SUPPORT OF IP QOS OVER UMTS NETWORKS

M. Ricardo[1,2], J. Dias[3,2], G. Carneiro[2], J. Ruela[1,2]

[1] FEUP - Fac. Eng. Univ. Porto, Rua Dr. Roberto Frias, 4200 Porto, Portugal
[2] INESC Porto, Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200 Porto, Portugal
[3] Instituto Superior de Engenharia do Porto, Rua Dr. Bernardino de Almeida, 4200 Porto, Portugal
{mricardo, jdias, gjc, jruela}@inescporto.pt

Abstract – The paper presents an end-to-end Quality of Service (QoS) architecture suitable for IP communications scenarios that include UMTS access networks. The rationale for the architecture is justified and its main features are described, notably the QoS management functions on the terminal equipment, the mapping between IP and UMTS QoS parameters and the negotiation of these parameters.

Keywords – IP, UMTS, QoS, PDP Context, RSVP, IntServ.

I. INTRODUCTION

The paper describes the architecture used in the IST project ARROWS (Advanced Radio Resource Management for Wireless Services) [1] for supporting IP Quality of Service over UMTS Networks.

ARROWS aims at providing Radio Resource Management (RRM) and Quality of Service (QoS) management solutions for the support of integrated services in the Universal Terrestrial Radio Access (UTRA). The project addresses packet access, asymmetrical traffic and multimedia services, all based on IP. A multimedia testbed, developed to validate the proposed concepts, supports four basic applications: 1) audio-video telephony, based on VIC (video) and RAT (audio) conferencing tools, 2) video streaming, based on the applications developed by the MPEG4IP group, 3) Web, based on *Mozilla* client and *Apache* server and 4) mail, based on *Mozilla* client and *sendmail*/*pop3d*/*imapd* servers for STMP/POP3/IMAP services, respectively.

Although ARROWS concentrates on the QoS aspects of UTRA, a global QoS framework is necessary for two reasons. First, from the applications point a view, QoS is an end-to-end issue [2] and, second, a mapping between UMTS and end-to-end QoS parameters is required.

II. QOS IN PACKET SWITCHED NETWORKS

The provisioning of adequate end-to-end transport services in an all-IP-based communications scenario where the access network is UMTS, the core networks may or may not provide QoS support and applications require differentiated transport services is a complex and challenging problem. Applications have different transport requirements while networks have also their own QoS characteristics. A review of these concepts is necessary in order to justify the rationale for the QoS architecture and strategy proposed.

A. The applications point of view

In packet networks, from a QoS point of view, applications can be classified according to two commonly accepted taxonomies. The first comes from the IP world, where applications are first classified as elastic or real-time.

An alternative taxonomy, proposed by 3GPP, classifies applications according to the characteristics of the traffic they generate: conversational, streaming, interactive and background [3]. A conversational class preserves the time relation between the information entities of the stream (the delay jitter must be tightly controlled), the end-to-end delay is required to be low and the traffic is usually almost symmetric. A typical example is audio conversation, but audio-video telephony is likely to become very common, as well. A streaming class is used for transferring data so that it can be processed as a steady continuous flow. Streaming applications, like video streaming, are usually highly asymmetric and end-to-end delay and jitter need not to be as tightly controlled as in the conversational class. The interactive class includes applications like web browsing, where traffic is asymmetric and round trip time as well as bit error ratio should be kept low. The background class serves applications like email, where delay is not an important issue.

The set of services selected for the ARROWS testbed is fully representative of UMTS traffic classes.

B. IP QoS

The transport of IP packets through a set of networks where at least one of them is UMTS, as represented in Fig. 1, can be in the first place seen from an IP QoS point of view. In this context, the network elements dealing with IP - terminal equipment, GGSN (Gateway GPRS Support Node) and server - need to cooperate and manage their packet queues, so that the packets generated by the applications can be transported with some predictable and differentiated guarantees. Two well-known approaches for dealing with IP QoS [4] are the differentiated services (DiffServ) model and the integrated services (IntServ) model.

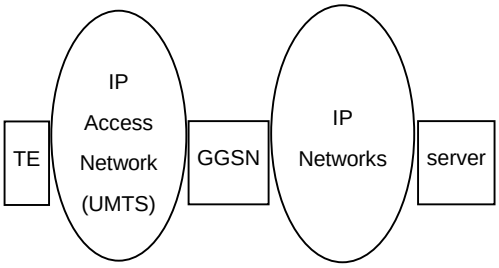

Fig. 1. Reference Architecture

Among others, the IntServ model supports a Guaranteed Service class aimed at providing packets with a maximum transfer delay between source and destination hosts. When a client process in the terminal needs to send a packet flow to a process in the server and these packets require a known maximum delay, resources must be reserved at every IP network element they traverse.

The characteristics of the flow are described by four traffic parameters: p (the peak rate), b (bucket size), r (token bucket rate) and M (maximum datagram size). The flow is regulated by the double token bucket shown in Fig. 2.

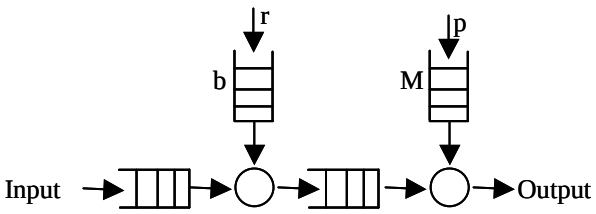

Fig. 2. Traffic regulator

The reservation characteristics are defined by two parameters: R, the reserved bandwidth and S, the slack term. The bound on the end-to-end delay depends on the reserved bandwidth, the traffic parameters and two terms (C_{tot} and D_{tot}) that are the summations of two error terms (C_i and D_i) accounted for in each router along the path.

The first network element in the path (e.g., the terminal equipment) estimates the time a packet waits to be served in its queue (related to C_i) plus the time it takes to be transmitted and propagated through the output link to the next router (D_i). This information is sent to the next router in a PATH (b, r, p, M, C_{tot}, D_{tot}) message.

Each router, when receiving this message, recalculates C_{tot} and D_{tot}, by adding its own C_i and D_i and forwards the PATH message to the next router. The last IP network element (e.g., the server), based on the maximum delay (D_{max}) tolerated by the application receiving the flow and on C_{tot} and D_{tot}, estimates the bandwidth R that must be reserved in the network.

For constant bit rate services, assuming that R is at least equal to the guaranteed rate (r = p), the bound on the end-to-end delay is given by $D_{EE} = (M + C_{tot}) / R + D_{tot}$ and $D_{EE} < D_{max}$ should be observed. Moreover, the receiver of the flow also calculates the slack term $S = D_{max} - D_{EE}$, and sends the RESV (R,S) message to the routers along the reverse path requesting them to reserve R for this flow. The process is summarised in Fig. 3.

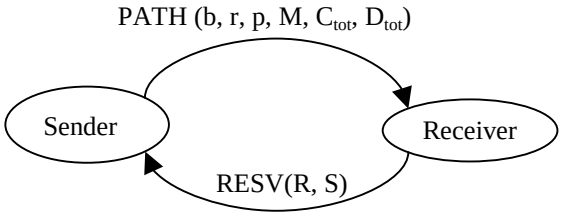

Fig. 3. RSVP negotiation

C. UMTS QoS

From the IP point of view, the UMTS access network is perceived as a link layer technology. When the terminal attaches itself to the network, a tunnel is established between the terminal and the GGSN. This tunnel, known in UMTS as PDP context [5], provides the UMTS Bearer Service. The terminal equipment endpoint of this tunnel moves with the terminal and new tunnels (Secondary PDP contexts) may be dynamically activated. In this case, packet filters, known as traffic flow templates (TFT), are associated with each Secondary PDP context.

QoS is one of the most important attributes of a PDP context and is negotiated during PDP context activation. Besides indicating the traffic class for which the context is optimised (conversational, streaming, interactive or background), a number of other attributes must be negotiated. Common to all traffic classes are the maximum bitrate, delivery order, maximum SDU (Service Data Unit) size, SDU error ratio, residual bit error ratio, delivery of erroneous SDUs and allocation/retention priority. Common to the conversational and streaming classes are the SDU format information, the transfer delay and the guaranteed bit rate. In the interactive class, the traffic handling priority can be negotiated, as well. For each QoS attribute some ranges of values are predefined in the standards.

III. PROPOSED QOS SOLUTION

ARROWS aims at defining RRM algorithms targeted at the UTRA portion of UMTS. Therefore, and to avoid other networks degrading the transport service, the IntServ guaranteed service class was selected as the IP QoS model for the testbed. Nevertheless, the proposed architecture can be easily adapted to other scenarios, such as the IP access network supporting IntServ and the IP core networks supporting DiffServ.

A. QoS architecture

The QoS architecture implemented in the ARROWS testbed is represented in Fig. 4. It is in line with Fig. 1 and all IP network elements (terminal equipment, GGSN and server) support IntServ and RSVP (Reservation Protocol) [6].

The six functional blocks represented in Fig. 4 are (1) Application, (2) Non-Access Stratum, (3) RSVP, (4) QoS Manager (5) IP BS Manager and (6) IP Traffic Control. The lines connecting them represent paths through which information is conveyed. The bold lines, such as that connecting Application to Traffic Control, represent data paths. The other lines are used for signaling. Functional blocks are aggregated in different ways in TE, GGSN and Server network elements.

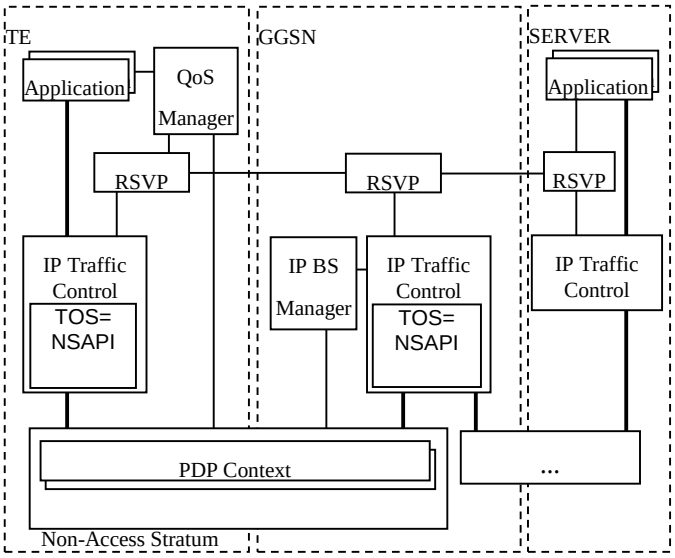

Fig. 4. Proposed QoS architecture

Application. The application block represents a client/server process that communicates through UDP or TCP sockets and usually provides a well-known service. Services like WWW and mail use TCP, while the unicast audio-video telephony as well as the video streaming services use UDP. These UDP services are real-time, use RTP and adopt the IETF multimedia service architecture. Applications requiring QoS must be able to communicate with RSVP. In ARROWS, however, a library was developed (included in the TE QoS Manager) which allows non-RSVP capable applications to be used with minor modifications.

Non-Access Stratum. This block, common to the edge UMTS network elements (terminal equipment and GGSN), represents the UMTS Non-Access Stratum, as defined in 3GPP standards. A block communicating with the Non-Access Stratum can activate, deactivate, modify or use (for packet transport) the PDP Contexts. Although opening a PDP Context may be requested from GGSN, in practice only the paying terminal can activate PDP Contexts. Contexts have a QoS associated.

RSVP. This block, which represents the standard RSVP daemon, is responsible for handling IP reservations on an IP flow basis. It offers an API for the applications, that is, the sender and the receiver of Fig. 3. In TE, this API is offered to the QoS Manager. The last interface alows RSVP to configure IP Traffic Control.

QoS Manager. This is the key block of the architecture. Besides allowing non-RSVP capable applications to use RSVP, the QoS Manager block is also in charge of (1) managing IP reservations, that is, selecting R and S values carried on RESV messages, (2) activating/deactivating PDP contexts with QoS attributes, (3) mapping RSVP QoS parameters into UMTS QoS parameters and (4) deciding the multiplexing of IP flows into PDP contexts.

IP BS Manager. It exists in the GGSN and is responsible for requesting the activation of the Primary PDP Context (if not yet activated) upon the arrival of the first datagram to be delivered to the terminal equipment.

IP Traffic Control. After a PDP context has been activated, the terminal may start transmitting datagrams. It is, however, necessary to direct the packets to the proper PDP Context, schedule the packets according to their priorities, and shape the traffic so that the flow is compliant with the QoS previously negotiated for that PDP Context. A typical traffic control configuration is shown in Fig. 5.

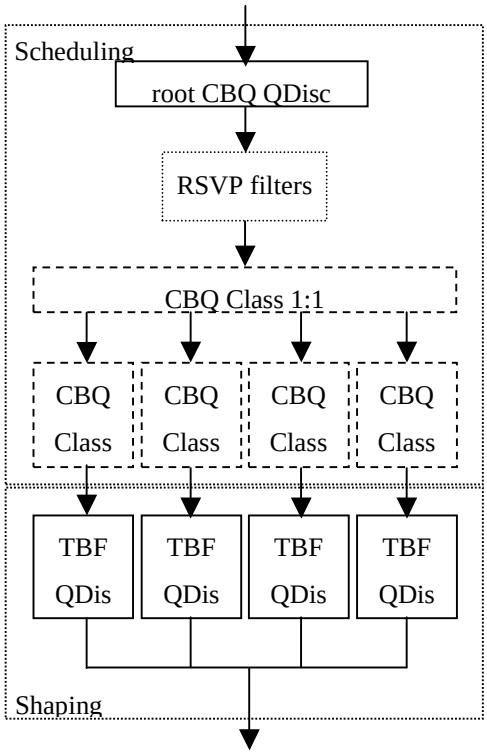

Fig. 5. IP Traffic Control

Scheduling and shaping of the flows are implemented with CBQ (Class Based Queueing) [7] and TBF (Token Bucket Flow) [8] disciplines, respectively. The location of the filters is also shown. For each PDP Context, one leaf class on the CBQ queuing discipline (e.g. 1:11) is created. Moreover,

each class has one TBF queuing discipline associated instead of the generic one installed by default.

Once the packet is sent to the Non-Access Stratum how is it known the PDP Context the packet belongs to? The solution adopted was to use the Type of Service (ToS) field of the IP datagram. The TFT associated with a Secondary PDP Context can be given a list of ToS. This solution, although more arduous to implement than to simply give the standard IP five tuple to a TFT, eases the migration of the QoS architecture towards, for instance, DiffServ.

B. IP and UMTS Bearers

In order to analyze the QoS mappings selected, the traffic generated by the applications must first be characterized. The results presented below are also summarized in Table 1.

Audio-video telephony. Two IP flows are generated in each direction (downlink and uplink), one for audio and the other for video. Audio is constant bit rate with r = p = 16 kbit/s. Video is also constant bit rate with r = p = 48 kbit/s. For each direction, the two IP flows are multiplexed (priority is given to audio), in a conversational Secondary PDP Context with both Maximum and Guaranteed bit rate of 64 kbit/s and a maximum SDU size of 570 octets, assuming compressed RTP/UDP/IP headers.

Video Streaming. Two IP flows, which correspond to base and enhancement layers, respectively, are generated in the downlink direction, from the server to the terminal. Both flows are constant bit rate. The base layer is characterized by r = p = 32 kbit/s and the bit rate of the enhancement layer (r = p) is negotiated from 0 to 96 kbit/s in steps of 16 kbit/s. These IP flows are, in turn, mapped into streaming Secondary PDP Contexts. For the former, the Guaranteed and Maximum bit rate are assumed to be 32 kbit/s. For the latter, the Maximum bit rate is requested to be equal to the negotiated value and the Guaranteed bit rate is requested to be 0 bit/s. The Maximum SDU size is 570 octets for both contexts, and RTP/UDP/IP header compression is applied.

Web Browsing. This service is modeled as two IP flows. The values selected for uplink and downlink directions were r = p = 64 kbit/s and r = p = 256 kbit/s, respectively. The support of these flows is achieved through a bidirectional interactive Secondary PDP Context. The Maximum bit rate was set to 64/256 kbit/s. The Guaranteed bit rate, as previously mentioned, is not applicable to this traffic class. A Maximum SDU size of 570 octets was selected.

Mail. This service is modeled as two IP flows, as well. The same bitrate value was selected both for the uplink and the downlink (r = p = 32 kbit/s). The support of these flows is achieved through a bidirectional background Secondary PDP Context, and the Maximum bit rate is 32 kbit/s in both directions. The Guaranteed bit rate is not applicable to this class and the Maximum SDU size was set to 570 octets.

Signaling. Besides the Secondary PDP Contexts, a Primary PDP Context is also established and is used to transport other IP datagrams than those generated by the above applications. Examples are packets carrying out the RSVP messages or the support of services such as *ping* or *telnet.* For that purpose, the primary PDP context is selected to be background, thus bidirectional, with a Maximum bit rate of 32 kbit/s. Unlike the Secondary PDP contexts whose activation and deactivation are triggered by the applications, the Primary PDP context does not depend on any application.

Table 1

IP flow and PDP Context characteristics in the ARROWS testbed

End-to-End Service	IP Bearer	UMTS Bearer
Videoconference	*RTP/UDP/IP* **UL:** Audio:16kb/s/ Video:48kb/s **DL:** Audio:16kb/s/ Video:48kb/s	**Sec. PDP Context:** *Conversational* UL: 64kb/s **Sec. PDP Context:** *Conversational* UL: 64kb/s
Video Streaming	*RTP/UDP/IP* **DL:** Base Layer: 32kb/s Enh. Layer: 0 to 96kb/s (16kb/s steps)	**Sec. PDP Context:** *Streaming* DL: Base L.: 32kb/s **Sec. PDP Context:** *Streaming* DL: Enh L.: 0 to 96kb/s (16kb/s)
Web Browsing	*TCP/IP* **UL:** 64kb/s **DL:** 256kb/s	**Sec. PDP Context:** *Interactive* *UL: 64kb/s* *DL: 256kb/s*
E-mail	*TCP/IP* **UL:** 32kb/s **DL:** 32kb/s	**Sec. PDP Context:** *Background* *UL: 32kb/s* *DL: 32kb/s*
Signaling Bearer	**UL:** 32kb/s **DL:** 32kb/s	**Sec. PDP Context:** *Background* *UL: 32kb/s* *DL: 32kb/s*

C. RSVP negotiation

For real-time services (audio-video telephony and video steaming) the receiver of the flow can select for R a value that must be at least equal to r = p. The other variable that can be selected is S, the Slack term. In the proposed solution S is used for negotiating the UMTS transfer delay attribute. The cost of a PDP context is inversely proportional to its transfer delay, since the UTRAN gains in flexibility for scheduling packets.

An RSVP negotiation example that takes into account this aspect is shown in Fig. 6.

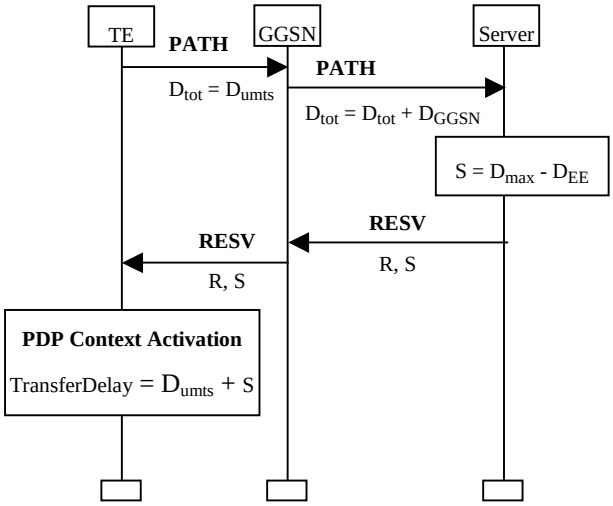

Fig. 6. RSVP negotiation in the proposed architecture

The PATH message sent by TE contains as D_{tot} the typical delay felt by a packet when sent through the Primary PDP Context. On the reception of the RESV message, the available slack delay S is used for negotiating the new Secondary PDP context that will accommodate the IP flow.

Fig. 7 shows the signaling process associated with a video streaming session where two IP flows and two Secondary PDP Contexts need to be established.

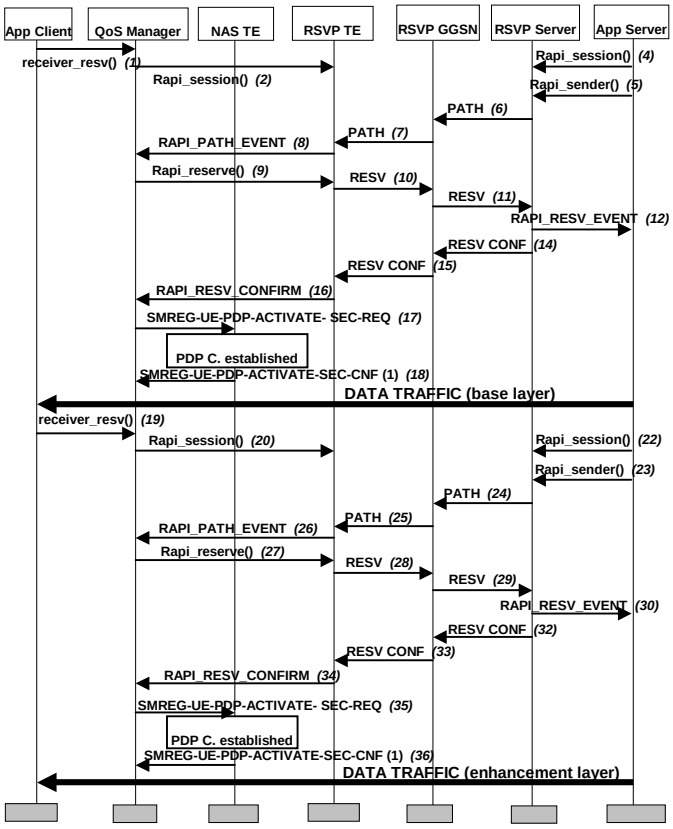

Fig. 7. RSVP and PDP Context signalling process

IV. CONCLUSIONS

The main contribution of this paper is a QoS architecture that is capable of providing IP based real-time applications with guaranteed services over UMTS access networks. To fulfil this goal it includes three important attributes.

First, a QoS Manager, which consists of a set of functions that ease QoS management by the terminal equipment. Besides allowing non-RSVP capable applications to work with RSVP, the QoS Manager is also responsible for managing IP reservations, activating/deactivating PDP contexts with QoS attributes, mapping RSVP QoS parameters into UMTS QoS parameters and deciding about multiplexing of IP flows into PDP contexts.

Secondly, the QoS mapping between IP and UMTS bearer services proposed for the relevant services. For this purpose, it was necessary to characterize IP traffic flows, to select UMTS bearers and to investigate multiplexing strategies.

Thirdly, the mechanism used by RSVP to negotiate R and S parameters. Although the PATH message is sent through the Primary PDP context, this mechanism allows the adequate use of Secondary PDP contexts and minimizes the cost of the link by maximizing the transfer delay.

ACKNOWLEDGMENTS

The authors wish to thank the support given by the IST European research programme and their partners within the ARROWS project: Universitat Politecnica de Catalunya, University of Limerick, Telefónica I+D and Telecom Italia Lab.

REFERENCES

[1] IST ARROWS project, http://www.arrows-ist.upc.es.

[2] 3GPP TS 23.207 v5.3.0, “End-to-End QoS Concept and Architecture”, March 2002.

[3] 3GPP TS 23.107 v3.8.0, “QoS Concept and Architecture”, March 2002.

[4] Xipeng Xiao, Lionel Ni, “Internet QoS: A Big Picture”, *IEEE Network Magazine*, Vol. 13, No. 2, pp. 8-18, March/April 1999.

[5] 3GPP TS 23.060 V3.11.0, “General Packet Radio Service (GPRS); Service description; Stage 2”, March 2002.

[6] Paul White, “RSVP and Integrated Services in the Internet: A Tutorial”, *IEEE Comm. Magazine*, Vol. 35, No. 5, pp. 100-106, May 1997.

[7] Sally Floyd, "Notes on CBQ and Guaranteed Service", July1995.

[8] K. Dovrolis, M. Vedarn, P. Ramanathan, “The Selection of the Token Bucket Parameters in the IETF Guaranteed Service Class”, Technical Report, Department of ECE, University of Wisconsin-Madison, November 1997.