Performance Estimation of Efficient UMTS Packet Voice Call Control

Gerry Foster, Maria Isabel Pous, Dirk Pesch, Amardiya Sesmun, Valerie Kenneally
System Engineering Analysis Group, Motorola & Cork Institute of Technology

Abstract— **UMTS was envisaged as a "Universal" Mobile telecommunications system with multiple access options and full IP packet support. However, 3GPP Rel99 is a single access technology, and its core is still switch centric. Rel4 further optimises the air interface. Rel5 first adds an IP Multimedia Core Network Sub-System (IMS) adjunct to the UMTS Packet Switched (PS) GPRS CN bearer. This adds basic PS call control trial capabilities. Rel6 will provide full UMTS packet call control (CC) capabilities (eg: security, emergency & QoS support).**

The Rel5/6 standards employ the text based call control protocol: SIP. Text based protocols are easier to develop than bit-wise presentations and likely to be more expediently taken to market. However, initial indications as to the delays associated with packet call control using SIP have concerned operators about the viability of IP-based services over a UMTS air interface. This has led UMTS vendors / operators to invest in the standardization of ways to wireless enable the efficient use of SIP with such measures as protocol compression. This paper explains GSM, UMTS circuit switched call control and SIP call control as applied to UMTS and then compares and contrasts the relative performance of both schemes using results derived from a UMTS system simulator called the Framework for Radio Architecture Modelling (FRAM).

I. BACKGROUND

With the advent of 3GPP R6 capability, all of the associated features of a circuit switched call and the additional features that come with a packet call control (CC) protocol become available to the PS user. 3GPP Release 6 (Rel6) completes IMS standardisation adding full security, emergency and QoS support. At this stage of UMTS development, we have a full PS call based system using the IETF defined text based, packet call control protocol, SIP.

Although text based protocols are easier to develop than bit-wise presentations, text over UMTS makes inefficient use of the air interface. This has led to UMTS vendors / operators investing in the standardisation of ways to make SIP more efficient for the wireless domain.

This paper presents and examines the trade-offs in the process of applying SIP and packet optimised voice to UMTS and provides estimates of likely optimised performance, as compared to the existing GSM/UMTS circuit switched call control performance. The paper explains 3GPP and SIP call control as applied to UMTS and then compares and contrasts the relative performance of both schemes using results derived from a UMTS system simulator called FRAM.

II. GSM CALL CONTROL

The GSM mobile standard has the largest installed base of any available today, worldwide. The architecture for this standard is as shown below. 3GPP TS 24.008 [1] mobile CC operates between the Mobile and the Mobile Services Switching Centre (MSC). Between the MSC and the edge of the PSTN ISUP is operated over the typical ISDN based PSTN. From the PSTN to the fixed phone either analogue TUP or a continuation of ISDN connects the Fixed Phone.

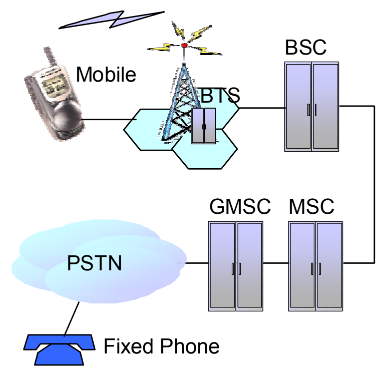

Figure 1. GSM Architecture

III. 3GPP UMTS CALL CONTROL

In UMTS, CC is enhanced from the 04.08 standard to the 3GPP 24.008 standard [1] to encompass the new radio resource protocol, Radio Resource Control RRC and its UTRAN based mobility. In fact a high level UMTS call involves two main specifications from the UMTS standard: 3GPP TS 25.331[2] for RRC management and 3GPP TS 24.008 [1], which includes the Circuit Switched (CS) Mobility Mgt (MM) and CC.

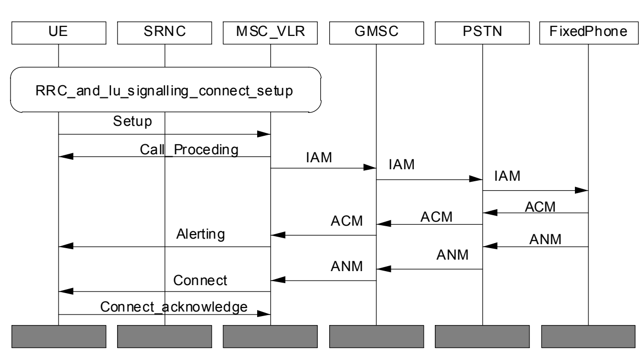

Figure 2. MO Call setup

Figure 2 shows a Mobile Originated (MO) call set up. This procedure requires a number of sub-procedures including several protocols RRC, Mobility Management (MM) and Call Control (CC), which makes them fairly complex procedures to execute. However, these protocols have bit-wise presentations and are usually coded in ASN.1, which is a universally understood industry standard and is fairly efficient for a band limited radio interface such as UMTS. The same procedures are required in the case of Mobile Terminated (MT) calls.

The UMTS Release 99 architecture is similar to GSM. However, with its CDMA interface the user can now be connected to multiple BTS or NodeB at a time allowing softer handover and these NodeBs are managed by a Radio Network Controller (RNC), not a BSC. The 3G-MSC is also very similar to the GSM MSC, with the exception that it is now ATM(STM-1) based and not TDM-PDH(E1/T1 etc).

Typical benchmarked call set-up delays for a UMTS call and a GSM call are as follows:

TABLE I. UMTS and GSM Benchmark Performance

Delay (s)	MO Call	MT Call	MM Call
GSM	2	2.2	4
UMTS	1.7	2	3.4

IV. SIP Call Control

A. *SIP in IP*

The Session Initiation Protocol (SIP) is an application-layer control protocol that can establish, modify and terminate multimedia sessions or calls [3] such as Internet telephony. SIP has been designed as part of the IETF suite of protocols to provide for multimedia services, including protocols for reserving network resources, for handling real-time data and quality-of-service. However, the operation of SIP is independent of any of these protocols and it can be used in conjunction with other call setup and signalling protocols.

SIP supports five aspects of session control: user location, user capabilities, user availability, call set-up and call handling.

B. *SIP in UMTS*

UMTS Release 99 includes two distinct core networks for support of PS (GPRS) and CS services separately, where the CS domain is very switch-centric. Call control for CS services, as mentioned earlier, is similar to call control in GSM. As UMTS evolves to Release 6, there is a move towards an 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 3 depicts the main components of the UMTS Release 6 architecture, including the UTRAN, the PS Core Network (CN) elements and the elements of the IP Multimedia Subsystem (IMS).

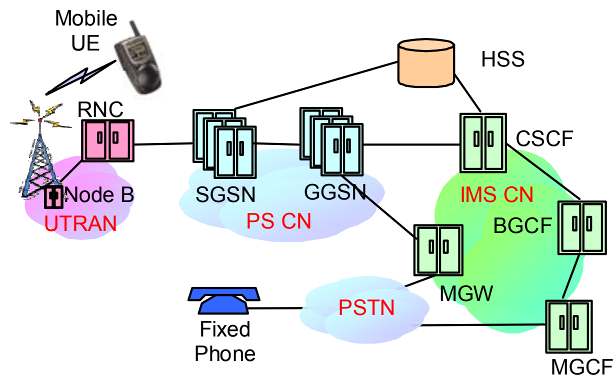

Figure 3. UMTS with PS CN & IMS Domain

The UTRAN consists of the Node B and the RNC, where the Node B operates as the base station and the RNC is the controlling unit for a number of Node Bs and for access to the radio network.

The Serving and Gateway GPRS Support Nodes constitute the PS Core Network. Each SGSN is connected to a number of RNCs, thereby acting as serving node for all mobiles that are under coverage of these serving RNCs. The GGSN is the point of PDN interconnection between external networks and the UMTS PLMN.

The IMS includes several components such as the CSCF, BGCF, MGW, MGCF and the HSS. The Call State Control Function may assume several functions, depending on whether it is operating as a Proxy, Interrogating or Serving CSCF [4]. The Proxy CSCF (P-CSCF) is the first point of contact between external networks and the IMS and behaves as a proxy, thereby processing requests for services and forwarding these, possibly after translation. The Interrogating CSCF (I-CSCF) acts as the point of contact within an operator's network for all connections destined to a subscriber of that network operator, or for a roaming subscriber currently located within that network operator's service area. The I-CSCF is used to provide routing to the correct S-CSCF. The Serving CSCF (S-CSCF) is responsible for performing session control services for the UE under its coverage and maintains session states as required by the network operator for support of the services.

The Breakout Gateway Control Function (BGCF) serves the purpose of selecting the network in which PSTN/CS domain breakout is to occur. The Media Gateway (MGW) and Media Gateway Control Function (MGCF) are nodes required for interworking with the PSTN or other CS domains.The Home Subscriber Server (HSS) is the database storing all subscriber related data, including application server subscription information, filter criteria and authentication data.

C. *UMTS procedures*

The following section provides a description of the main UMTS procedures relevant to the current investigation.

1) *UMTS GPRS Attach*

On power on, a UE first needs to attach to the network using a UMTS GPRS Attach procedure, which causes MM contexts 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

2) PDP Context Activation

In UMTS, in order to enable any transfer of data in the PS domain, a PDP context needs to 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 specific QoS on a specific NSAPI. It contains routeing information that is used to transfer the PDP PDUs between the UE and the GGSN. Once a primary PDP context has been established for a given PDP address, a secondary PDP context can be activated that re-uses the same PDP address and other information associated with the already active PDP context, but with a different QoS profile.

D. IMS Procedures

IMS makes use of SIP signalling procedures and the flows required for registration and call setup are described below.

1) Application level Registration

In order to request the services provided by the IM domain, the user needs to 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 for transfer of IM related SIP signalling. The QoS parameters specified in activation of the context are appropriate for IM subsystem related signalling.

Besides acquiring a PDP context within the PS CN, the UE also needs to find a Proxy CSCF. The process of P-CSCF discovery can be done as part of the PDP Context Activation procedure. During registration, the UE has to be authenticated by the S-CSCF.

2) Call setup flows and use of PDP contexts

After registration with the IMS, the user is set up to access multimedia services using SIP call control procedures. The MO call setup flow is shown below (Figure 4). It is assumed that a primary PDP context has already been activated for transfer of IP signalling. The session set-up starts with the INVITE message being sent from the caller to the callee. The two-end parties negotiate the media characteristics that they will support for the session. After these are determined, resource reservation is required, which entails creating a secondary PDP context for transport of the required media, and setting up the corresponding radio bearers. If resource reservation and H.248 connection creation and modification are successful, the terminating point sends a SIP 200 OK final response and the originating mobile replies with a SIP ACK message to confirm the session set-up.

In the MT case, if there is no PDP context currently active, as the GGSN receives SIP signalling destined for the UE, a network initiation PDP context activation procedure is executed and radio bearers are set up for transfer of IP signalling. Subsequently, as the call set-up proceeds, after media characteristics are negotiated, resource reservation is performed.

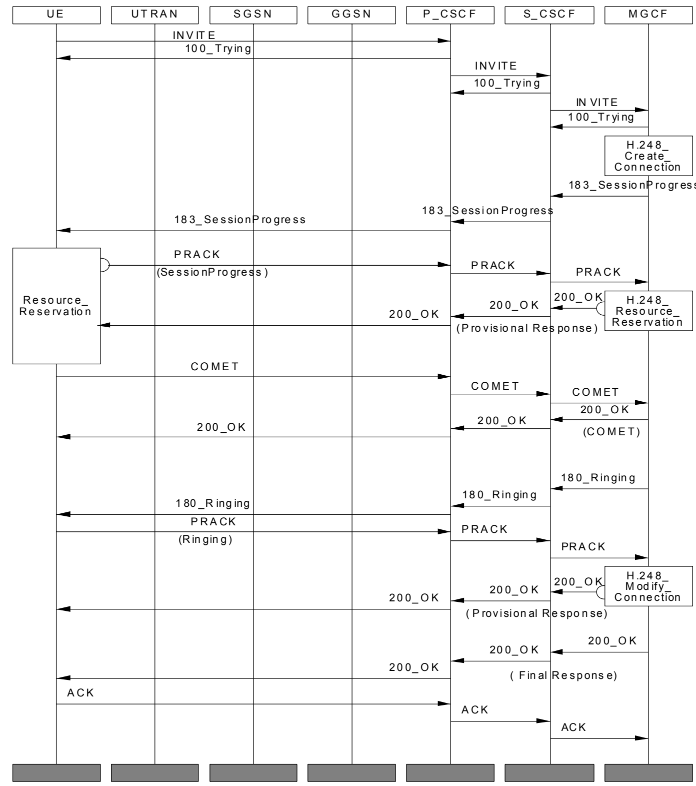

Figure 4. SIP MO Call Set Up

V. Model Reference

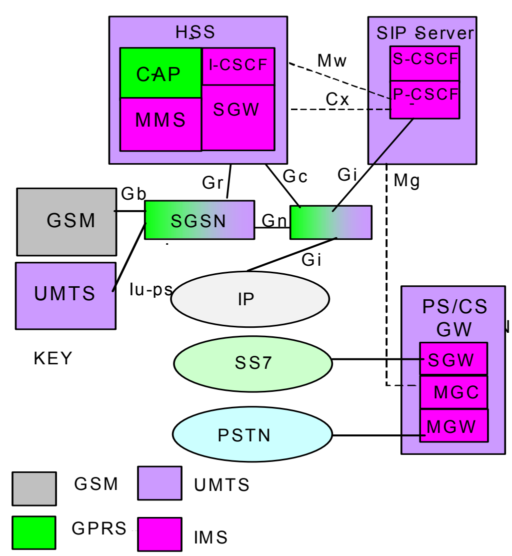

Figure 5. Model Reference of Physical IMS

Figure 3 shows the functional model of the IMS domain. However, it does not provide a representative physical model required to model delays realistically in order to determine 3GPP UMTS SIP based call set up.

To model physical delays, a more practical realisation of a UMTS IMS is proposed in Figure 5. It shows a physical SIP server Network Element, which hosts the S-CSCF and the P-CSCF. This seems a reasonable concept as every network will require both of these functionalities and interaction between them is likely to be largest for the local network. We assume this is typically a box of 16 SUN Netra 20 type servers, providing a 25ms SIP to SIP message turnaround duration for up to 800k subs per NE. We make similar assumptions for the HSS. Co-locating the I-CSCF and the HLR together provides for fast interrogation behaviour and we estimate interrogations and HSS Cx data retrieval to take ~55ms per transaction (read, read/write and forward averaged).

VI. TEXT-BASED COMPRESSION USING CACHE AND BLANK APPROACH (TCCB) COMPRESSION

TCCB is a compression scheme defined for text-based protocols [5] like SIP and SDP. It removes all redundant header and payload information between the User Agent and the Peer Core Network entity(Proxy-CSCF). It also caches this information in the local memory (indexed by a unique identifier) for future compression/ decompression purposes.

The TCCB compression modelled uses the "From" and "To" fields of the SIP headers as the table index. A numerical representation of the SIP headers is used in the compressed messages. The first time that the user sends an INVITE message its SDP body is uncompressed, as shown in Table II.

As shown, the initial INVITE message can only be compressed to about 9% because most of the information is not redundant. A compression level between ~30% and ~50% can be achieved on other messages.

TABLE II. Initial MO Compression (bytes)

UPLINK		**UE**		**%**
		No compression	**Compressed**	**Compressed**
INVITE	***Header***	497	433	12.9
	SDP	219	219	0
	Total	*716*	*652*	*8.9*
COMET	*Header*	346	231	33.2
	SDP	185	40	78.4
	Total	*531*	*271*	*49*
PRACK (180)	*Total (SDP=0)*	*365*	*245*	*32.9*
ACK	*Total (SDP=0)*	*311*	*222*	*28.6*
DOWNLINK		**P-CSCF**		**%**
		No compression	**Compressed**	**Compressed**
183-Session Progress	*Header*	600	402	33
	SDP	223	106	52.5
	Total	*823*	*508*	*38.3*
180-Ringing	*Total (SDP=0)*	***354***	***225***	***36.4***
200-OK	*Header*	355	216	39.2
	SDP	185	89	51.9
	Total	***540***	***305***	***43.5***

Table III shows that the level of compression achieved with subsequent calls is even better, where results show that the INVITE message can now be compressed by 30%.

TABLE III. Subsequent MO Compression

UPLINK		**UE**		**%**
		No compressed	**Compressed**	**Compressed**
INVITE	*Header*	497	370	25.6
	SDP	219	134	38.8
	Total	***716***	***504***	***29.6***
DOWNLINK		**P-CSCF**		**%**
		No compressed	**Compressed**	**Compressed**
183-Session Progress	*Header*	600	382	36.3
	SDP	223	106	52.5
	Total	***823***	***488***	***40.7***

VII. PRE-CACHE

Another technique that has been analysed with TCCB compression engaged, is P-CSCF pre-caching. This pre-caches the UL and DL SDP fields and the FROM field as each user registers at their P-CSCF, and a set of Register message extensions to pre-select say up to 3 SDP types. For UMTS this is estimated to be fairly easy to estimate for the next few years and is unlikely to change for a GPRS PMM Attach duration for one day. Typically the UE would store a PDP pre-cache for UMTS-AMR (Speech), MPEG2000(Video Clips) and JPEG2000(Images). These fields are then not transmitted at MS_UE origination of a message but are reinserted in the standard SIP message in the network at the P-CSCF. The compression achieved by applying both of these techniques is detailed in Table IV. In this case, a compression level ranging from ~45% to ~70% can be achieved.

TABLE IV. MO Compression & Pre-Cache

UPLINK		**UE**		**%**
		No compressed	**Compressed and Pre-caching**	**Compressed**
INVITE	*Header*	497	253	49.1
	SDP	219	50	77.2
	Total	***716***	***303***	***57.7***
COMET	*Header*	346	125	63.9
	SDP	185	40	78.4
	Total	***531***	***165***	***68.9***
PRACK (180)	*Total (SDP=0)*	***365***	***140***	***61.6***
ACK	*Total (SDP=0)*	***311***	***116***	***62.7***
DOWNLINK		**P-CSCF**		**%**
		No compressed	**Compressed and Pre-caching**	**Compressed**
183-Session Progress	*Header*	600	296	50.7
	SDP	223	106	52.5
	Total	***823***	***402***	***51.2***
180-Ringing	*Total (SDP=0)*	***354***	***120***	***66.1***
200-OK	*Header*	355	198	44.2

SDP	185	89	51.9
Total	***540***	***287***	***46.9***

In all three cases considered, similar compression levels have been observed for MT calls.

VIII. Model Implementation

A typical UMTS network topology is analysed for the dense urban environment, where the UTRAN consists of 784 Node Bs and 4 RNCs. The CS domain is comprised of 2 MSCs and 1 GMSC whereas the PS domain consists of 2 SGSNs and 1 GGSN. With this configuration, all core network-related and mobility signalling are accounted for. In the IMS, we assume that there are 1 P-CSCF, 1 S-CSCF, 1 I-CSCF and 1 MGCF. These elements are co-located with the GGSN. Note that in this analysis, the caller is assumed to be located in its home network.

In this paper, the scenarios considered for comparative analysis are: GSM circuit-switched (CS), UMTS circuit-switched, UMTS with a GPRS CN and IMS, UMTS (with IMS) with call control compression engaged and UMTS (with IMS) with compression and pre-cacheing engaged. The systems are analysed using the call model described in the 3GPP TS 24.008 for GSM and UMTS CS performance estimation and the SIP call model as detailed in the 3GPP TS 24.228 [4] for the Release 5 IMS-domain using UMTS (GPRS) as the reference system. The protocol compression scheme analysed in this paper is TCCB [5].

The simulation results are obtained, assuming a subscriber population of 210,000 users, accessing SIP voice service only. The call types supported are mobile originated (MO) and mobile terminated from/to a fixed user.

IX. Results

In this section, results are presented to show call set-up delays for different scenarios in the IMS, namely without any compression, with compression as in TCCB and with pre-cacheing. The split between the delay spent in the radio access network and the core network is also detailed.

The following tables present delays for PDP context activation and for SIP call set-up. The SIP call set-up delay includes the time taken for secondary PDP context activation, which is approximately 1.94 seconds (1.18 seconds in the RAN and 0.76 seconds in the CN). In the MT cases, the NI PDP Activation delays include paging delays.

TABLE V. Initial SIP call set-up delay

Delay (s)	Primary NI PDP Activation	MT SIP Call Set-up	
		Compress	No Compress
RAN delay	2.619	3.153	4.228
Core delay	1.583	2.396	2.397
Total delay	4.202	5.55	6.624
	Primary PDP Activation	**MO SIP Call Set-up**	
		Compress	**No Compress**
RAN delay	1.564	3.256	4.169
Core delay	0.77	2.692	2.692
Total delay	2.334	5.948	6.861

As expected, without compression, the RAN delay contributes a significant portion (68%) of the total delay. When TCCB compression is applied, the RAN delay is reduced by approximately 23% for both MO and MT cases. For subsequent calls, results show a 25% reduction in the RAN delay for MO calls and a 29% reduction for MT calls.

TABLE VI. Subsequent SIP call setup delay

Delay(s)	MT SIP Call Setup		MO SIP Call Setup	
	Compress	No Compress	Compress	NoCompress
RAN	3.02	4.228	3.137	4.169
Core	2.396	2.397	2.692	2.692
Total	5.417	6.624	5.828	6.861

With pre-cacheing and TCCB compression engaged, further reduction of delays are achieved. As shown in Table VII, RAN delays are reduced by about 40%.

TABLE VII. SIP call setup delay with TCCB compression and pre-cacheing

Delay (s)	MT SIP Call Setup		MO SIP Call Setup	
	Compress Pre-cacheing	No Compress	Compress Pre-cacheing	No Compress
RAN	2.551	4.228	2.517	4.169
Core	2.396	2.397	2.691	2.692
Total	4.947	6.624	5.209	6.861

X. Conclusions

This paper has presented results on call set-up delays for GSM, UMTS Release 99 and Release 5. As shown, without compression, SIP call set-up delays are excessive compared to GSM or UMTS delays. Techniques of reducing the call setup delays have been investigated and results indicate that with TCCB compression enabled, the RAN delay can be reduced by about 24%. Note that with TCCB, the level of compression achieved and hence the reduction in delays improves with subsequent calls. Using pre-cacheing with TCCB compression enables a further reduction in delays to about 40%. Although an improvement is noted, the call setup delays are still high.

Possible solutions for further investigation include using different cache memory for uplink and downlink at the CSCF, to enable further blanking of fields in messages. Reducing the number of message transfers in the core to decrease the core delay is another option.

References

[1] 3GPP TS 24.008 UMTS Mobile radio interface layer 3 specification; Core Network Protocols – Stage 3

[2] 3GPP TS 25.331 UMTS Radio Resource Control Protocol Specification

[3] RFC 2543, "SIP: Session Initiation Protocol", March 1999

[4] 3GPP TS 23.228, IP Multimedia System (IMS) Stage 2

[5] IETF Draft, "Text-based Compression Using Cache and Blank Approach (TCCB)", July 2001