END-TO-END UMTS QUALITY OF SERVICE ARCHITECTURE FOR THE SUPPORT OF REAL-TIME IP MULTIMEDIA SERVICES IN UMTS R5

S Leroy, L Bos, J De Vriendt
Network Strategy Group, Alcatel, Francis Wellesplein 1, 2018 Antwerp, Belgium

ABSTRACT

3G promises to be an enabler for many new services: videoconferencing, interactive gaming, streaming, etc. Such applications are typically QoS (Quality of Service) sensitive. This requires new functionality in the UMTS QoS architecture. In 3GPP work on QoS is ongoing; the basic mechanisms are included in UMTS release 5, the more advanced features will be part of UMTS release 6.

This document highlights UMTS QoS aspects. The main focus is on UMTS R5 as this release will introduce IP multimedia services. We clearly make the distinction between 3GPP stable decisions, topics under discussion within 3GPP, and the Alcatel position on the topics.

1. INTRODUCTION

The UMTS network will be deployed in different releases, each corresponding with a standardisation phase and a well-defined functionality. The first UMTS release is called R3 (previously also called Release 99). R3 introduces a new radio interface protocol (CDMA) together with a cleaner split, compared to 2G, between access and core network functionality. The R3 CS core network is an evolution of the GSM core network. R4 introduces the NGN (Next Generation Network) concept in the core network. The traditional MSC is split in a control part -the MSC server- and a packet based transport part – MGWs. Release 5 introduces the IP Multimedia Subsystem (IMS). The IMS system [5] comprises all the core network elements necessary to control/setup IP multimedia sessions. The GPRS [6] infrastructure is reused for the transport of the IMS data and session control packets within the UMTS network.

1.1. UMTS End-to-end QoS architecture

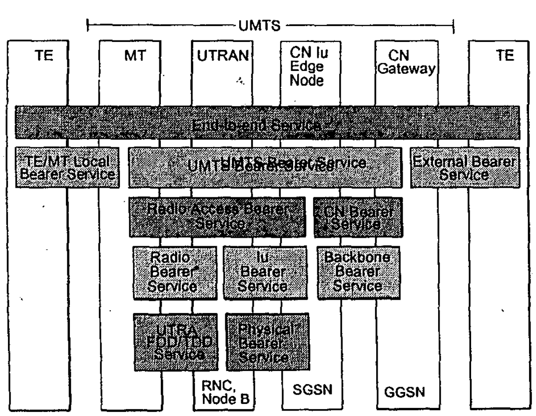


Figure 1, UMTS End-to-End QoS architecture

3GPP considers the network services end-to-end when they are from one TE (Terminal Equipment, e.g. laptop or PDA connected to mobile handset) to another TE. The UMTS end-to-end QoS architecture is a layered 'bearer service' architecture [7] (see figure 1). Each bearer service on a specific level offers its individual services using services provided by the layers below. To satisfy the end-to-end QoS demanded by the users, several bearer services (possibly located in different networks) have to be set up between the different network elements. These bearer services are TE/MT Local, UMTS and External Bearer Service.

The external bearer service depends heavily on the characteristics of the external network to which the session is routed. This can be a QoS enabled or best effort IP network. The CN gateway (GGSN) has to offer all required mappings between the UMTS QoS parameters and the specific QoS parameters necessary to offer QoS in each external network connected to it.

The UMTS bearer service provides the various services related with QoS that a UMTS operator may offer. It uses the GPRS PDP (Packet Data Protocol) context procedures to negotiate the QoS parameters that will define the channel between the MT and the GGSN. The UMTS bearer service uses the services of both the Radio Access Bearer and the Core Network Bearer. A DiffServ enabled IP Backbone bearer service is used in order to fulfil the QoS requirements of the CN bearer service. The Radio Access Bearer Service provides encrypted transport of signalling and user data between MT and CN Iu Edge Node (SGSN) with the Radio Access QoS adequate to the negotiated UMTS bearer service. This service is based on the characteristics of the radio interface and is maintained for a moving MT.

1.2. UMTS QoS Classes

When setting up a PDP context the end-user has the choice between four UMTS QoS classes, every class corresponding to a certain set and range of QoS parameters. The four UMTS QoS classes [7] are: conversational, streaming, interactive, background.

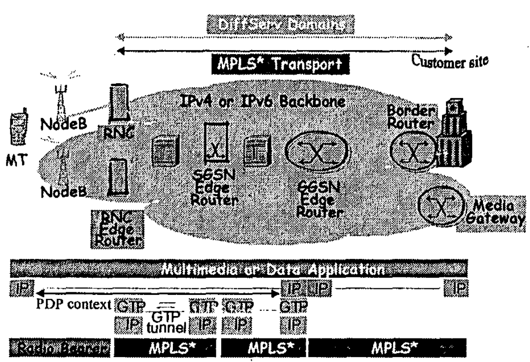
Class	Transfer Delay	Transfer Delay Variation	Low bit Error Rate	Guaranteed Bit Rate
Conversational	Stringent	Stringent	No	Yes
Streaming	Looser	Constrained	No	Yes
Interactive	No	No	Yes	No
Background	No	No	Yes	No

Table 1, UMTS QoS classes

Table 1 illustrates the different classes with their QoS requirements. Note: The difference between Interactive and Background class is that in the Interactive class, a traffic handling priority parameter can be signalled. The activation of the GPRS connection (PDP context) with the correct QoS is handled by the MT. To efficiently support applications that need a certain level of QoS in UMTS, a certain level of QoS can be requested - via the QoS parameters (e.g. UMTS QoS class) specified in the QoS profile- during PDP context activation. For IMS these QoS parameters will be derived from QoS information carried in IMS signalling (see chapter 6).

1.3. Physical Architecture of the PS domain

Figure 2 shows the overall physical architecture for the R5 PS domain. Alcatel's position for the core network as well as for an IP based Iu interface, is IP over MPLS for the transport of user and control data. Alcatel chooses MPLS because it integrates the performance and traffic management of layer 2 (ATM) with the scalability and flexibility of layer 3 (IP).



* The use of MPLS here is not a standards requirement but an Alcatel position
Note: GTP = GPRS Tunneling Protocol

Figure 2, Physical architecture PS domain (IP option Iu)

1.4. IMS elements involved in QoS provisioning

The introduction of IP multimedia services in the PS domain on a large scale puts high demands on the QoS infrastructure. The most important IMS elements that participate in the QoS provisioning are the CSCFs. The Call State Control Function (CSCF) is a SIP [1] server providing multimedia services to UMTS IP terminals.

- The *P-CSCF (Proxy-CSCF)* is the first contact for a UE in the visited network. The P-CSCF is responsible for the resource management in the visited network for IMS sessions.
- The *S-CSCF (Serving-CSCF)* is always located in the home network of the subscriber. The S-CSCF holds a copy of the IMS subscription profile of the subscriber and interfaces with the service platform.

2. QUALITY OF SERVICE SUPPORT IN THE DIFFERENT UMTS RELEASES

2.1. Release 3

In Release 3 the Quality of service profile is backward compatible with GPRS for GSM. One important addition to the 2G GPRS specification is the possibility to setup more than one PDP context per user (IP address), with each PDP context corresponding to a different QoS profile. The TFT (Traffic Flow Template) in the GGSN contains the necessary filters to send the data packets over the correct PDP context.

Upon reception of a PDP context activation message from the UE, the SGSN checks the requested QoS with the GPRS subscription profile and can accordingly downgrade the requested QoS. The (modified) PDP context activation message is forwarded by the SGSN to the GGSN. The GGSN has the possibility to downgrade the requested QoS based on the available resources. The (modified) PDP context activation message is sent back by the GGSN to the SGSN. The SGSN asks the RNC to setup a RAB (Radio access bearer) towards the UE with the requested QoS. In Rel-3 the UTRAN cannot modify the QoS specified by the SGSN in the RAB signalling. Rel-3 also does not allow the UTRAN to renegotiate RAB QoS parameters during the session or at handover.

2.2. Release 4

Asymmetric and uni-directional PDP contexts will be supported from *Rel-4* onwards, which allows to separately set the QoS profile for the up- and downlink connection. This feature is very useful for streaming applications where information with totally different QoS requirements is flowing in the two directions (downlink = data, uplink = control).

Release 4 will also introduce RAB QoS negotiation and renegotiation over the Iu interface.

2.3. Release 5

Release 5 marks the introduction of IP Multimedia services. Both IMS signalling (SIP) and user data (audio/video media flows) are transported over the same GPRS bearer network. Nevertheless it should be possible for the network to distinguish a signalling packet from a user data packet. This allows operators to handle those packets differently with respect to traffic handling priority (in case of congestion increase priority of signalling packets) and charging (no charging for signalling). The two solutions currently discussed in 3GPP are defining a separate flag or a separate QoS class for IMS signalling.

Another important novelty in R5 is the linking between the session layer (SIP) and the GPRS bearer layer. For this purpose an additional policy control interface - Go - is standardised [8]. The GGSN contains a Policy Enforcement Function (PEF) that has the capability of policing packet flow into the IP network, and restricting the set of IP destinations that may be reached from/through a PDP context according to a packet classifier. Policy control is performed by a Policy Control Function (PCF), which in R5 is considered a logical entity of the P-CSCF.

Thanks to the Go interface between PCF and GGSN an operator is able to enforce policy on the PDP contexts (see figure 3). In previous releases the GGSN could downgrade the QoS requested in PDP context activation only based upon available resources. Section 5.2 explains how in R5 the P-CSCF/PCF can give via Go policy instructions to the GGSN, which enables the GGSN to verify that the requested PDP QoS does not exceed the QoS authorised by the P-CSCF.

3. RADIO OPTIMISATION TECHNIQUES

In the UTRAN there is a need to optimise the use of the radio resources and the voice quality for VoIP. ROHC (RObust Header Compression) is a compression algorithm used to avoid that a single packet loss results in loss of multiple voice frames [9]. ROHC provides an efficient and fault-tolerant compression suite, suitable for both IPv4 and IPv6, that can be used to compress IP/UDP/RTP, TCP, SIP, ... ROHC is implemented in the GPRS PDCP layer and runs between RNC and UE. Furthermore Unequal Error Protection (UEP) is used to protect the bits of a voice codec according to their importance to reconstruct the audio signal. UEP is currently not yet supported on the Iu-ps interface. It will be a feature for R6.

4. END-TO-END QOS SCENARIOS

Originally 3GPP identified 6 different end-to-end QoS scenarios [8]. The scenarios described the different options to specify the QoS level for the end-to-end bearer establishment. Which scenario would apply depends mainly on terminal capabilities, the operator's choice to perform QoS policy control or not and the QoS mechanisms supported by the IP backbone interconnecting both access networks.

It was envisaged that some low-end terminals wouldn't implement any of the existing IETF QoS protocols like RVSP or DiffServ. In this case the QoS for the GPRS network is specified using the GPRS PDP mechanisms and relying on the GGSN to do the correct mapping to DiffServ for Gi interface. The other extreme is the scenario where typically you would have a laptop using RSVP connected to a mobile handset. RSVP is used to signal end-to-end the required QoS and is sent over the radio interface.

Alcatel from the beginning recognised the standardisation overhead of 6 scenarios. Recently 3GPP limited the number of scenarios. General consensus was reached to support the scenarios where the GGSN has DiffServ edge capabilities and can remark the IP packets coming from the UE (with and without policy control in the GGSN). The end-to-end RSVP scenario is still under discussion in 3GPP but it is likely to disappear. From the technical point RSVP would introduce an overhead on the radio interface, also RSVP raises scalability problems when used in backbones both in terms of bandwidth and processing power. The operational argument is that the current IP backbones do not support end-to-end RSVP and probably never will.

5. QoS FOR UMTS MULTIMEDIA: A BEARER AS WELL AS SESSION LAYER ISSUE

Providing QoS to IMS services is not just a bearer level (GPRS) issue. Not only is there a need for involving the session layer (SIP/SDP) but also for co-ordinating bearer and session layer QoS. Section 5.1 explains the general steps to establish a multimedia session between two UMTS R5 IMS users, focussing on QoS related events linking session and bearer level. Section 5.2 describes the policy control mechanism and authorisation token used to synchronise QoS at session and bearer level. Finally, section 5.3 proposes QoS extensions to the session layer protocol SDP [2] that allow an end-user to express and negotiate with the network the QoS 'as he wants to perceive it'.

5.1. Establishing a UMTS multimedia session

The UMTS R5 architecture is a layered architecture with a clean split between bearer (e.g. SGSN, GGSN), session (e.g. P-CSCF, S-CSCF) and service level. The way to set up an IMS session [5] is based on the principles of *interleaving session and bearer level signalling* proposed in [3]. Figure 3 shows the call flow.

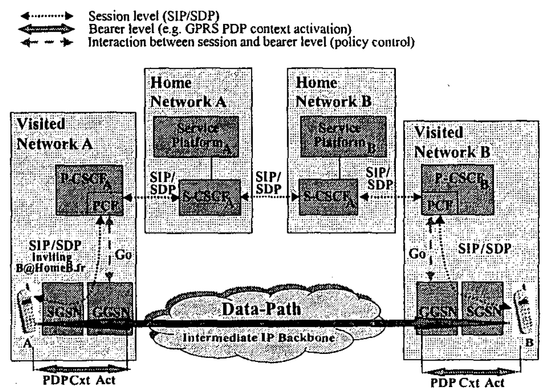


Figure 3, Bearer-Session QoS interactions, session setup

Establishing an IMS session involves three steps:

1. <u>Session Initiation: Negotiation of the media components between calling and called party.</u>

In this first step of session initiation, the two end-users negotiate the characteristics of the media components (e.g. codecs) they would like to use in the session. End-users communicate their preferences (e.g. list of preferred codecs) to each other by attaching SDP fields to the sequence of session initiation messages. Section 5.3 proposes new SDP extensions that allow end-users to not only negotiate codec lists but also the Quality of Service characteristics of each media component. Although the negotiation happens basically end-to-end between the two end-users, S-CSCF and P-CSCF are allowed to 'screen' the negotiated media characteristics. The S-CSCF in each end-user's home network checks for compliance with the end-user's multimedia subscription profile. If P-CSCF finds the result of the SDP negotiation compliant with local visited network policies, it informs the GGSN, via the Go interface as explained in section 5.2, for each media component of the 'QoS authorised by the session layer'.

2. Resource Reservation: on the UMTS access networks and on interconnecting backbone(s).

After the two end-users have agreed at session level on the media characteristics to be used for the session, the resources for the media flows can be reserved at bearer level. On the UMTS network, resource reservation means PDP context activation. The UE has to derive the QoS parameters to be put in the PDP context activation from the media characteristics negotiated via SIP/SDP in step 1. As explained in section 2.1 SGSN, GGSN and RNC can modify/downgrade the QoS requested by the end-user in the PDP context activation. The SGSN checks whether the QoS requested is consistent with the GPRS subscription profile. The role of the GGSN is important with respect to session-bearer level co-ordination. Based on the policy instructions received from P-CSCF, the GGSN can reject/downgrade the QoS requested via PDP context activation, and the policy enforcement point in GGSN can furthermore discard/mark the associated user plane traffic. If calling and called party are not located in the same UMTS access network, resources also have to be reserved in the external backbone connecting the two access networks. This is based on one of the mechanisms explained in section 4. Note that although the result of this step is that the resources for the media flows are reserved, it is not allowed to actually start transmitting user data until charging commences, after Step 3.

3. Session Completion: resource reservation OK, ringing and opening of the gates.

First the calling party notifies the called party that resources are reserved. The called party is rung and calling party is informed. When the called party hooks-off, the final session establishment is carried out. During this process, the P-CSCF instructs the GGSN via the Go interface to let the user plane traffic flow through and to start the charging of the session.

5.2. Policy control & authorisation token for bearer-session level synchronisation

In order to offer "carrier grade" IMS services, an operator should be able to correlate the QoS requested by the user at the session layer with QoS actually provided by the network at the bearer level. Although the use of policy control is optional for R5, Alcatel believes that it is of capital importance for fraud prevention and theft of service.

The binding between the media flows specified in the SIP signalling and the corresponding PDP contexts at the GGSN is ensured by using an authorisation token as defined by [4]. At session setup the P-CSCF inserts the authorisation token in the first SIP message towards the end-user. Via the Go policy control interface, the P-CSCF/PCF informs the GGSN for every media component about the 'QoS authorised by the session layer' together with the 'authorisation token and corresponding sequence number'. The MT insert the authorisation token together with the sequence number in the IP specific field of the PDP context activation message. At reception of the PDP activation message, the GGSN verifies if the requested PDP QoS does not exceed the QoS authorised by the P-CSCF.

5.3. SDP extensions for end-to-end QoS negotiation

Current SDP doesn't carry enough information to allow the P-CSCF to extract all the QoS parameters per media flow needed for policy control. The only information carried in SDP from which QoS parameters can be derived is the codec and an optional B (bandwidth) parameter. Based on current SDP specification the only way to ensure coherency between all session control elements is to standardise a one-to-one mapping between codec and QoS parameters.

The final goal of the QoS provisioning architecture however should be to deliver the user a QoS for each media component that corresponds exactly to the level of "quality" he wishes to experience. An end-user should have the possibility to request a different "quality" for a media component based on for instance the destination, expected duration of the session, pricing, etc. The best quality could be preferred for an important business session, while the lowest could be chosen when the caller knows by advance that the call will be quite long, or not very important, or if the callee is far away, etc. The flexibility to let customers request 'the quality as they want to perceive it' creates opportunities for differentiation between operators.

The Quality is requested by the end-user at session setup time and negotiated end-to-end at the signalling level. The SDP QoS extensions Alcatel proposes [10] are split into two categories: the Traffic Information (TI) and the Sensitivity Information (SI).

1) The 'Traffic Information' characterises the traffic type of the bearer associated with the media component. If there is more than one codec per media component the traffic information can be specified per codec. Typically TI is related to bandwidth and packet size. There are situations where some media streams are not associated with codecs (e.g. white board) or situations where the intermediate session control entities do not have any a priori knowledge of the codec being used (e.g. use of new codecs). Carrying TI per media stream via SDP still provides intermediate SIP proxies with knowledge/control on the bearer requirement of the session being established. It relieves the requirement for all involved SIP proxies to know the mapping from a codec to the bearer level requirement (TI) of this codec. This facilitates the introduction of new codec types.

2) The 'Sensitivity Information' determines the tolerance/sensitivity of a certain media stream specified by the traffic information. The SI information directly influences the end-users perceived Quality level. Typically SI information is characterised by maximum end-to-end delay and delay variation and maximum packet loss. For a given bearer defined by the TI, the SI unambiguously determines the quality 'as the end-user wants to perceive it'. A user can request per TI more than one acceptable SI level, which are specified in decreasing order of preference in SDP to allow prioritisation during negotiation.

TI is always represented in SDP under the form of parameters with their respective values. For SI this is different. Three possible representation formats can be used to unambiguously describe a given SI level for a given media stream in a given session:

1) Standard QoS parameter format: set of well known parameters allowing to precisely describe for a given medium stream the SI requirement. The standard QoS parameters are maximum packet loss ratio, maximum end-to-end delay and maximum end-to-end delay variation.

2) Standard QoS Class format: is an abbreviated representation of a standardised set of standard QoS parameters with well-defined values. In ETSI TIPHON different classes and their mapping to QoS parameter values for voice have already been standardised; best, high, medium, low and acceptable.

3) QoS flavour format: is an abbreviated way of representing operator specific (non-standard) QoS information describing the SI requirement. The usage of the QoS flavour is always assuming a pre-defined and well-understood interpretation of the QoS information over the considered interface. Examples are: Gold, Silver, Bronze, ... type of QoS.

The real added value of the SDP extensions proposed by Alcatel can best be explained from an end-user perspective by looking at their impact on the call flow in Figure 3. When setting up a session the A party can request a level of QoS as she wants to perceive it, using the TI and SI information as explained above. As in current SIP/SDP for each medium a list of codecs preferred by the A party is defined. New is that for each codec also a TI and an ordered list of SI values (e.g. Silver, Bronze) is carried in the SDP of the SIP-Invite message. This allows A and B party to negotiate for each medium in the session not only about the codecs but also about the QoS levels they would prefer. In that negotiation process also A and B's P-CSCFs and S-CSCFs can modify the requested QoS. The P-CSCFs at both sides will check the QoS with the local network policies. The S-CSCF can reduce the requested QoS level based on the end-user's subscription profile, the specificity of the service, the time of day/week, etc. If the A party used the QoS flavour form (Silver, Bronze), the S-CSCF of A will have to translate the Gold and Silver QoS flavour in the SIP-Invite towards the S-CSCF of B to one of the two remaining standardised forms. After all the CSCFs checked and/or modified the QoS requested by the A party, the B party receives the SIP-Invite message. B is the last one to modify the list of codecs and QoS levels. When receiving the final SDP message, the P-CSCFs ensure the coherence of the negotiated QoS levels at session and bearer level through the policy interface with the GGSN as explained in section 5.2.

Suppose B is roaming and the call arrives during peak hours. As this type of calls is normally quite expensive, normally B would decide not to take that call. Thanks to the SDP extensions B can choose to accept the call at a less expensive – lower Quality- level. The opposite as well can happen. Maybe B is abroad for business reasons using a company phone and A is an important customer calling him with an urgent question about an important business deal. In that case the SDP extensions even allow B to propose A to upgrade the Quality of the call - on B's expense of course - to for example Gold quality. In both cases the operator clearly makes more money thanks to the SDP extensions for end-to-end service level QoS negotiation.

6. CONCLUSION

With the introduction of IP multimedia services in UMTS on a large scale, new challenging demands are imposed on the QoS infrastructure. Quality of Service is provided in the IP based UMTS core network via the support of Diffserv -and MPLS- in GGSN, SGSN, RNC. Furthermore several enhancements are needed to the way Quality of Service can be negotiated via GPRS (asymmetric PDP contexts, QoS re-negotiation over Iu interface, differentiation of IMS signaling versus user data traffic, etc). At the border of the UMTS network, the GGSN has to be able to perform the mappings between UMTS QoS parameters and IP QoS parameters supported by all possible neighboring IP networks.

However, the main message to remember is that providing 'carrier grade' IP multimedia services is not just a transport level issue. Not only is there a need for involving the session layer but also for co-ordination between transport (GPRS) and session (SIP/SDP) layer QoS. Policy control mechanisms are introduced in P-CSCF/PCF and GGSN to synchronise the QoS requested by the end-user at transport level with the QoS authorised at session level. Finally, Alcatel proposes QoS extensions to SDP that allow an end-user to express and negotiate end-to-end at session level the QoS 'as he wants to perceive it'.

7. REFERENCES

[1] J. Rosenberg et al, "SIP, Session Initiation Protocol", draft-ietf-sip-rfc2543bis-05.txt, Work in Progress.

[2] M. Handley et al, "SDP: Session Description Protocol", draft-ietf-mmusic-sdp-new-03.txt, Work in progress.

[3] W. Marshall et al. "Integration of Resource Management and SIP", draft-ietf-sip-manyfolks-resource-02.txt, Work in progress.

[4] W. Marshall et al. "SIP Extensions for Media Authorization", draft-ietf-sip-call-auth-02.txt, Work in progress.

[5] 3GPP TS 23.228 "IP Multimedia (IM) Subsystem (Stage 2) - Release 5".

[6] 3GPP TS 23.060: "General Packet Radio Service (GPRS); Service Description; Stage 2".

[7] 3GPP TS 23.107: "QoS concept and architecture".

[8] 3GPP 23.207: "End-to-end QoS concept and architecture".

[9] Bormann C. et al, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed", IETF RFC 3095

[10] Bos L. et all, "A Framework for End-to-End User Perceived Quality of Service Negotiation", draft-bos-mmusic-sdpqos-framework-00.txt, Work in Progress.

[11] Baudet S. et al, "QoS implementation in UMTS networks", Alcatel Telecom Review, 1st Quarter 2001