the IMS model to particularly suit BT's requirements, the BT model is called the 21st Century Network (21CN).

At the heart of the IMS model is the Home Subscriber Server (HSS), the master database which holds both the authentication and service user profiles. It is the 'brain' of the network, the high performance of which is critical to the whole network. The HSS uses the Diameter protocol [2] for Authentication, Authorisation and Accounting (AAA) dialogues with different components in the network, mostly with Call Session Control Functions (CSCFs) during session establishment. Diameter, like SIP, is sent in clear text and can be heavy. It is generally understood that the weight of signalling protocols is not an issue because most of the links are within the operator's own IMS network. However a quantification of traffic is required for various reasons, but mainly to accurately dimension the entire system. Although the links that use SIP are mostly within an operators internal network, for an operator that relies on FMC some links from the access networks may be beyond its control, for example from public hotspots. Hence it is essential that the signaling traffic in an IMS network is analysed and quantified in a form of a representative traffic model. Since there are no known large scale IMS networks, a representative signaling traffic model is still unavailable. The discussion in this paper will be mainly limited to the traffic created to and from the HSS, the main element of the IMS.

To the best of the authors' knowledge, there are no known traffic models in the public domain that describes the signaling in IMS. Several authors have written about use of SIP on IMS [3], but they are limited to issues such as interactions with Mobile IP [4] and how it supports real-time multimedia [5], but they all stop short of analysing and quantifying the IMS signaling traffic. This paper is divided as follows; section II discusses the IMS signaling call flows, section III introduces our model, section IV critically compares IMS traffic to other schemes and finally in section V conclusions are made and points to where the research can be directed in the future.

## II. IMS Call Flows

The SIP signaling in an IMS *always* flows through the home network of each party. Therefore if both parties are roaming, the SIP messages could possibly flow through 4 networks. This is to ensure that proper services are triggered at all times. However, the media traffic in IMS flows end-to-end between the two parties. The IMS network has numerous functional elements in it, but figure 1 shows a very simplified view of its session control layer. This paper will concentrate only on

1550-2252/$25.00 ©2007 IEEE

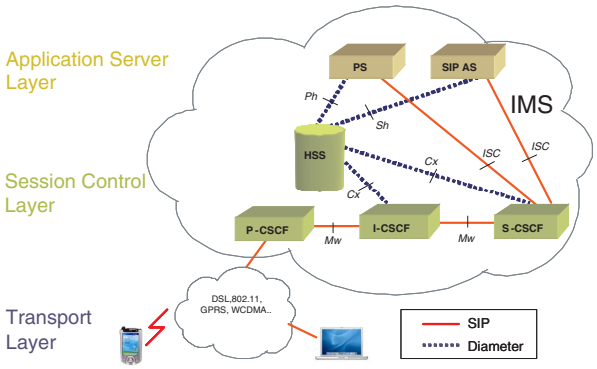


Fig. 1. A simplified view of the session control in IMS

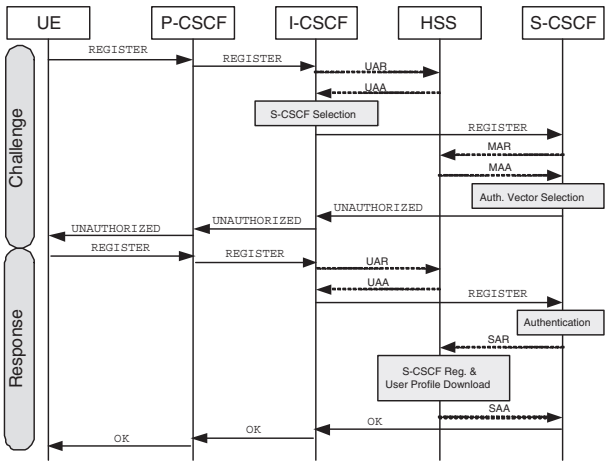


Fig. 2. IMS Registration call flows (when user not registered)

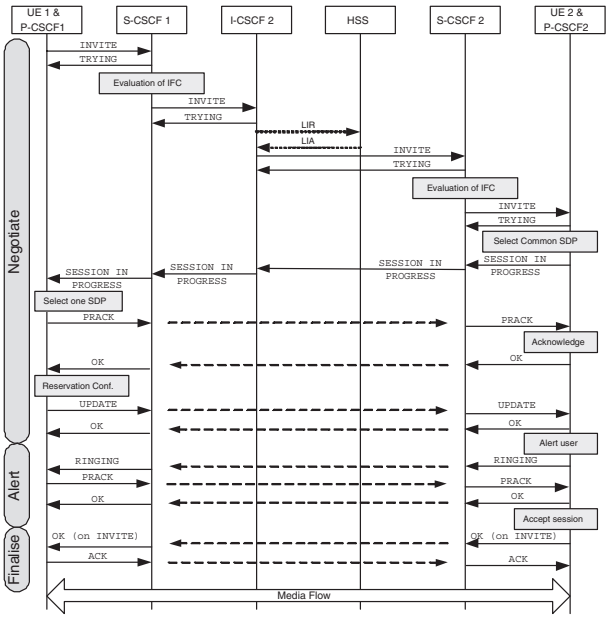


Fig. 3. IMS Session setup call flows (both users in the same network)

those elements shown. The CSCFs come in three kinds. Firstly the Proxy-CSCF (P-CSCF), which may reside in a roaming network, is the first point of contact for the User Equipment (UE). It forwards SIP to and from the home network and may also perform encryption and compression. The Interrogating-CSCF (I-CSCF) is the entry point to the home network. It may function similar to a firewall and hide the internal topology. Lastly, the Serving-CSCF (S-CSCF) is the main element in session control. It is fully responsible for registration and controlling of sessions to the UE. It also decides which Application Servers (AS) that need to be triggered, depending on the Initial Filter criteria (IFC). The IFC is part of the user profile which is held in the HSS and downloaded to the S-CSCF upon registration. One other important element is the Presence Server (PS), that holds the presence status of each subscriber and a list of 'watchers' that are interested in that information.

The procedures that need to be carried out in IMS are clearly defined in [6], [7]. Since the paper concentrates on the traffic created towards the HSS the discussion will be mainly limited to 3 procedures in this paper, namely registration, call setup and presence watcher subscriptions. The basic registration call flow (i.e. when user registers for the first time) is shown in figure 2. A SIP `REGISTER` message is used to initiate a registration. Once it flows to the I-CSCF, it does a Diameter UAR/UAA [8] dialogue with HSS to download a list of S-CSCFs. Once it flows to the S-CSCF, it does an MAR/MAA lookup to download authentication vectors from the HSS. The credentials required for the authentication challenge are sent in a SIP `UNAUTHORIZED` message. The response to the challenge is sent by the UE in another `REGISTER` message. If the authentication is successful, the S-CSCF does a Diameter SAR dialogue to register its name in the HSS and subsequently downloads the user profile in a SAA message. Call flows attributed to registrations in other scenarios as well as de-registrations are presented in [6].

Figure 3 shows the call flows for a typical IMS session set up. The UE and the P-CSCF columns have been merged for brevity. The session setup goes through three phases, negotiation, alert, and finalisation. There is only one LIR/LIA interaction with the HSS during negotiation to locate the S-CSCF address assigned to the UE at the terminating end. A SIP message carries an Session Description Protocol (SDP) part in the body of the message that describes the media available at either end. The first `INVITE`, `SESSION IN PROGRESS` and `PRACK` messages are used by the two UEs to negotiate the media that will be used to establish the session. Please refer [6] for more details. Please note that some intermediate flows are not shown after the `SESSION IN PROGRESS` message in figure 3 and are denoted by dashed arrows. They should flow similar to the detailed ones shown on top of the figure, but without any HSS interactions.

Presence is one of the most important services that could be provided by the IMS and could easily be reused by other services. The PS holds presence information of 'presentities', which 'watchers' subscribe to. Any changes of presence state

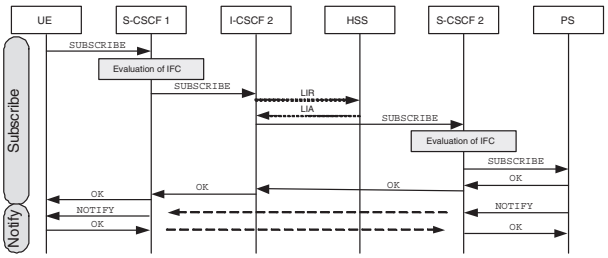


Fig. 4. Presence call flows (New watcher subscription)

of presentities are informed to the watchers through SIP NOTIFY messages. Call flows for presence service in IMS are found in [9], but the discussion will be limited to the scenario of a watcher subscribing to a new presentity as shown in figure 4. The presence service itself is bound to create a lot of SIP traffic but this is the only scenario that will generate traffic towards the HSS. Similar to the call set up scenario, a LIR/LIA is required to find out the S-CSCF assigned to the presentity. If an LIR/LIA message is generated by the presence service, the interface between S-CSCF and the HSS is denoted as 'Px', rather than 'Cx'.

Apart from the call flows and the interfaces discussed so far, one interface that is worth noting is the Sh, between the HSS and the ASs. It uses the Diameter protocol. The ASs can download the user profile over the interface or subscribe for notification of any changes to the user profile. The call flows for those will not be discussed in more detail as they are relatively straight-forward and can be found in [10].

## III. IMS Traffic Model

To produce generalised calculations for the IMS, access network delays, such as PDP context activation for GPRS networks, have been omitted. A typical IMS subscriber can be of various types, such as residential, enterprise, WiFi user etc. and each will have a different profile. For purposes of simplicity, we have made the assumptions shown in table I that we think is representative of all users. Some of the values quoted are for Busy Hour (BH).

### A. Traffic flow calculations

The assumptions that were presented in table I were used to calculate the number of call flow instances/sec in BH, assuming an even spread of traffic in it. Consequently, the call flow instances/sec in BH was calculated as shown in table II. A sensitivity analysis is also presented for 2 and 4 million (M) subscribers. These call flows were chosen because they were considered likely to provide the bulk of the traffic to and from the HSS.

The first column in table II also shows the calculations that gives the *proportion* for each entry. Here, $t = 3600$, which is the number of seconds in an hour. Please note that although presence notification generates a large volume of traffic, it does not create any traffic towards the HSS. The only use case that will be discussed hereafter within the presence service is new watcher subscriptions, which causes a LIR/LIA lookup from the HSS.

| Description | Amount |
|---|---|
| Perc. of users registered at BH, $a$ | 80% |
| Perc. of users that are permanently registered, $b$ | 60% |
| Perc. of users that register everyday, $c$ | 20% |
| No. of times a user registers per day, $d$ | 2 |
| Re-registrations per registered user per hour, $e$ | 1 |
| Perc. of registrations in BH, $f$ | 25% |
| Originating sessions/reg. user in BH, $g$ | 1.5 |
| Overall terminating session/user in BH, $h$ | 1.5 |
| Perc. sessions terminated while user not registered, $i$ | 10% |
| Perc. of UE-initiated de-registrations, $j$ | 95% |
| Perc. of NW-initiated de-registrations, $k$ | 5% |
| AS contacts/user in BH, $l$ | 0.125 |
| AS subs. to notification/user in BH, $m$ | 0.0625 |
| Perc. of users that use presence, $n$ | 20% |
| No. of presence watchers per user, $o$ | 10 |
| No. of new watcher subscriptions/reg. user in BH, $p$ | 0.25 |
| No. of presence status changes/reg. user in BH, $q$ | 1 |

TABLE I
ASSUMPTIONS MADE FOR THE IMS TRAFFIC MODEL

| Call flow instances/sec in BH | 2M subs | 4M subs |
|---|---|---|
| Registrations $(c.d.f/t)$ | 56 | 111 |
| Re-registrations $(a.e/t)$ | 444 | 889 |
| Terminating calls - user registered $(h.(1-i)/t)$ | 750 | 1500 |
| Terminating calls - user unregistered $(h.i/t)$ | 83 | 167 |
| De-registrations - UE initiated $(c.d.f.j/t)$ | 53 | 106 |
| De-registration - NW ini.(timeout) $(c.d.f.k/t)$ | 3 | 6 |
| AS user profile download $(l/t)$ | 69 | 139 |
| Subs. to notification of profile change $(m/t)$ | 35 | 69 |
| User presence registration $(c.d.f.n/t)$ | 11 | 22 |
| User presence de-registration $(c.d.f.n/t)$ | 11 | 22 |
| Presence notify $(o.n(q.a + (g + h.(1-i)).2)/t)$ | 7222 | 14444 |
| Presence new watcher subscriptions $(a.n.p/t)$ | 22 | 44 |

TABLE II
MAJOR TRAFFIC FLOWS

The analysis of call flows in section II yields the Diameter messages/interface/use case as shown in figure III. It shows Diameter messages for 3 interfaces, namely the Cx, Sh and the Px. Note, only the request message is shown in the table for brevity (e.g. UAR), but it is implicit that each Diameter message contains a request-answer dialogue (e.g. UAR/UAA).

Using messages/interface/use case information in table III and the major traffic flows for use cases shown in table II, the messages/interface was calculated as shown in table IV. For instance, the number of call flow instances per second for Cx UAR/UAA was calculated as (Re-registrations + UE-initiated de-registrations + 2 x Registrations). All values shown have been rounded to the nearest 10.

SIP and Diameter are IP based protocols and hence can travel in any transport protocol. Most common implementations use UDP for SIP transportation because of simplicity and speed.

| Use case | Cx | | | | | Sh | | Px |
|---|---|---|---|---|---|---|---|---|
| | UAR | MAR | SAR | LIR | RTR | UDR | SNR | LIR |
| Registration | 2 | 1 | 1 | | | | | |
| Re-registration | 1 | | | | | | | |
| Session termination-user registered | | | | 1 | | | | |
| Session termination-user unregistered | | | 1 | 1 | | | | |
| De-registration-UE initiated | 1 | | 1 | | | | | |
| De-registration-NW ini.(timeout) | | | 1 | | | | | |
| De-registration-admin initiated | | | | | 1 | | | |
| User profile download | | | | | | 1 | | |
| Subscribe to notification | | | | | | | 1 | |
| New watcher subs. | | | | | | | | 1 |

TABLE III
DIAMETER MESSAGES PER USE CASE

| Interface | Message type/sec in BH | 2M subs | 4M subs |
|---|---|---|---|
| Cx | UAR/UAA | 600 | 1200 |
| Cx | MAR/MAA | 60 | 110 |
| Cx | SAR/SAA | 190 | 390 |
| Cx | LIR/LIA | 830 | 1700 |
| Sh | UDR/UDA | 70 | 140 |
| Sh | SNR/SNA | 40 | 70 |
| Px | LIR/LIA | 20 | 40 |

TABLE IV
MESSAGES PER INTERFACE

### B. Latency calculations

The second part of the model calculates the latency of the procedures. IMS session setup is of particular interest since it is critical in terms of user experience (i.e. until the `RINGING` message). Both users were assumed to be roaming to find the scenario that would have the longest possible call flow for session set up. Please refer to [6] for the complete call flows. The figures presented in section II do not assume any roaming scenarios.

Table V lists the typical SIP message sizes taken from a lightly loaded test network at the UE to P-CSCF interface. Processing times at each node for each message type were also measured. It should be noted that due to the light load, these processing times are a 'best case'. Since the test measurements did not cover all the message types that appear in the three call flows being considered, the processing times used for the latency calculations are merely approximate best guesses based on the available measurements. It was noted that the processing times for each message at each node were much greater than the transmission times (hop latencies). Note that all links were assumed to be wired. Hence, it was decided to assign a hop latency based on link type, rather than precisely calculating it based on message lengths and transmission link speeds.

Each link was assigned to be either a LAN or a WAN,

| SIP Message | size (bytes) |
|---|---|
| `INVITE` (no authentication) | 930 |
| `INVITE` (digest authentication) | 1280 |
| `SESSION INPROGRESS` | 910 |
| `PRACK` | 450 |
| `OK` | 990 |
| `RINGING` | 450 |
| `ACK` | 630 |
| `BYE` | 510 |
| `REGISTER` (no authentication) | 490 |
| `REGISTER` (digest authentication) | 810 |
| `UNAUTHORIZED` | 680 |
| `SUBSCRIBE` (digest authentication) | 900 |
| `NOTIFY` | 550 |

TABLE V
SIP MESSAGE SIZES

with the former having a typical hop latency of 1 ms and the latter 10 ms. Only a P-SCSF and I-CSCF lying in the same network were assumed to be connected via a LAN and all other elements were assumed be connected via WAN. Also the link between the UE and the P-CSCF was assumed to be via a WAN because the latency calculations were carried out for Digital Subscriber Line (DSL) customers. To estimate the end-to-end latency for these three call flows a summation was made of all the processing times at destination taken for each message on each link. These were based on the measurements on processing times taken at the live network. Table VI shows the final latency values for the three procedures that were considered in the paper. The first column lists the summation of the hop latencies for all of the messages in the call flow. The second column is a summation of the destination processing times for each scenario. Finally, the third column is the summation of the first two columns showing the overall latency. Table VI assumes a HSS latency of 50 ms and table VII shows a sensitivity analysis with the HSS processing time at 100 ms.

| Procedure | Total hop (s) | Total destination (s) | Overall total (s) |
|---|---|---|---|
| Cx Registration | 0.164 | 0.435 | 0.599 |
| Cx Call setup | 0.434 | 0.762 | 1.196 |
| Px subs. to watcher | 0.122 | 0.255 | 0.377 |

TABLE VI
LATENCY FOR EACH PROCEDURE (HSS LATENCY IS 50 MS)

| Procedure | Total hop (s) | Total destination (s) | Overall total (s) |
|---|---|---|---|
| Cx Registration | 0.164 | 0.635 | 0.799 |
| Cx Call setup | 0.434 | 0.812 | 1.246 |
| Px subs. to watcher | 0.122 | 0.305 | 0.427 |

TABLE VII
LATENCY FOR EACH PROCEDURE (HSS LATENCY IS 100 MS)

## IV. Comparison of Signalling Protocols

It has been very difficult to compare the results shown in section III with other systems. In [11], the time taken to establish a SIP call over GSM is calculated to be approximately 4.13 seconds. However, virtually all of that time is taken up for transmission of messages over the GSM radio access network. The authors calculated the transmission times precisely, using message sizes and the speed used to transmit data over the air interface, which was 9.6 kb/s. Since the authors were mainly interested in latencies that arise in radio networks, they assigned a constant round trip time for each message, which was 70 milliseconds. The same values cannot be used for comparison with our model because this paper does not assume any wireless links. Besides an IMS call setup involves more messages being transmitted compared to a simple SIP session set up and typically and IMS network will include more intermediate elements making the round trip time higher. However, if crude yet conservative estimates are made using the method adopted in [11] while associating a constant round trip time of 70 milliseconds for the call flows used in latency calculations of section III the overall total time for Cx registration, Cx call set up and Px subscribe to watcher takes 0.419, 0.552 and 0.175 seconds, respectively. Note that HSS latency was set as 50 milliseconds (i.e. it is added for each HSS access) and for the presence watcher registration a penalty of 15 milliseconds is included for the processing in the presence server. Main reason for the discrepancies between these values and the values presented in table VI is because of the latter calculation takes into account the individual processing times for each message at each node.

In [12] latency estimations are made for a typical SIP call over GERAN network. Here the overall time taken is estimated as 7.9 seconds. However, as in [11], the authors were mainly interested in the latency over the wireless interface. Again comparisons cannot be directly made with the calculations presented in this paper.

## V. Conclusions and Future Work

The purpose of this modelling was to create a traffic model for the HSS and to gain an appreciation of the performance in terms of processing times that the HSS would need to be capable of to ensure that the call flows could be kept within latency targets. From the results it was clear to see that in fact the performance of the HSS would have to be very poor to have a big impact on the end-to-end latency. However, the performance of the S-CSCF, I-CSCF and P-CSCF is crucial, since all of these are visited a large number of times during the call flow.

The work presented in this paper had to be limited mainly to three procedures that involved the HSS. It could be extended to cover all other procedures in IMS. The access network was assumed to be a DSL link. This was mainly due to the lack of measurements of test networks. The work can also extend to cover those access networks that are more latency critical such as WiFi.

## References


[1] J. Rosenberg *et al.*, "SIP: Session Initiation Protocol," Tech. Rep. RFC 3261, IETF, June 2002.

[2] P. Calhoun *et al.*, "Diameter base protocol," Tech. Rep. RFC 3588, IETF, September 2003.

[3] H. Schulzrinne and J. Rosenberg, "The session initiation protocol: Internet-centric signaling," *IEEE Communication Magazine*, pp. 134–141, October 2000.

[4] S. Faccin, P. Lalwany, and B. Patil, "IP multimedia services: analysis of Mobile IP and SIP interactions in 3G networks," *IEEE Communication Magazine*, pp. 113–120, January 2004.

[5] K. Wong and V. Varma, "Supporting real-time IP multimedia services in UMTS," *IEEE Communication Magazine*, pp. 134–141, October 2000.

[6] 3GPP TS 24.228 v5.13.0 (2005-06), "Signaling flows for the IP Multimedia call control based on SIP and SDP; Stage 3 (Release 5)," June 2005.

[7] 3GPP TS 24.229 v6.9.0 (2005-12), "IP Multimedia call control protocol based on SIP and SDP; Stage 3 (Release 6)," December 2005.

[8] 3GPP TS 29.228 v6.9.0 (2005-12), "IP Multimedia subsystem Cx and Dx interfaces; Signaling flows and message contents (Release 6)," December 2005.

[9] 3GPP TS 24.141 v6.5.0 (2005-09), "Presence service using IP Multimedia core network subsystem; Stage3 (Release 6)," September 2005.

[10] 3GPP TS 29.328 v6.8.0 (2005-12), "IP Multimedia subsystem Sh interface; Signalling flows and message contents (Release 6)," September 2005.

[11] H. Hannu, "Signaling compression (SigComp) requirements and assumptions," Tech. Rep. RFC 3322, IETF, January 2003.

[12] Nortel Networks, "A comparison between GERAN packet-switched call setup using SIP and GSM circuit-switched call setup using RIL3-CC, RIL3-MM, RIL3-RR and DTAP, Rev. 0.4," Tech. Rep. GP-000508, 3GPP TSG GERAN 2, November 2000.