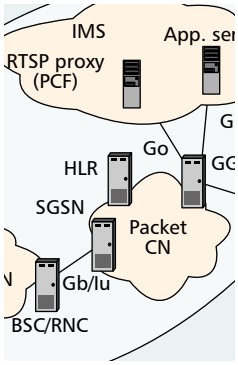
DEPLOYMENT OF IP MULTIMEDIA STREAMING SERVICES IN THIRD-GENERATION MOBILE NETWORKS

HÉCTOR MONTES, GERARDO GOMEZ, RENAUD CUNY, AND JOSE F. PARIS
NOKIA NETWORKS



Multimedia streaming services are receiving considerable interest in the mobile network business. Supporting reliable real-time services is a decisive factor for the increasing migration toward packet based mobile networks.

ABSTRACT

In this article, an end-to-end quality of service framework for streaming services in 3G mobile networks is considered. Under this scenario, the interaction between UMTS and IETF's protocols and mechanisms for a streaming session is analyzed. By signaling flowcharts, it is shown that both groups of protocols and mechanisms can co-operate to provide seamless end-to-end real-time services. Specifically, the article proposes to make the IP Multimedia Subsystem aware of Real Time Streaming Protocol, in order to extend its control from SIP to RTSP-based services, such as multimedia streaming services. Supported by this proposed framework, provisioning of audio streaming services over 3G mobile networks is also outlined.

INTRODUCTION

Multimedia streaming services are receiving considerable interest in the mobile network business. Supporting reliable real-time services is a decisive factor for the increasing migration toward packet-based mobile networks. For Universal Mobile Telecommunications System (UMTS), deploying an all-IP architecture is a promising standardization trend due to the convergence of IP technologies and telephony services. Multimedia streaming services are also technically applicable over evolving second- and third-generation (2G, 3G) wireless networks; thus, streaming clients will soon be incorporated into advanced wireless communication devices.

Although a few proprietary streaming technologies rule the Internet today, the proliferation of Internet Engineering Task Force (IETF) standardized protocols, such as RTSP, and aims to standardize an open streaming concept in major wireless standardization organizations (3G Partnership Project, 3GPP, and 3GPP2) will bring a strong open standards-based service to the wireless marketplace [1]. However, it seems inevitable that early adopters among operators will pilot the service with modified proprietary streaming technologies that are fitted to reliable wireless streaming, or rather progressive file downloading that will be transported over Hypertext Transport Protocol (HTTP) connections. One important advantage of supporting an existing commercial service platform (e.g., a RealNetworks™ or QuickTime™ server) is to provide added value from access to an existing service/content provider, besides its brand awareness.

One key issue is how mobile networks can support these kinds of services. In these "pre-all-IP" service cases the used radio bearers can be chosen from either a circuit-switched (CS) or packet-switched (PS) bearer set. The first commercial streaming services may well utilize existing CS bearer services, but in 3G the services will be offered over PS bearers. For those approaches where PS bearers are to be used, an open standardized approach will provide operators a better environment for creating productive business with widespread wireless streaming services. These wireless services will still utilize relatively low transmission bandwidths due to overall capacity restraints in the air link capacity. Thus, they should benefit from standardized and robust IP header compression methods while achieving acceptable quality of service (QoS) for end users.

Providing end-to-end QoS for multimedia streaming services implies harmonized interworking between protocols and mechanisms specified by IETF and 3GPP. Both groups of protocols and mechanisms are involved in QoS provisioning within the different 3G network subdomains and either the external IP-PDN or PLMN-hosted application servers through which the service is accessed. Even though the existing Release 99/Release 4 QoS concept provides sufficient support for streaming services [2], currently few Internet service providers (ISP) offer or apply end-to-end QoS in their Internet backbone networks. In fact, although several protocols and mechanisms, such as Resource Reservation Protocol (RSVP) and differentiated services (DiffServ), are proposed by IETF in order to manage the QoS in IP-PDN [3], they have not been completely deployed by ISPs.

1070-9916/02/$17.00 © 2002 IEEE

3GPP Release 5 introduces the IP Multimedia Subsystem (IMS) concept, which consists of network elements used in Session Initiation Protocol (SIP) based session control [4]. This article proposes to extend such control to RSTP-based services like multimedia streaming services.

In this article the end-to-end QoS management of streaming services in 3G mobile networks is considered. Particularly, the possibility of employing a PLMN-hosted multimedia streaming service is studied to avoid accessing streaming services through an external IP-PDN. With this solution the mobile operator hosts a streaming server or proxy server within the PLMN, allowing the operator to provide sufficient QoS to users of wireless streaming terminals. Supported by this proposed framework, different types of streaming services can be offered. More specifically, in this article provisioning of audio streaming services over 3G mobile networks is tackled. In addition, the presented analysis of the multimedia streaming session is chronologically divided in two phases: session initiation and session in progress.

The remainder of this article is organized as follows. First, a multimedia streaming service, mobile network architecture, and protocol stack are overall described. Second, session initiation is depicted in detail, highlighting both application and UMTS level signaling procedures. After that, the mechanisms involved in QoS provisioning in the UMTS network while a session is ongoing are outlined. Finally, the conclusions summarize the main ideas presented in this article.

Overall Scenario Description

Description of the Service: Multimedia Streaming

The content creation system of multimedia streaming services may have one or more media sources (e.g., a camera and a microphone). In order to compose a multimedia clip consisting of different media types, the raw data captured from the sources are edited. It should be noted that multimedia content could also be synthetically created without a natural media source. Animated computer graphics and digitally generated music also belong to this category. Typically, the storage space required for raw media data is quite large. In order to facilitate attractive multimedia retrieval service over commonly available transport channels such as low-bit-rate modem connections, the media clips are also compressed in the editing phase before they are handed to a server. Typically, several clients can access the server over a determined network. Then the client decompresses and plays the clip. In the playback phase, the client utilizes one or more output devices, most often the screen and the loudspeaker of the client.

By streaming, 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 media receiving, the client plays the media with a small delay or no delay at all. This technique not only frees up precious terminal memory, but also allows for media to be sent live to clients as the media event happens.

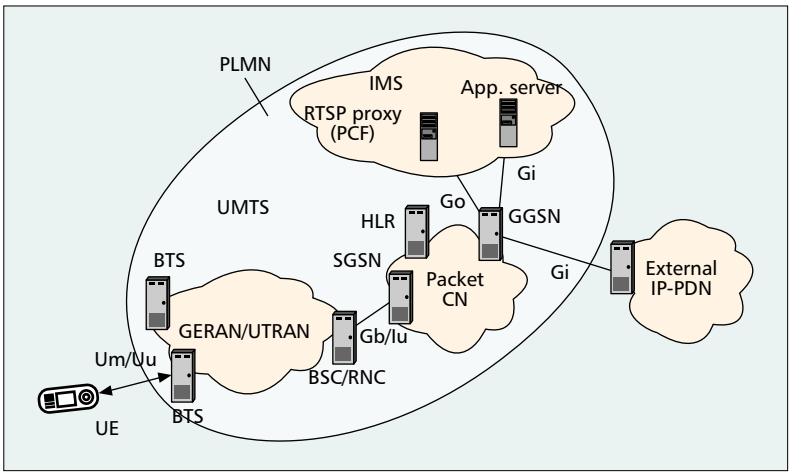


Figure 1. *End-to-end network architecture.*

Nowadays, since access to Internet services is moving fast to wireless devices, the available computing capacity in mobile devices is increasing, and user rates for cellular subscribers are approaching those of wired terminals, streaming service is also technically feasible in wireless handsets.

Generally, multimedia streaming by definition is seen to include one or several media streamed or transported to the client over the network. Some example services are:

- Audio streaming (offering music playback on the terminal); studied in this article
- Streaming with audio and video components (e.g., news reviews, music videos)
- Audio streaming with simultaneous visual presentation comprising still images and/or graphic animations, video clips presented in a predefined order (e.g., surfing through an interactive map)

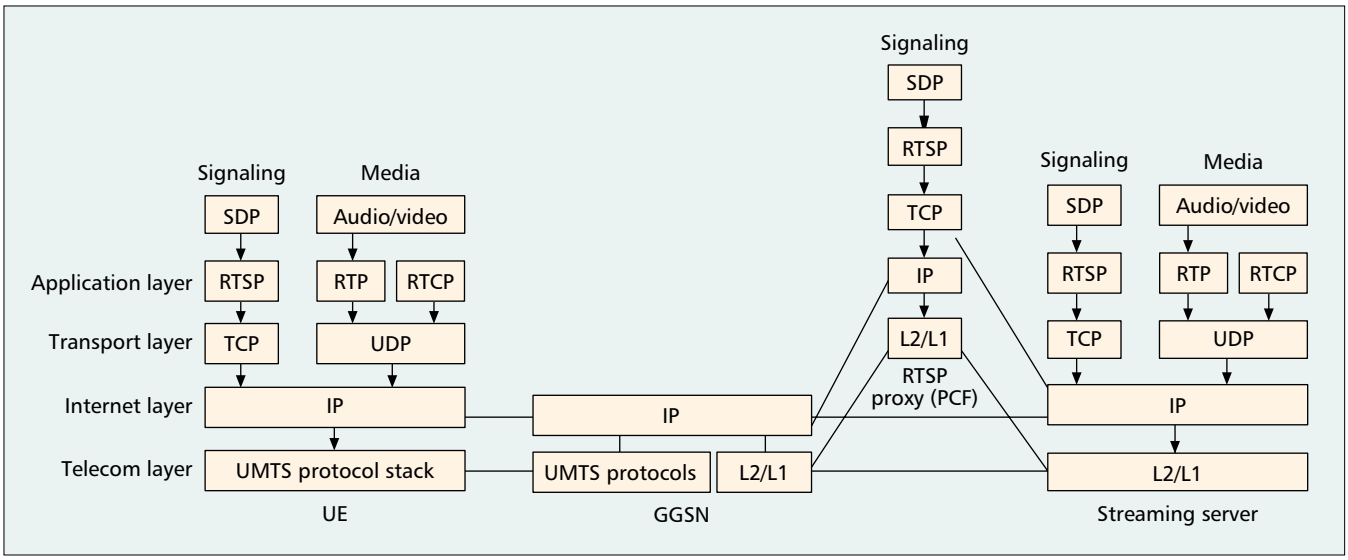
A statistical model of this kind of traffic depends quite a lot on the service. In the audio case, the generated traffic is rather nonbursty, whereas video traffic has a more bursty nature. In either case, PS bearers give more multiplexing gain and better resource utilization, while CS bearers offer better performance for those services that require stringent delay.

The Architecture of the Mobile Network

A general overview of the considered UMTS network architecture is depicted in Fig. 1. Detailed descriptions of the entities, interfaces, and protocols in UMTS are given in [5].

In addition to the user equipment (UE), the main entities in Fig. 1 that are involved in QoS management are:

- UMTS terrestrial radio access network (UTRAN) and Global System for Mobile Communications (GSM)/Enhanced Data Rates for GSM Evolution (EDGE) radio access network (GERAN)
- Serving GPRS support node (SGSN): the node that serves the UE and supports GPRS for GSM and/or UMTS (i.e., the Gb and/or Iu interface is supported by the SGSN)
- Home location register (HLR), which contains packet domain subscription data and routing information
- Gateway GPRS support node (GGSN): the



■ Figure 2. *The protocol stack for signaling and media flows of streaming services.*

first point of PDN interconnection with a GSM PLMN supporting GPRS (i.e., the Gi reference point is supported by the GGSN)

- Application server and RTSP proxy within the IMS, which includes GPRS/UMTS enhancements in Release 5 for the support of SIP-based voice over IP and data multimedia services

End-to-end QoS in UMTS Release 5 is based on the IP bearer service (BS) concept, which consists of the necessary extension of the UMTS BS defined in UMTS release 99/release 4 [2] to take into account QoS in both the external IP-PDN and the IMS domain. In our model, the GGSN is connected to an RTSP proxy, which is also connected to the streaming server. Therefore, no external IP-PDN is involved in providing the streaming service.

The IMS, where the RTSP proxy and streaming server are located, enables mobile network operators to offer their subscribers multimedia services based and built on Internet applications, services, and protocols. The IMS should enable convergence of, and access to, voice, video, messaging, data, and Web-based technologies for the wireless user, and combine the growth of the Internet with the growth in mobile communications. The IMS consists of network elements used in SIP-based session control, such as the policy control function (PCF). The PCF is standardized as a logical part of the proxy call state control function (P-CSCF) in 3GPP Release 5 specifications [4]. The P-CSCF/PCF interfaces with the GGSN via the standardized Go interface as well as the Gi reference point. PCF and GGSN interworking is based on the IP policy model, which allows the creation of a complete framework for IP BS management. Policies represent established service level agreements (SLAs) between service providers and users. SLAs specify a set of agreed rules for performing admission control that are based not only on the availability of the requested resources but also on accessibility, security, and other network performance issues expected by the UE.

The entity in charge of the IP BS policy management is the PCF, which is collocated with the RTSP proxy. The GGSN and RTSP proxy use Common Open Policy Service (COPS) protocol to interact and negotiate the IP BS [4].

Protocol Stack: Signaling and Media

The 3GPP multimedia streaming service is being standardized [1] based on control and transport Internet Engineering Task Force (IETF) protocols such as RTSP, Real-Time Transport Protocol (RTP) and Session Description Protocol (SDP), as Fig. 2 shows.

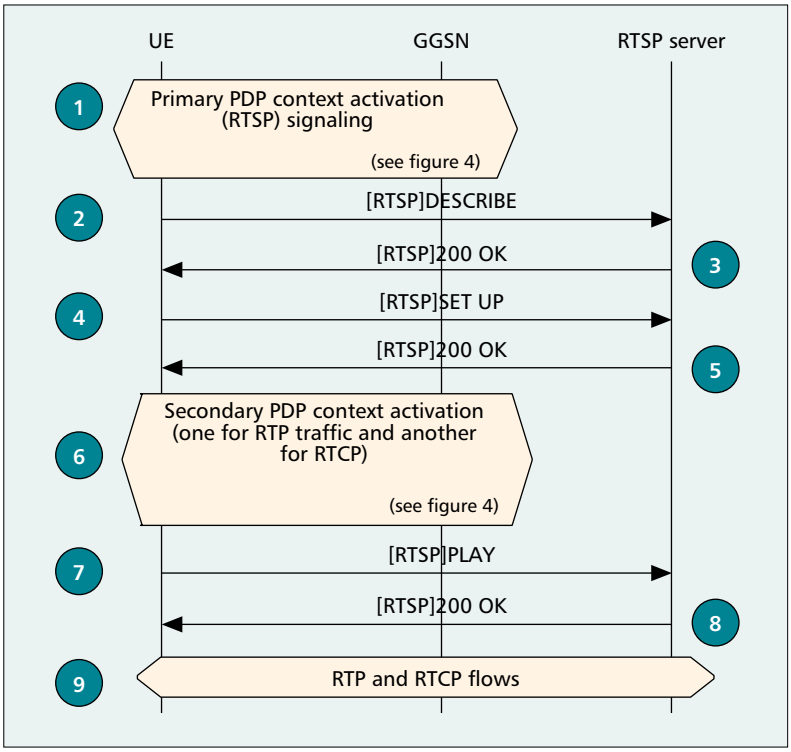
RTSP is an application-level client-server protocol used to control the delivery of real-time streaming data [6]. It establishes and controls one or several streams of continuous media but does not convey the media streams itself. The media streams may be conveyed over RTP, but the operation of RTSP is independent of the transport mechanism of the media streams.

A presentation description defines the set of media streams controlled by RTSP. The format of the presentation description is not defined in [6], but one example is the session description format, SDP, which is specified in [7]. SDP includes information on the media encoding and port numbers used for the media streams. Each media stream can be identified with an RTSP uniform resource locator (URL). This URL points to the media server that is responsible for handling a particular media stream. Thus, the RTSP specification allows separate media streams to reside in different servers.

RTSP may be sent over TCP while the media streams normally use UDP as the transport mechanism. Thus, the continuity of the media stream is not affected by delays in RTSP signaling. The RTSP request is a signaling message from the client to the server. The server sends responses back to the client by RTSP response status codes that are mainly reused from HTTP. Some methods in RTSP, similar to HTTP, play a central role in defining the allocation and usage of stream resources on the server. For instance, the DESCRIBE method retrieves the description of a presentation or media object identified

by the request URL from a server, constituting the media initialization phase of RTSP; SETUP causes the server to allocate resources for a stream and start an RTSP session; PLAY and RECORD start data transmission on a stream allocated via SETUP; PAUSE temporarily halts a stream without freeing server resources; and TEARDOWN frees resources associated with the stream.

Both RTP and its related control protocol called Real-Time Transport Control Protocol (RTCP) convey media data flows over UDP [8]. RTP carries data that has real-time requirements, while RTCP conveys information on the participants and monitors the quality of the RTP session. The RTP and RTCP services together provide payload type identification, sequence numbering, timestamping, and delivery monitoring. It should be pointed out that RTCP only affects the media encoding adaptation process. RTP defines a flexible framework for real-time data transport for multimedia services. However, a complete RTP specification for a particular application requires additional profile specification and payload format specification. The profile defines a set of payload type codes and their mapping to the payload formats that specify how a particular payload such as audio encoding is to be carried in RTP. RTP does not ensure timely delivery or provide any QoS guarantees.



■ Figure 3. *The RTSP session initiation procedure in a UMTS network.*

Session Initiation

This phase is described at three levels. In the first place, the session initiation procedure from the UE viewpoint is briefly outlined. Second, the signaling interchanges between application entities using RTSP in order to establish the session are 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 detail.

User Equipment Operation

Session initiation from the user viewpoint can be described as follows. At first, a 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 allocate the IP address for the UE as well as the access point. The primary PDP context, generally used in accessing either the IMS domain or an external network, is activated with interactive UMTS traffic class and other suitable UMTS QoS parameters. A socket is opened for RTSP negotiation, and it is tied to the interactive PDP context. The user then selects audio streaming content. The application activates a streaming handler to take care of the streaming content. Once the RTSP 200 OK message is received, the RTSP negotiation completes the SETUP phase (Fig. 3). Afterwards, new sockets are opened for RTP and RTCP traffic and tied to two secondary PDP contexts. One of them is activated with QoS parameters suitable for audio streaming (RTP traffic) and the other for transport signaling (RTCP traffic). The secondary PDP contexts reuse the same IP address and access point as the primary, but they may have different QoS profiles compared to the primary PDP context. The secondary PDP contexts must be activated before the RTSP PLAY command, because after that the RTP flow will start running through the streaming PDP context. The streaming handler launches a user interface to let the user control the audio stream, including, for example, play, pause, and stop knobs.

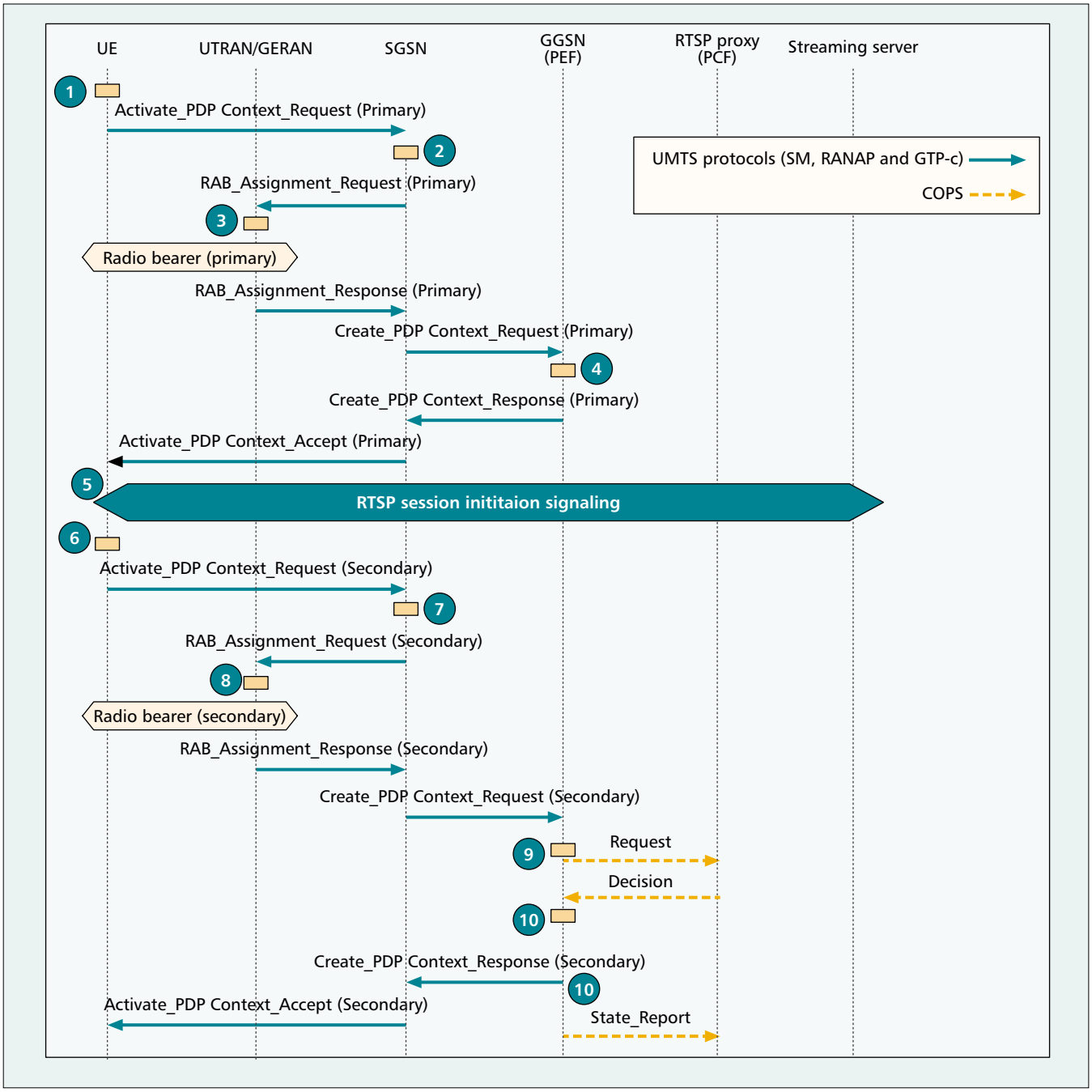
Application Layer Signaling

The application layer signaling interchange between the UE streaming client and the content provider streaming server is outlined in Fig. 3. A more detailed description is given in the following steps:

Step 1 — A primary PDP context is activated for the RTSP signaling between the terminal and the streaming server. The UE finds out the address of the streaming server when the user selects a link that points to streaming content residing in the streaming server. This step is presented in more detail in Fig. 4.

Step 2 — After creating a TCP connection to the streaming server, the UE sends an RTSP DESCRIBE request to the server. This request indicates that the server should send the UE information about the media it is going to send. This information includes the encoding of the media and the corresponding UDP port numbers.

Step 3 — The streaming server sends a 200 OK response containing a presentation description in the form of an SDP message. The SDP describes the streaming media the UE is about to receive. It should be noted that the RTSP specification as defined in the 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 sys-



■ Figure 4. *The multimedia streaming session initiation procedure in a UMTS network.*

tem must receive a description of the media one way or other. The 3GPP standard [1], which defines the protocols and codecs for the transparent end-to-end packet-switched streaming service in 3G networks, mandates the use of the DESCRIBE method for the conveyance of the media description.

Step 4 — The UE sends a SETUP request to the server. This message indicates the transport information of the stream including the UDP port numbers the UE is going to use for the RTP stream and the RTCP control traffic. The port numbers indicated in the DESCRIBE and the corresponding 200 OK messages are recommended values that can be overridden by the values given in the SETUP request and the corresponding 200 OK response.

Step 5 — The server acknowledges the SETUP request by sending a 200 OK response back to the UE.

Step 6 — In this phase the two secondary PDP contexts for the streaming media (RTP and RTCP flows) are activated. This step is presented in more detail in Fig. 4.

Step 7 — When the resources for the media are successfully reserved, the UE sends the streaming server a PLAY request in order to start to receive the stream.

Step 8 — The server replies with a 200 OK response to the UE client.

Step 9 — The server starts to send the stream in the form of an RTP flow. Likewise, RTCP

traffic is sent for QoS control of the corresponding RTP data flow.

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 and capability). Figure 4 illustrates the different steps of session initiation within the UMTS network.

In the first place, as mentioned earlier, a primary PDP context is activated for RTSP signaling. When a user clicks on the streaming icon in the mobile station, the streaming client application in UE requests an RTSP connection to the streaming server. Therefore, these messages are sent through the primary PDP context with interactive UMTS traffic class management [2]. The interactive traffic class has priority-based handling instead of guarantees-based handling, reliability being the target in this case.

The control plane functions are distributed in different layers of several network entities. Assuming that the service session establishment is successful, a detailed description of this phase is shown in the flowchart of Fig. 4.

Step 1 — The QoS requirements of the application signaling in the UE are mapped on UMTS QoS attributes. Since the primary PDP context is used for RTSP signaling, it requires high reliability. Therefor, a UMTS 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.

Step 2 — After the SGSN has validated the service for that user by querying the HLR, local admission control is performed (based on the state of the buffers, CPU load, etc.). Then the SGSN maps the UMTS QoS attributes on radio access bearer (RAB) QoS attributes and triggers an RAB assignment procedure in the RAN by using the RAN Application Protocol (RANAP). The RAB service provides confidential transport of signaling and user data between UE and the core network (CN) with QoS adequate to the negotiated UMTS BS.

Step 3 — In the RAN, admission control is mainly based on the availability of radio resources. Once a new PDP context is locally accepted in the 3G SGSN, the RAB attributes are mapped on radio bearer (RB) parameters used in the physical and link layers (e.g., spreading codes, retransmission requirements). An RB according to these parameters is established and reported to the SGSN. Since the GGSN is the entity in charge of managing the PDP contexts, the SGSN employs GPRS Tunneling Protocol for Control Plane (GTP-c) to indicate to the GGSN that a new PDP context has to be created.

Step 4 — Since the primary PDP context is not intended for real-time traffic, no resource reservations are needed in the CN. The GGSN accepts creating 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.

UMTS QoS attribute name	Attribute value
Traffic class	Streaming
Traffic handling priority	Not applicable
Maximum bit rate for uplink	0 kb/s
Maximum bit rate for downlink	90 kb/s
Delivery order	No
Maximum SDU size	1060 bytes
Delivery of erroneous SDUs	No
SDU error ratio	10^{-2}
Residual BER	10^{-3}
Transfer delay	2 s
Guaranteed bit rate for uplink	0 kb/s
Guaranteed bit rate for downlink	72 kb/s

Table 1. *The proposed UMTS QoS profile for audio RTP traffic.*

Step 5 — Once the streaming server accepts the RTSP connection request, by sending a 200 OK message responding to the RTSP Setup message, the UE triggers two secondary PDP context activation procedures, one for unidirectional RTP traffic and one for bidirectional RTCP traffic. The reason for the use of different secondary PDP contexts is that RTCP traffic must be separated from RTP if header compression is going to be applied for RTP/UDP/IP.

Step 6 — The UE converts user data application requirements into a QoS profile for the streaming class. Thus, Table 1 shows an example of a QoS profile for audio RTP data traffic.

The QoS parameters requested for the PDP context take into account the full RTP, UDP, and IP headers. Thus, no header compression is assumed in the IP level when requesting QoS. Some assumptions have been made for this proposed QoS profile. A bit rate of 64 kb/s is assumed (e.g., MPEG-AAC codec). This bit rate achieves good stereo quality. However, bit rates of 24–48 kb/s could be also tolerable. If mono audio were preferred, 32 kb/s would be good. The payload size from the streaming application in this example is assumed to be between 500–1000 bytes. The downlink bit rate of 72 kb/s is calculated by including the impact of the following header sizes: RTP 12 bytes, UDP 8 bytes, and IPv6 40 bytes. Since RTP flow is unidirectional, guaranteed bit rate for uplink is set to 0 kb/s.

Since the maximum bit rate attribute is used for policing and shaping at the GGSN, its value is above the guaranteed bit rate requirement to leave a safety margin for possible bit rate fluctuation.

Due to the existence of jitter compensating buffers at the application layer of duration around 5 or 6 s, transfer delay is set to 2 s. Since the delay requirement is not stringent, retrans-

The QoS parameters requested for the PDP context take into account the full RTP, UDP, and IP headers. Thus, no header compression is assumed in the IP level when requesting QoS. Some assumptions have been made for this proposed QoS profile.

The UMTS QoS profile is mapped to RAB QoS attributes. The UMTS QoS attributes are exactly the same as the RAB QoS attributes but the values are not typically the same for the following parameters: Residual BER, SDU error ratio and Transfer Delay, due to the packet loss and the delay inside the Core Network.

missions at the radio link level are allowed.

For a given service data unit (SDU) error ratio, the larger the SDU size, the smaller the bit error ratio (BER). The reliability requirements for the radio link are therefore stringent. Since a more protective coding scheme must be used, the bit rate 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 the network viewpoint, smaller SDUs allow easier compliance to reliability requirements by relaxing the radio link adaptation. However, too small SDUs increase the associated overhead leading to larger delays. Moreover, a trade-off between the relevance of reliability and delay should be found. This compromise needs to be communicated from the UE application to the network, or the application criteria for SDU size should be always conservative.

The delivery order attribute is disabled because RTP does not assume that the underlying network is reliable and delivers packets in sequence. The sequence numbers included in RTP allow the receiver to reconstruct the sender's packet sequence [8].

In a similar way, UE converts transport control requirements into a QoS profile for RTCP traffic.

Once the QoS profiles are derived, the secondary PDP contexts are activated. This procedure, which is performed for both secondary PDP contexts, is outlined in Fig. 4. The main steps of a secondary PDP context activation procedure are described below. The RTP traffic PDP context is used for this example.

An *Activate Secondary PDP Context Request* SM message is sent from UE to SGSN. This message contains, among other parameters, the requested QoS attributes and the traffic flow template (TFT). The TFT is sent transparently through 3G-SGSN to 3G-GGSN to enable packet classification for downlink data transfer.

Step 7 — The 3G-SGSN validates the service request. The HLR is queried to check if the service can be provided to this subscriber, and then the 3G SGSN performs admission control. There are two configurable parameters that control the maximum amount of streaming traffic: real-time bandwidth and streaming bandwidth. These parameters are used to check if there is enough bandwidth for the new PDP context. If there is, the flow is accepted and the resource reservation is performed by decreasing the bandwidth quota by the guaranteed bit rate of the new PDP context. In addition to configuration parameters, the admission control procedure checks the CPU load so that there is processing capacity for an additional flow before accepting the PDP context activation. The allocation/retention priority parameter is given by the HLR to the SGSN. This parameter is used for admission precedence, that is, to select which is accepted among several users when streaming bandwidth is limited.

The UMTS QoS profile is mapped to RAB QoS attributes. The UMTS QoS attributes are exactly the same as the RAB QoS attributes, but the values are not typically the same for the following parameters: residual BER, SDU error ratio, and transfer delay, due to the packet loss and delay inside the CN.

The 3G-SGSN sends an RAB *Assignment Request* RANAP message to the RAN through the Iu interface. This message contains the RAB QoS attributes.

Step 8 — The RAN (both UTRAN and GERAN) performs admission control. RAN must do the mapping between RAB and RB parameters. When requesting a QoS profile for a PDP context, the parameters should be requested for full header IP packets. Since the header compression applies only at the PDCP layer in the radio Uu interface, the impact of header compression is only taken into account in RAB to RB mapping, when the resources are requested. Therefore, RB resources should be reserved when RAN applies header compression according to the corresponding bit rate. Later, the RAN establishes the RAB with the selected cell. After that, the 3G-SGSN receives a *RAB Assignment Response* RANAP message.

Step 9 — The 3G-SGSN sends a *Create PDP Context Request* GTP-c message to the 3G-GGSN with the negotiated QoS. The GGSN generates a new entry in its PDP context table and stores the TFT. Local admission control and resource reservation are performed in the 3G-GGSN in the same way as in the 3G-SGSN. The allocation/retention priority parameter is also used in a similar manner as for 3G-SGSN. Once the local admission control is performed in the GGSN based on its own capability, it outsources the admission control to the PCF in the RTSP proxy by sending a COPS message. The PCF applies appropriate rules to the streaming service and sends its decision back to the GGSN.

Step 10 — The 3G-GGSN replies with a *Create PDP Context Response* GTP-c message to the 3G-SGSN. Likewise, the 3G-SGSN replies with an *Activate PDP Context Accept* SM message to the UE. The SGSN is now able to route PDP PDUs between the GGSN and the UE, and to start charging.

Step 11 — The 3G-GGSN reports the success of the secondary PDP context activation procedure to the SGSN and the PCF in the CSCF. Finally, the SGSN sends the corresponding SM message to the UE so that it knows of the end of the service session establishment.

Session in Progress

Once the connection is established, the RTP data flow needs appropriate QoS provisioning, in both IP transport and radio domains.

In the IP transport domain, a DiffServ mechanism [3] is employed (i.e., both in the CN and between GGSN and RTSP server, since the GGSN us responsible for mapping UMTS QoS parameters into DiffServ parameters). DiffServ mechanism is based on different per-hop behaviors (PHBs). Each PHB consists of the rules used to treat packets in specific ways inside the network. More specifically, PHB denotes a combination of forwarding, classification, scheduling and drop behaviors at each hop. For streaming traffic, two groups of PHBs can be applied: expedited forwarding (EF) or assured forwarding (AF). The EF PHB target is to provide tools to build a low-loss low-latency low-jitter assured-

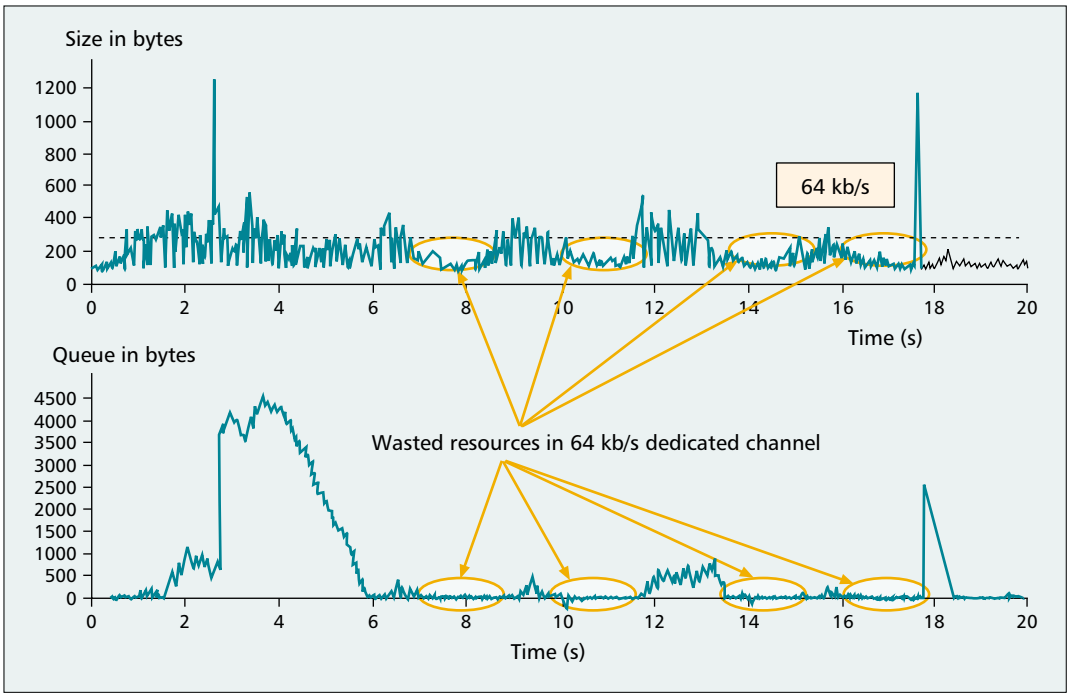


Figure 5. *The bursty nature of streaming traffic.*

bandwidth end-to-end service within the DiffServ domain, with the drawback of the complexity it introduces in the system. Due to the undemanding QoS requirements of streaming services, mainly in comparison with other real-time traffic like VoIP services, AF PHB can be used. Inside an AF PHB group there are a number of PHB delay classes, each with a number of drop precedence levels. For streaming traffic the highest priority should be used.

In the radio domain there are basically two options for conveying streaming traffic: CS or PS bearer. The CS approach has the inherent drawback of waste of resources, mainly in the bursty traffic case. Streaming traffic analysis shows its bursty nature, as Fig. 5 depicts. The graph above in Fig. 5 presents the fluctuation of the bitrate for such a traffic source (assuming constant packet interarrival time and variable packet size), whereas the graph below depicts the queue status. The source generates traffic at an average rate of 64 kb/s. Due to the variation of the source rate, the existence of less activity periods (bit rate below the average one) is observed. Likewise, when these periods are large enough the queue gets empty (see highlighted parts of Fig. 5). Therefore, when a dedicated capacity of 64 kb/s is allocated for such a connection there is a waste of resources. In other words, if resources are shared, multiplexing gain is obtained.

Since 3G mobile networks are going to support multiradio technologies, such as wideband code-division multiple access (WCDMA) [9] and EDGE [10], in this article the QoS provisioning in both radio technologies is briefly outlined. In UTRAN there are basically two types of bearers: dedicated channel (DCH) or data shared channel (DSCH). Otherwise, GERAN provides different bearers that can support streaming services: traffic channels (TCHs) like high-speed CS data (HSCSD) and enhanced CS data (ECSD) from the CS domain, or packet data channel (PDCH) from the PS domain. As mentioned earlier, the use of shared resources gives operators higher multiplexing gain. The challenge comes from the need to guarantee certain bandwidth on shared channels whose radio link capacity is continuously varying, so enhanced QoS mechanisms are needed for that purpose. This requires coordination between admission control and resource allocation as well as packet scheduling and link adaptation algorithms [11].

When the QoS negotiated during service establishment cannot be maintained by any network entity, different QoS control mechanisms have to be employed. Subsequently, some control plane signaling activity is needed to coordinate all these mechanisms, especially in order to provide a seamless end-to-end service bearer from the user point of view. The control plane activity when QoS degradation occurs can be divided into two different groups of mechanisms:

- QoS preserving mechanisms, which are transparent to UE. For example, some RAN internal mechanisms are able to detect radio link degradation so that specific control plane signaling procedures are triggered to successfully recover the negotiated QoS (e.g., by means of radio resource reallocation or cell reselection).
- QoS renegotiations mechanisms. When the first type of mechanisms cannot successfully keep the negotiated QoS, it is possible to renegotiate a downgraded QoS profile with the UE. Therefore, this group of mechanisms is not transparent to the UE.

Conclusions

Since supporting reliable real-time services is a decisive aspect in packet-based telephony networks, an end-to-end QoS framework for streaming services in 3G mobile networks is considered.

Due to the undemanding QoS requirements of streaming services, mainly in comparison with other real-time traffic like VoIP services, the AF PHB can be used. Inside AF PHB group there are a number of PHB delay classes, each with a number of drop precedence levels. For streaming traffic the highest priority should be used.

The 3GPP Release 5 introduces the IMS, which consists of network elements used in SIP based session control. The authors propose to extend such control to RTSP based services such as multimedia streaming services.

This article addresses a solution based on a PLMN-hosted multimedia streaming service. Signaling flowcharts have shown that UMTS and IETF's protocols can cooperate to provide seamless end-to-end real-time services. Thus, session initiation has been described at three levels: initiation from the UE viewpoint, the RTSP signaling interchanges between application entities, and the UMTS signaling procedures.

3GPP Release 5 introduces the IMS, which consists of network elements used in SIP-based session control. This article proposes to extend such control to RTSP-based services such as multimedia streaming services. This solution avoids adding specific methods to SIP, such as PLAY and STOP, when a protocol, RTSP, specified in IETF already exists for this purpose. The only reason in the scope of 3GPP to add these methods to SIP is that the IMS supports SIP and that it might be beneficial for the operator to include streaming as one service in the IMS. However, the problem is that there are other streaming servers outside the IMS that still use the industry standard RTSP. Making PCF aware of RTSP solves this problem.

Provisioning of audio streaming services over 3G mobile networks has also been tackled in this article. Results from traffic behavior analysis have shown the convenience of using PS bearers in the radio domain. In shared channels, the challenge of assuring capacity for such traffic has also been pointed out.

Acknowledgments

This work has been performed as part of the cooperation agreement between Nokia and the University of Malaga. This agreement is partially supported by the Program to Promote Technical Research (Programa de Fomento de la Investigación Técnica, PROFIT) of the Spanish Ministry of Science and Technology. The authors wish to thank J. Jouppi and K. Ahvonen for their contribution, and to J. M. Melero and J. Muñoz for their support and useful comments.

References

[1] 3GPP, "Transparent End-to-End Packet Switched Streaming Services (PSS); Protocols and Codecs," Release 4, TR 26.234 v4.3.0, Mar. 2002.
[2] 3GPP, "QoS Concept and Architecture," TS 23.107 v5.3.0, Jan. 2002.
[3] S. Blake *et al.*, "Architecture for Differentiated Services," IETF RFC 2475, Dec. 1998
[4] 3GPP, "End-to-End QoS Concept and Architecture," TS 23.207 v5.2.0, Jan. 2002.
[5] 3GPP, "General Packet Radio Service (GPRS); Service Description; Stage 2," TS 23.060 v4.4.0, Mar. 2002.
[6] H. Schulzrinne, A. Rao, and R. Lanphier, "Real Time Streaming Protocol (RTSP)," IETF RFC 2326, Apr. 1998.
[7] M. Handley and V. Jacobson, "SDP: Session Description Protocol," IETF RFC 2327, 1998
[8] H. Schulzrinne *et al.*, "RTP: A Transport Protocol for Real-Time Applications," IETF RFC 1889, 1996.
[9] H. Holma and A. Toskala, *WCDMA for UMTS Networks*, Wiley, 2000
[10] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS*, Wiley, 2002.
[11] H. Montes and D. Fernández, "An Enhanced Quality Of Service Method for Guaranteed Bitrate Services over Shared Channels in (E)GPRS Systems," *IEEE 54th VTC*, Birmingham, AL, May 2002.

Biographies

HÉCTOR MONTES (ext-hector.montes@nokia.com) received his engineer degree (M.Sc.) in telecommunications at the E.T.S.I.T (Escuela Técnica Superior de Ingenieros de Telecomunicación), from the University of Málaga in 2001. He joined Nokia Networks in 2000 where he is managing a project related to QoS in mobile networks. His research interests include the field of mobile communication systems, specifically QoS and radio resource management.

GERARDO GOMEZ (ext-gerardo.gomez@nokia.com) received his engineer degree (M.Sc.) in telecommunications at the E.T.S.I.T (Escuela Técnica Superior de Ingenieros de Telecomunicación), from the University of Málaga in 1999. He was involved in different European projects before joining Nokia Networks in 2000. His research interests include the field of mobile communication systems, especially QoS and radio resource management.

RENAUD CUNY (renaud.cuny@nokia.com) received an engineer degree (M.Sc) in computer sciences from the EFREI (Ecole Francaise d'Electronique et d'Informatique) in 1997. He started at Nokia Research Center in Helsinki in 1998 defining traffic management methods in wireless networks. He joined Nokia Networks in 2001 and is now involved in a research program that studies end to end quality of service in 3G networks.

JOSE F. PARIS (paris@ic.uma.es) received his engineer degree (MSc) in telecommunications in the E.T.S.I.T (Escuela Técnica Superior de Ingenieros de Telecomunicación), from the University of Málaga in 1996. In 1996 he developed R&D activities related to wireless telephony in Alcatel. Since 1997, he has worked at the Universidad de Málaga and currently is a associate professor who collaborates with Nokia Networks. His research interests are different topics on mobile networks such as adaptive modulation, equalization, and QoS.