PERFORMANCE EVALUATION OF A SIP BASED PRESENCE AND INSTANT MESSAGING SERVICE FOR UMTS

M. Pous¹, D. Pesch¹, G. Foster², A. Sesum²

¹Cork Institute of Technology; Ireland  ²Motorola Global Network, United Kingdom

ABSTRACT

With ever increasing penetration of IP technologies and the tremendous growth in wireless data traffic, the wireless industry is evolving the mobile core networks towards IP technology. The 3G Generation Partnership Project (3GPP) is specifying an IP Multimedia Core Network Sub-System (IMS) in UMTS Release 5/6, which is adjunct to the UMTS GPRS CN. This IP-based network will allow mobile operators to provide commonly used Internet applications to wireless user.

The UMTS IMS uses the Internet Engineering Task Force (IETF) defined text based Session Initiation Protocol (SIP), to control a wide range of anticipated IP-based services. Text based protocols are easier to develop than bit-wise presenttions and likely to be more expediently taken to market, as they allow vendors/operators to open the session/application space to a much broader development community. This is likely to offer new services such as wireless multimedia calls, chat, presence and push services.

This paper presents results of a performance evaluation of the SIP based Presence and Instant Messaging service as being standardised by 3GPP and IETF based on a model implemented in a UMTS system simulator called the Framework for Radio Architecture Modeling (FRAM).

INTRODUCTION

The tremendous success of the GSM Short Message Service (SMS) motivates the development of a messaging service for UMTS that can provide a richer and faster way of communication. Presence and Instant Messaging services have a strong following on the Internet with services such as AOL IM, Yahoo Messenger, Jabber, and ICQ. A similar service does not exist in the mobile domain yet, but efforts by 3GPP are underway to define such a service in the mobile environment, which will utilise the IETF Session Initiation Protocol (SIP) [8] and its SIMPLE extension for Presence and Instant Messaging. This messaging service combined with presence awareness (always on-line paradigm) will compliment and may even replace the present day SMS.

In order to provide insight into the performance that can be expected from such as service, a system model has been implemented in a computer simulation environment. Initial results indicate that this service will put a significant burden on the UMTS Radio Access Network (RAN) as well as the Core Network (CN) to the large message sizes of the text based signalling protocols used.

UMTS NETWORK ARCHITECTURE RELEASE 5/6

While the first phase of UMTS based on Release 99 still includes two distinct core networks, one for circuit-switched (CS) and one for packet-switched (GPRS) support, UMTS Release 5/6 moves towards an all IP Multimedia Core Network Subsystem (IMS), with full IP packet support including full UMTS packet call control capabilities. The IMS operates in conjunction with the PS Core Network.

Figure 1 depicts the main components of the UMTS Rel6 architecture, including the UTRAN, the PS CN elements and the elements of the IMS.

Figure 1- UMTS with PS CN & IMS Domain

The IMS consist of Call State Control Function (CSCF), Media Gateway Control Function (MGCF) and Media Gateway (MGW). Basically the MGCF controls the MGW used in connections to external networks. The CSCF provide the logic of how transactions using IMS are treated. The CSCF may assume several functions, depending on whether it is operating as a Proxy, Interrogating or Serving CSCF [2, 3]. UMTS Rel.5/6 standardises the IETF Session Initiation Protocol (SIP) as the standard session, call, and wide area mobility management protocol for the Non-Access Stratum between the mobile user equipment (UE) and the IMS core network, providing direct transfer of signalling messages and transparent flow of data. Several extensions of SIP are being specified by IETF IMPP and SIMPLE Working Groups, in order to enhance its capabilities.

PRESENCE AND INSTANT MESSAGING SERVICE

Presence and Instant Messaging service, as defined in RFC 2778 [7], allows users to subscribe to each other and be notified of changes in state, e.g. going off-line, changing contact details, etc., providing a simple communication between online users. Such services are based on the IETF SIP protocol. SIP provides useful mechanisms for presence and session-oriented communication applications but not for messaging and asynchronous notification. SIMPLE and IMPP working groups have extended SIP's capabilities by introducing methods for messaging and notification [10].

© 2003 The Institution of Electrical Engineers. Printed and published by The IEE, Michael Faraday House, Six Hills Way, Stevenage, SG1 2AY
PRESENCE SERVICE

Figure 2 depicts the 3GPP Release 6 specified presence architecture model [6] with the major functional entities.

![Figure 2- Presence Service Model](image)

The Presence Server, which resides in the presence entity's home network, manage and distribute the information to interested parties, called watchers. The two set of entities involved, presentity and watchers, are either internal or external to the home network and access network. Watchers access the server through presentity proxies.

INSTANT MESSAGING SERVICE

The exchange of content between the participants in near real time is realised with instant messaging. Generally, the content is short text messages and its transfer fast enough in order to maintain an interactive conversation. Each message can be sent independently using the SIP MESSAGE method, or messages can be associated into sessions that are initiated using SIP. The first approach is often referred as pager-mode, due to its similarity to the behaviour of two-way pager devices, and is used when small short IMs are sent to a single or reduced number of recipients. On contrast the second approach, called session-mode messaging, is required for extended conversations, joining chat groups, etc. Both approaches, defined by SIMPLE, are considered in our model.

UMTS SIGNALLING PROCEDURES

The following section provides a description of the main UMTS procedures [11,12] relevant to the current investigation outlined in Figure 3.

![Figure 3 – UMTS Procedures](image)

On power on, a UE first needs to attach to the network using a UMTS GPRS Attach procedure, which causes Mobility Management (MM) context to be established at the UE and the SGSN. During an Attach, the UE is authenticated and the HLR is updated with its location information.

PDP CONTEXT ACTIVATION

In UMTS, in order to enable any transfer of data in the PS domain, a PDP context must be established between the UE and the GGSN using the PDP Context Activation procedure. This procedure may be initiated by the UE or by the network depending on the direction of the session. A PDP context establishes an association between the UE and the CN for a given QoS on a specific NSAPI. It contains routing information that is used to transfer the PDP PDUs between the UE and the GGSN. Activation of PDP context entails checking of the UE's subscription selection of the APN and the host configuration. Once a primary PDP context has been established for a given PDP address, a secondary PDP context can be activated re-using the PDP address and other information associated with the already active PDP context, but with a different QoS profile.

IMS SIGNALLING PROCEDURES

Once the connection is established, the UE needs to access the IM sub-system. IMS makes use of SIP signaling flows and procedures required for the provision of presence and IM service detailed below.

PROXY CSCF DISCOVERY

In the PDP context activation procedure, besides acquiring a PDP context within the PS CN, the UE also can find a Proxy CSCF. This is a SIP proxy, as defined before, and is the contact point of the UE and is located in the same network as the GGSN, i.e. in the home or visited network, depending on whether the mobile is or not roaming.

APPLICATION LEVEL SIP REGISTRATION

In order to request the services provided by the IM domain, the user must perform an application level registration. This can only be done after registration with the access network is complete and after a signalling connection has been established for transfer of IP signalling. In other words, the user needs to activate a PDP context to transfer IM related SIP signalling. The QoS parameters specified in activation of the context are appropriate for IM subsystem related signalling.

Figure 4 shows the flow of messages for registration of the UE with its Serving CSCF, assuming the UE was not previously registered. As shown, the S-CSCF authenticates the mobile before registration is successful.
Once the mobile has purchased all the signaling required, instant messages can be exchanged between the two or more peers involved in the session. Traffic models govern the generation of bearer traffic within a user session. Each service is characterized by a traffic model. A service session may last for the duration of the PDP session or several service sessions may be initiated within a PDP session.

An IM session consists of several file downloads/uploads, as shown in the diagram below. Each file is made up of packets. The traffic model characterizing SIP IM defines the average number of files within an IM session, the mean time between file downloads/uploads, the average size of a file, the average packet size, and the average time between packets.

Considering the similarity of IM session-mode service and internet chat, the traffic model parameters have been obtained from an intranet study on chat messaging.

**SIMULATION ENVIRONMENT**

The simulation environment is event-driven and implemented in C++. An overview of the structure of the simulation is presented in Figure 6, which can be viewed as consisting of...
four main phases: reading i
(core simulation) and post-processing of data collected.

The user profile information is required to activate a mobility
model and a session/call generator model. The mobility
model is used to determine the residence time of a mobile in a
cell and therefore the time at which the mobile crosses a cell
boundary. The call and session models determine when the
mobile initiates and terminates a session/call. These actions
then drive the simulation by invoking the UMTS and SIP
signalling flows described above, depending on the RRC,
PMM and presence states of the mobile. In addition, for PS
simulations, once a session is activated, traffic models, based
on the service-related data, are used to determine the periods
of bearer activity and inactivity during the PDP session.
These dictate times at which transitions between RRC/PMM
states take place. The call/session model used in the
simulation assumes a Poisson arrival of PDP sessions/ CS
calls and a negative exponential distribution for session and
call holding times. The rates of arrival and mean holding
times are specified as user inputs. The Signaling flows
(UMTS and SIP) and traffic model in our investigation are
described in more detail in the next sections.

MODEL REFERENCE AND ASSUMPTIONS

The IMS domain functional model, defined in 3GPP TS
23.228 [2], does not give us the representative physical model
that we require to model realistic transmission and queuing
delays during message exchange. We propose a practical
realisation of the UMTS IMS as shown in Figure 7.

![Figure 7 - Model Reference of physical IMS](image-url)

A physical SIP Server Network Element (NE) is shown,
which host the S-CSCF, the P-CSCF and the Servers required
for the Presence and IM services. This seems a reasonable
concept as every network using Presence will require those
functionalities and the large number of interactions between
them. By locating these two entities together we reduce the
transmission delay. We assume the computing platform,
providing a 25ms SIP-to-SIP message turnaround duration for
up to 800k subscribers per NE. We make a similar
assumptions for the HSS. Co-locating the I-CSCF and the
HLR with the 3G extensions provides for fast interrogation
behaviour and we estimate that for similar sized platform to
the SIP server, interraations and HSSCx data retrievals can take
~55ms per transaction (read, read/write and forward
averaged).

Based on this reference model, a typical UMTS network
configuration is analysed for the dense urban environment,
where UTRAN employs 784 NodeBs and 4 RNCs. The PS
domain consists of two SGSNs and one GGSN. With this
configuration, all core network-related and mobility
signalling scenarios including handover are accounted for.
For this study we assume that mobiles are always located in
their home network, which requires only one P-CSCF, one S-
CSCF, one I-CSCF and one Server in the IMS. We consider
that these elements are co-located with the GGSN. The
system is analysed using the SIP signalling call control model
described in Rel5 3GPP TS 24.228[3]. The Presence and IM
service functional model is based on Rel6 3GPP TR 24.841
[5] and the work of the IETF SIMPLE and IMPP working
groups [10]. The simulation results are obtained, assuming a
subscriber population of 25,000 users, accessing the Presence
and Instant messaging service only. The IM model used is the
'session mode' described previously. The session type
supported is mobile-to-mobile (MM), however the M2M
scenario is analysed by separating the two counterparts
involved in the session and its networks elements, Mobile
Origination (MOO) and Mobile Termination (MMT). A
single network operator performs both parts, origination and
termination.

RESULTS

Results collection is confined to session UMTS and IMS
signaling delay analysis, where the scenario described above
is simulated. The model keeps track of the Radio Access
Network delay (RAN delay) and the Core Network delay for
the different procedures required in our case. We consider
that all UMTS (PMM) signaling uses a 3.8Kbit/s Radio
Bearer. However both IMS (SIP) signaling and data bearer
(i.e. Instant Messages exchange) use a DCH at 64kbit/s.

Table 1 show the time required to activate the primary PDP
context that carries all the IMS SIP signaling. It also shows
the SIP signaling delays for the Registration method.

<table>
<thead>
<tr>
<th>Delays (sec)</th>
<th>Primary PDP Activation</th>
<th>Registration</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAN</td>
<td>1.471</td>
<td>0.795</td>
</tr>
<tr>
<td>Core</td>
<td>0.770</td>
<td>0.876</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2.241</td>
<td>1.671</td>
</tr>
</tbody>
</table>

Table 1 – PDP Activation & Registration Delays

As explained earlier, before the MMO communicates with the
MMT, both need to subscribe to the presence server in order
to know about the state of the other entity. We assume that the MMT has already been registered and subscribed and only the MMO delay for these procedures is considered as represented in Table 2.

<table>
<thead>
<tr>
<th>Delays (sec)</th>
<th>Subscription Presence List Server</th>
<th>Subscription Presentity</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAN</td>
<td>0.449</td>
<td>0.380</td>
</tr>
<tr>
<td>Core</td>
<td>0.525</td>
<td>0.400</td>
</tr>
<tr>
<td>TOTAL</td>
<td>0.974</td>
<td>0.780</td>
</tr>
</tbody>
</table>

Table 2 – SIP Subscription Delays

The SIP Invite delay shown in Table 3 includes the time taken for a secondary PDP context activation, which is approximately 1.94 seconds (1.18 seconds in the RAN and 0.76 seconds in the CN).

<table>
<thead>
<tr>
<th>Delays (sec)</th>
<th>MMO Invite</th>
<th>MMT Invite</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAN</td>
<td>2.698</td>
<td>4.624</td>
</tr>
<tr>
<td>Core</td>
<td>2.012</td>
<td>2.569</td>
</tr>
<tr>
<td>TOTAL</td>
<td>4.709</td>
<td>7.193</td>
</tr>
</tbody>
</table>

Table 3 – SIP INVITE method Delay

The higher delay in the MMT part is expected as it includes signaling for UMTS paging.

As expected, the required SIP signaling set-up introduces considerable delays. The IM service may not be regarded as instant as 16.6 sec are required to establish the session connection. This is partly due to the large number of SIP messages required to establish the session and partly due to the large size of the text based SIP messages. Therefore, SIP signaling optimization techniques are required. The introduction of SIP message compression over the air interface as proposed in [13] can reduce these delays. However, as the SIP signaling transmission rate is higher than UMTS signaling one, only the 29% of the total delay is SIP RAN delay.

CONCLUSIONS

We have presented a performance analysis of a UMTS Instant Messaging and Presence Service as specified by 3GPP and IETF. The service uses the Session Initiation Protocol as its main session management protocol. The performance analysis focuses on the instant messaging session set-up delay on an end-to-end basis. Simulation results of the service within a comprehensive model of the UMTS radio access, PS core network, and the IMS show that instant messages are not necessarily transmitted in near instant fashion but that substantial delays with an average of about 16.6 seconds are encountered. This delay renders the service as a basis for an online text base chat unsuitable. In order to improve the performance of the service mechanisms such as SIP message compression can be introduced to reduce the signaling delay in the RAN. However, optimization of the core network and IMS architecture as well as computing platform performance for these network elements are also required as 70% of the total end-to-end delay is encountered in the core network.

The authors acknowledge the support of the Irish Department of Education and Science Technological Sector Research Programme Strand 3 in funding parts of the work reported in this paper under grant CRS/00/CR02.

REFERENCES

[3] 3GPP TS 24.228, “Signalling flows for the IP multimedia call control based on SIP and SDP; Stage 3”
[5] 3GPP TS 24.841, “Presence service based on Session Initiation Protocol (SIP); Functional models, information flows and protocol details” (Release 6)