An End-to-End QoS Framework for Multimedia Streaming Services in 3G Networks

H. Montes, G. Gómez and D. Fernández
Nokia Networks, IP Mobility Networks, NSR-Málaga SCT, P.T.A, Málaga (Spain)
E-mail: {ext-hector.montes, ext-gerardo.gomez, ext-daniel.fernandez}@nokia.com

I. Introduction

Multimedia streaming services are receiving considerable interest in the mobile network business [1]. Supporting reliable real time services is a decisive aspect for the increasing migration towards packet based telephony networks. For UMTS, deploying an all-IP architecture is a promising standardization trend due to the convergence between IP technologies and telephony services [2]. Streaming services are also technically supported over evolving second (2G) and third generation (3G) wireless networks, thus streaming clients will soon be deployed in advanced wireless communication devices.

Inside this new group of services, there exist a variety of applications (e.g. audio and video on demand) with different traffic source statistical characteristics [3]. In case of audio streaming, the generated traffic is rather non-bursty whereas video traffic has a more bursty nature. One key issue is how mobile networks can support this kind of services. In these "Pre-All-IP" service cases the used radio bearers can be chosen from either 2G or 3G circuit switched (CS) or packet switched (PS) bearer set. PS bearers provide more trunking gain and better resources utilization while CS bearers offer better performance for those services with stringent delay requirements. All the multimedia services are mainly characterized by the necessity from the network point of view to guarantee certain Quality of Service (QoS) requirements.

Providing end-to-end QoS for multimedia streaming services implies the harmonized interworking between protocols and mechanisms specified by IETF and 3GPP [4] and involved in QoS provisioning within the different 3G network subdomains and the external IP-Packet Data Networks (IP-PDN) through which the user accesses to the service.

In this paper, the end-to-end QoS management of streaming services in 3G mobile networks is considered. Particularly, the possibility of employing a Public Land Mobile Network (PLMN) hosted multimedia streaming service is studied to avoid accessing through an external IP-PDN to streaming services. By this solution, the mobile operator hosts a streaming server or a proxy server within the PLMN, allowing the provision of sufficient QoS to users of wireless streaming terminals. The presented analysis of the multimedia streaming session is chronologically divided in two phases: service activation and service utilization.

The transmission of multimedia services with stringent QoS requirements implies a conversion of the existing GSM/EDGE (Enhanced Data rates for GSM Evolution) network into real 3G networks [5]. This conversion entails the enhancement of the radio interface, with the necessity of a complete Radio Resource Management (RRM) functionality, aware of the QoS requirements of the new aforementioned services.

II. Overall scenario description

A. *Description of the service: Multimedia Streaming*

A generic framework for a typical multimedia streaming service consists of content creation and retrieval system. When providing a streaming service, a media server opens a connection to the client terminal and begins to stream the media to the client at approximately the playout rate. During the media receiving, the client plays the media with a small delay or no delay at all. This technique does not only free up limited terminal memory, but also it allows to the media to be sent live to clients as the media event happens. The user needs a player, which is a special program that decompresses and sends video data to the display and audio data to the speakers. This client application must be able to control the streaming flows (control plane) and manage the media flows (user plane). Moreover, the client also has to interface with the underlying transport network technology, its specific protocols and data bearers dedicated to the service.

The 3GPP PS multimedia streaming service is being standardized based on control and transport IETF protocols as Real-Time Streaming Protocol (RTSP), Real-Time Transport Protocol (RTP) and Session Description Protocol (SDP), as fig. 1 shows. RTSP is an application level client-server protocol, which is used to control the delivery of real-time streaming data [6]. RTP transports media data flows over UDP, in the same way as its related control protocol called Real-time Transport Control Protocol (RTCP) [7]. RTP carries data with real time requirements while RTCP conveys information of the participants and monitors the quality of the RTP session.

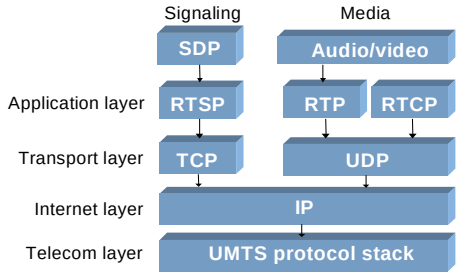

Fig. 1. Protocol stack for signaling and media flows of streaming services

B. *Architecture of the Mobile Network*

A general overview of the considered UMTS network architecture is depicted in fig. 2. Detailed descriptions of the entities, interfaces and protocols in UMTS are given in [8] and [9].

In addition to the User Equipment (UE), the main entities involved in QoS management are: UMTS Terrestrial Radio Access Network (UTRAN) and GSM/EDGE Radio Access Network (GERAN), Serving GPRS Support Node (SGSN), Home Location Register (HLR), Gateway GPRS Support Node (GGSN) and Application Server and RTSP Proxy.

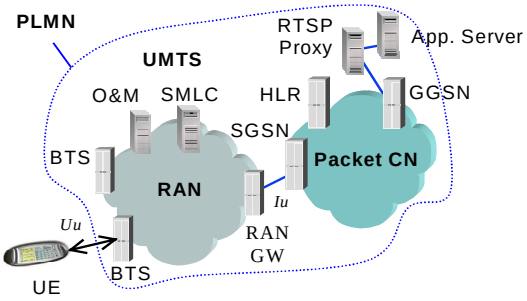

Fig. 2. End-to-end network architecture

End-to-end QoS in the UMTS Release 5 is based on the IP bearer service (IP BS) concept [10], which consists of the necessary extension of the UMTS BS defined in the UMTS release 1999 [4] to take into account the QoS in the external IP-PDN. In our model, the GGSN is connected to a RTSP proxy, which is also connected to the Streaming Server. Therefore, no external IP-PDN is involved in providing the streaming service.

III. Service Activation

Multimedia service session set up is described at three levels. In the first place, the service activation procedure from UE viewpoint is briefly outlined. Secondly, the signaling interchanges between application entities by using RTSP in order to establish the session is presented, as well as the media codec negotiation. Finally, all the signaling messages and mechanisms at lower layers (i.e. UMTS protocols) are explained in details.

A. *User Equipment Operation*

The service activation from user viewpoint can be described as follows. At first, user initiates the streaming client application, which connects to the UMTS network by using a socket Application Program Interface (API). The application requests a primary Packet Data Protocol (PDP) context which is opened to an specific access point with interactive UMTS traffic class and other suitable UMTS QoS release 99 parameters. A socket is opened for RTSP negotiation and it is tied to the interactive PDP context. The user then selects an audio streaming content. The application activates a streaming handler to take care of the streaming content. When the RTSP negotiation reaches the SETUP phase, a secondary PDP context is activated with QoS parameters suitable for audio streaming (RTP traffic) and for transport signaling (RTCP traffic). The RTP flow will start running through the streaming PDP context. New sockets are opened for RTP and RTCP traffic and they are tied to the streaming secondary PDP context.

B. *Application Layer Signaling*

The application layer signaling interchange between the UE and the streaming server is outlined in fig. 3. A primary PDP context is activated for the RTSP signaling between the terminal and the streaming server. By means of RTSP messages, information about the encoding of the media and the corresponding User Datagram Protocol (UDP) port number is interchanged. The SDP message [11] describes the streaming media the UE is about to receive. It should be noted that the RTSP specification defined by IETF [6] does not mandate the use of the DESCRIBE method for this media initialization phase. However, in order to function properly, any RTSP-based system must receive the description of the media one way or other. The 3GPP standard defining the protocols and codecs for the transparent end-to-end packet switched streaming service in 3G networks[12], mandates the use of the DESCRIBE method for the conveyance of the media description.

Afterwards, a secondary PDP context for the streaming media (RTP and RTCP flows) is activated. When resources for the media are successfully reserved, the UE sends the streaming server a PLAY request in order to start to receive the stream. The server sends the stream in form of a RTP flow. Likewise, RTCP traffic is sent for the QoS control of the corresponding RTP data flow.

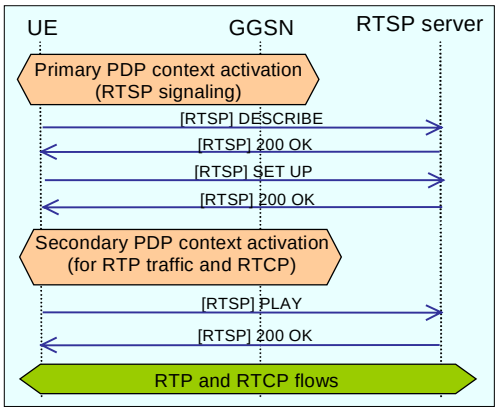

Fig 3. RTSP session initiation procedure in UMTS network

C. *UMTS Signaling Procedures*

Once the application level signaling procedure is presented, further insight about the UMTS signaling is provided. In UMTS, all signaling associated with service session establishment is carried out by the control plane through different QoS management functions (i.e. bearer service management, subscription, translation and admission&capability).

In the first place, a primary PDP Context is activated, as aforementioned, for RTSP signaling using interactive UMTS traffic class [4]. The interactive traffic class has a priority based handling instead of guarantees based handling, being the reliability requirement the target in this case. The control plane functions are distributed in different layers of several network entities.

The QoS requirements of the application in the UE are mapped into 3G QoS attributes. Since the primary PDP context is used for RTSP signaling, a 3G QoS profile with interactive traffic class, high priority and low error rate is appropriate. A Session Management (SM) protocol message from the UE to the SGSN initiates the PDP context activation procedure.

After the SGSN has validated the service for that user by querying HLR, local admission control is performed (e.g. based on the state of the buffers, the CPU load, etc.). Then, the SGSN maps the 3G QoS attributes into Radio Access Bearer (RAB) QoS attributes and triggers a RAB assignment procedure in the RAN by using the Radio Access Network Application Protocol (RANAP).

In the RAN, the admission control is basically based on the availability of radio resources. Once a new PDP context is accepted, RAB attributes are mapped into Radio Bearer (RB) parameters used in the physical and link layers (e.g. spreading codes, transmission modes, etc.). A RB according to these parameters is established and it is reported to the SGSN, which employs GPRS Tunneling Protocol for Control Plane (GTP-c) to indicate the GGSN that a new PDP context has to be created.

As the primary PDP context is not intended for real time traffic, no resource reservations are needed in the Core Network (CN). The GGSN accepts to create the primary PDP context based on similar admission criteria to those employed by the SGSN. Thereafter, the GGSN notifies the SGSN that the primary PDP context for RTSP has been successfully created and the SGSN sends a SM message to the application in the UE.

Once the streaming server accepts the RTSP connection request, the UE triggers a secondary PDP context activation procedure, used for both unidirectional RTP traffic and bidirectional RTCP traffic.

The UE converts user data application requirements into QoS profile for streaming class. Thus, table I shows an example of QoS profile for both RTP and RTCP data traffic. The QoS parameters requested for the PDP context take into account the full RTP/UDP/IP headers. Thus, no UDP/IP header compression is assumed in the IP level when requesting QoS.

Table I. Example of QoS Profile for RTP and RTCP traffic

QoS99 Parameter name	Parameter value
Traffic Class	Streaming
Maximum bitrate for uplink	4 kbps
Maximum bitrate for downlink	90 kbps
Maximum SDU size	1060 bytes
Delivery of erroneous SDUs	No
SDU error ratio	10^{-2}
Transfer Delay	1 s
Guaranteed bitrate for uplink	1 kbps
Guaranteed bitrate for downlink	72 kbps

For a given *SDU Error Ratio*, the larger the SDU size, the smaller the Block Error Rate (BLER), meaning the reliability requirements for radio link are stringent. Since a more protective coding scheme must be used, the bitrate is lower (for the same radio blocks sent), implying larger delay. Therefore, *maximum SDU size* should be commonly considered with the required *SDU error ratio*. From network viewpoint, smaller SDUs allow easier compliance to reliability requirements by relaxing the radio link adaptation. Moreover, a trade off between the reliability and delay relevancy should be found. This compromise needs to be communicated from UE application to the network or the application criteria for SDU size should be always conservative.

Once the QoS profile is derived, the secondary PDP context is activated. This procedure is quite similar to the above explained for the primary PDP context. The main differences in the secondary PDP context activation procedure are located in the RAN (UTRAN and GERAN) admission control.

IV. Service Utilization

Once the connection is established, the RTP data flow needs an appropriate QoS provisioning.

In the radio subdomain there are basically two options for conveying the data flow: CS bearer or PS bearer. The CS approach has the inherent drawback of the waste of resources, mainly in case of bursty traffic, as it is the case of streaming traffic. In other words, if resources are shared, trunking gain is obtained. The challenge comes from the need of guaranteeing certain bandwidth on shared channels whose radio link capacity is continuously varying, so enhanced RRM mechanisms are necessary for that purpose.

The Enhanced Quality of Service (EQoS) framework is a complete RRM system designed for the transmission of Guaranteed Bitrate (GBR) services (as streaming services) over EGPRS networks [13]. The EQoS scheme consists of different functionalities: in the establishment phase, an Admission Control and Channel Allocation scheme is used to accept/reject new allocation requests, according to the QoS requirements and the available radio resources; a Packet Scheduler, with the goal of providing to each allocated connection the needed air transmission time in order to guaranteed its bitrate requirements; and a Quality Control functionality, in charge of monitoring whether the provided QoS to each connection is in accordance with the negotiated QoS.

Main characteristics of EQoS feature are: the use of acknowledged Radio Link Control (RLC) mode of operation, due to undemanding delay requirements for streaming services; multiplexing of several GBR users over the same physical channel is allowed, by means of the use of shared Medium Access Control (MAC) mode. This use of shared channels leads to a better trunking efficiency, when comparing with dedicated channels. The reason for that trunking gain is the bursty nature of the incoming streaming traffic.

The EQoS RRM is designed to guarantee the required bitrate for different GBR requirements, as shown in fig. 4. However, the GBR requirement determines the maximum load that the system is able to support. Plainly, the system is able to manage more users with lower bitrate, since the lower the guaranteed bitrate, the higher the statistical multiplexing gain (i.e. a low bitrate involves a higher number of connections multiplexed over the same timeslot).

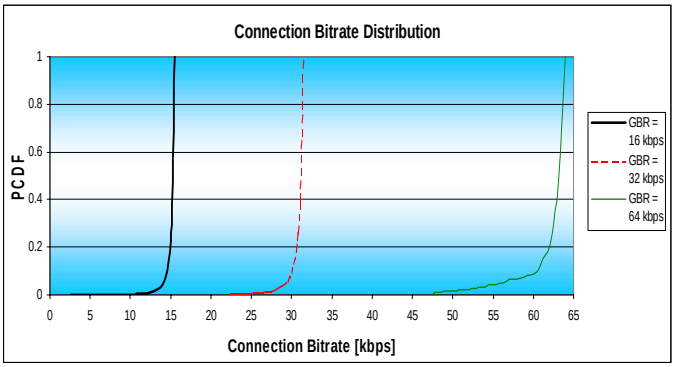

Figure 4. Connection Bitrate Distribution for different GBR Cases

Other advantage of the use of shared channels is that non-real time traffic can be transmitted over the same channels, making use of the remaining capacity left by streaming connections. Obviously, the higher the streaming capacity supported by the network, the lower the non-real time capacity multiplexed with it.

As above stated, streaming traffic does not have stringent delay requirements. However, in fig. 5, the transfer delay distribution for IP packets at link level is shown. In that figure, it can be observed how transfer delay requirements are also fulfilled. In addition, this figure provides useful information for the dimensioning of the compensating buffers used in application layer for streaming services, for EGPRS mobile networks.

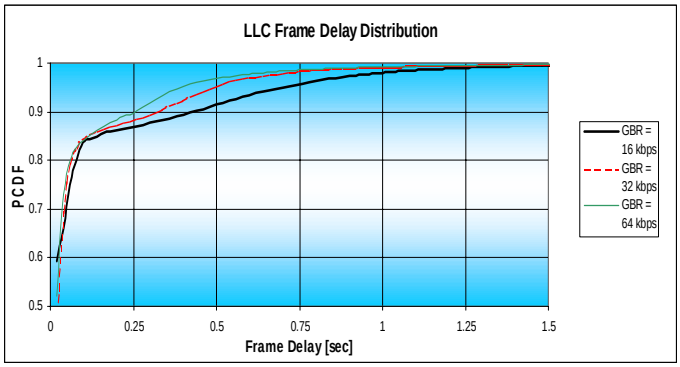

Figure 5. Frame Transfer Delay Distribution for different GBR Cases

In case connections with lower delay requirements are going to be transmitted through EGPRS networks, the maximum transfer delay can be controlled by means of the Admission Control, restricting the number of multiplexed connections in each timeslot. In fig. 5, transfer delay for a GBR requirement of 64 kbps is lower than for a GBR requirement of 16 kbps. The reason is that number of streaming connections multiplexed over the same timeslot in the 64 kbps case is much lower than in the 16 kbps case. By means of reducing the number of 16 kbps connections accepted in the system, transfer delay can be reduced to the desired requirement.

REFERENCES

[1] M. Margaritidis, and G. C. Polyzos, "MobiWeb: Enabling Adaptive Continuous Media Applications over 3G Wireless Link," IEEE Pers. Commun., vol. 7, Dec. 2000, pp. 36-41.

[2] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects, "Architecture for an all IP network," TR 23.922 v1.0.0, October 1999

[3] A. Mena and J. Heidemann, "An Empirical Study of Real Audio Traffic," IEEE INFOCOM 2000, vol. 1, pp. 101-110

[4] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects, "QoS concept and architecture," TS 23.107 v5.3.0, January 2002

[5] T. Halonen, J. Romero, J. Melero, GSM, GPRS and EDGE Performance, Evolution Towards 3G/UMTS, Wiley, 2002

[6] H. Schulzrinne, A. Rao, and R. Lanphier, "Real Time Streaming Protocol (RTSP)," IETF RFC 2326, April 1998

[7] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," IETF RFC 1889, 1996

[8] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects, "General Packet Radio Service (GPRS); service description; stage 2," TS 23.060 v4.3.0, January 2002

[9] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects, "Architectural principles for release 2000," TR 23.821 v1.0.1, July 2000

[10] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects, "End-to-end QoS concept and architecture," TS 23.207 v5.2.0, January 2002

[11] M. Handley and V. Jacobson, "SDP: Session Description Protocol," IETF RFC 2327, 1998

[12] 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects, "Transparent End-to-End Packet Switched Streaming Services (PSS); Protocols and Codecs," Release 4, TR 26.234 v4.2.0, December 2001

[13] D. Fernández, H. Montes, "An Enhanced QoS method for guaranteed bitrate service over Shared Channels in (E)GPRS," VTC2002 Spring