

End-to-end QoS Signaling for Future Multimedia Services in the NGN

Lea Skorin-Kapov¹ and Maja Matijasevic²

¹ R&D Center, Ericsson Nikola Tesla, Krapinska 45, HR-10000 Zagreb, Croatia

² FER, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia
lea.skorin-kapov@ericsson.com, maja.matijasevic@fer.hr

Abstract. The paper presents a model for Quality of Service (QoS) signaling for complex multimedia services in the next generation network (NGN), from establishing necessary QoS support at session setup, to further QoS modifications in response to dynamic changes during the session. The model aims to maximize user perceived quality, while taking into account constraints imposed by the user, network, and service itself. While the model is independent of the particular service and network scenario, we take a Networked Virtual Reality (NVR) service as a fairly good representative of intricate media-rich services in the NGN. We further propose a mapping onto the IP Multimedia Subsystem (IMS), as the most prominent NGN-driven part of the current Universal Mobile Telecommunications System (UMTS) architecture; and identify its possible functional enhancements for supporting services such as NVR.

1 Introduction

One of the key requirements for the next-generation network (NGN) is to provide seamless Quality of Service (QoS) for converged mobile and fixed multimedia services “anywhere-anytime” [6]. The provisioning of end-to-end (E2E) QoS, in general, involves signaling and dynamic negotiation/re negotiation between involved parties in order to agree on a common and feasible set of QoS parameters at session set-up, as well as reacting to possible changes during the session. In this paper, we present a model supporting the signaling, negotiation, and adaptation of QoS for media-rich interactive services, addressing issues arising due to the heterogeneous environment of NGNs. We focus on control signaling at the application/session level, rather than on the underlying mechanisms and architecture that provide actual service delivery. For reference purposes, we use the model proposed in our previous work [12], which takes into account the following issues: (1) user terminal and access network constraints; (2) way(s) of expressing user preferences in terms of application components (media elements); (3) dynamic resource availability and cost; and, (4) mapping of user/application requirements to transport QoS parameters. While the proposed model is service independent, we discuss it in the context of Networked Virtual Reality (NVR), looking to address one of the most “complex multimedia service” scenarios. Characterized by 3D graphics, integrated multimedia components, and user interaction (perceived

as) in real-time, examples of NVR services include networked 3D games, virtual worlds for training and collaborative work, and many more.

The paper is organized as follows. Related work on E2E QoS signaling and negotiation for multimedia services is discussed in Section 2. A generic QoS model is described in Section 3, with sequence diagrams provided to illustrate session QoS setup and (re)negotiation. Section 4 presents a case study involving the mapping of model entities to the IMS architecture. In Section 5, we give concluding remarks and identify open issues for ongoing and future work.

2 Related work

Issues of E2E QoS signaling and negotiation regarding the NGN architecture have been addressed in ITU-T, 3GPP, IETF, and ETSI/TISPAN, as well as in recent research literature. The ITU-T defines an NGN as a packet-based network able to provide telecommunication services and make use of multiple broadband, QoS-enabled transport technologies [3]. A generic architecture for E2E QoS control and signaling for multimedia services is defined by the ITU-T in [4]. In parallel to ITU-T, ETSI/TISPAN has embraced the 3GPP IP Multimedia Subsystem (IMS) architecture, which has become an internationally recognized standard for offering multimedia services in the packet switched domain. With regards to application level QoS signaling, the IETF Session Initiation Protocol (SIP) [10] has been adopted by 3GPP as the key session control protocol for IMS. The common idea of all mentioned approaches is a horizontally layered architecture, with QoS-signaling interfaces between service-level call/session functionality and the underlying transport/connectivity functions. The relationship and efficient mapping of QoS parameters across layers, however, remains a challenge, especially for complex, highly interactive, and media-rich services. As such, it has been investigated from various points of view, with those mentioned next the most relevant to our approach. An evaluation of scenarios involving relationships between application-level and network-level QoS signaling during session negotiation, renegotiation, and handover can be found in [9]. The work [8] proposes an End-to-End Negotiation Protocol (E2ENP) for active negotiation of QoS for multimedia, based on SIP and SDPng (new generation), the successor of the Session Description Protocol (SDP) [7]. The authors assume an end-user application deriving valid QoS Contracts (service configurations) negotiated between users and enforced based on dynamic resource availability and user expectations. Further dealing with the issue of convergence between user preferences and expectations and network resource constraints, the authors in [5] propose an adaptive QoS control architecture for multimedia wireless applications. The referenced approaches have been analyzed with the goal of inspiring and contributing to the proposed model supporting signaling and dynamic QoS adaptation for complex multimedia such as NVR. Our model aims to provide a general framework within which various methods for QoS negotiation and/or parameter matching may be applied.

3 QoS Control Model

The presented model represents an adaptive QoS system supporting the process of E2E QoS control based on negotiation/renegotiation and service adaptation from the moment a user accesses a service until service termination. The model, shown in Fig. 1, is composed of a number of functional entities grouped logically into three components: *Client*, *Access and Control*, and *Application Server*. In the figure, solid arrows between model entities represent data flows, while dashed lines represent control (signaling) flows.

Upon a user's initiated service request, the *Client* sends a request to the *Access and Control* entities. The request contains user preferences and terminal capabilities in a client profile, or, a reference to a profile in a *client profile repository*. The *Access and Control* entities are responsible for identifying the client and service requirements, and authorizing necessary network resources. Negotiated and authorized parameters serve as input to an optimization process designed to dynamically calculate the resource allocation and *application operating point*. The application operating point refers to the final application configuration (included media components, corresponding codecs/formats, etc.) to be delivered to the user from the *Application Server*. After calculation, reservation mechanisms are invoked to reserve network resources, hence we are assuming an underlying network with implemented QoS mechanisms. A high-level view of the QoS control and negotiation process is given in Fig. 2. Once the service starts, further optimization and renegotiation/adaptation procedures may be invoked during service lifetime in response to changes in resource availability and/or resource cost; the client profile (user preferences, terminal capabilities, access network); and service requirements.

3.1 Access and Control

The *Access and Control* component, shown in Fig. 1, represents a logical grouping of service control functionalities. In an actual network scenario, proposed functionalities may be distributed among different entities in the network.

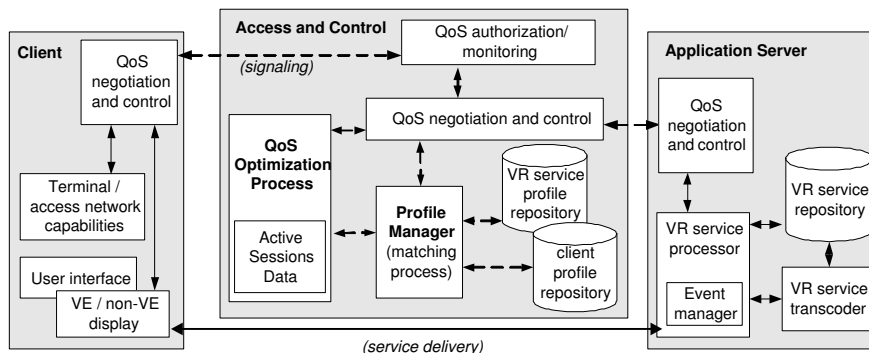


Fig. 1. Model for dynamic negotiation and adaptation of QoS for NVR services.

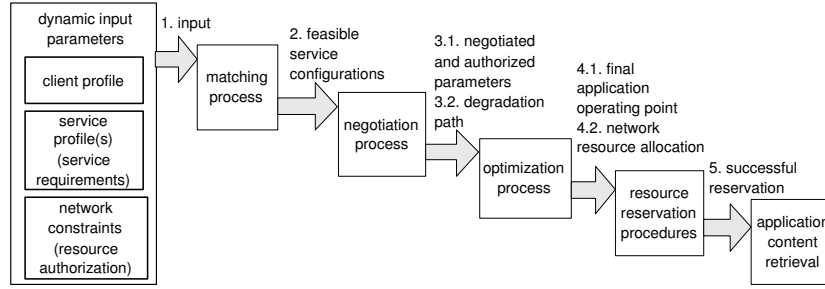


Fig. 2. QoS negotiation and control process.

Client profile determination. A user’s initial service request is constructed by way of a user interface enabling a user to request a service and set (possibly predefined) preferences. Options include enabling a user to specify desired quality (e.g. high, medium, low) of media components, preferences such as “audio has priority over video”, budget constraints, minimum acceptable framerate, and maximum acceptable download time. Preferences, together with terminal hardware, terminal software, and access network characteristics are incorporated into a client profile. The proposed generic client profile is shown in Table 1. For the purposes of this paper, we assume that communication end points are capable of QoS-related signaling. It should be noted, though, that in a general case, where the user terminal does not support QoS signaling mechanisms, QoS requests may be issued through network session control entities.

Table 1. Generic client profile

<i>Client Profile</i> parameters			
Terminal hardware	Network Characteristics	Terminal Software	User Preferences
<ul style="list-style-type: none"> - Model - Display size - Processor type - Processor MIPS - Memory - Color depth - Sound option - Input character set - Output character set 	<ul style="list-style-type: none"> - Current bearer service - Supported bearers - Downlink (bandwidth, delay, loss, jitter, BER) - Uplink (bandwidth, delay, loss, jitter, BER) 	<ul style="list-style-type: none"> - OS (name, vendor, version) - Browser info - Supported software - Supported media types 	<ul style="list-style-type: none"> - Acceptable service format(s) - Textures (accept, desired quality) - Audio (accept, desired quality) - Video (accept, desired quality) - Text (accept, desired quality) - Data (accept, desired quality) - Max. download time - Budget - Min. acceptable framerate - Degradation code

Profile Manager. The client request, including or referencing the client profile, passes via the *QoS authorization/monitoring* module and is received by the *QoS Negotiation and Control (QNC)* module of the *Access and Control* component. The profile, or a reference to it, is passed to the *Profile Manager (PM)* module. The PM matches the restrictive parameters of the client profile with parameters of the retrieved VR service profile(s) corresponding to the requested service in order to determine feasible service configurations.

The *VR service profile repository* contains profiles specifying supported configu-

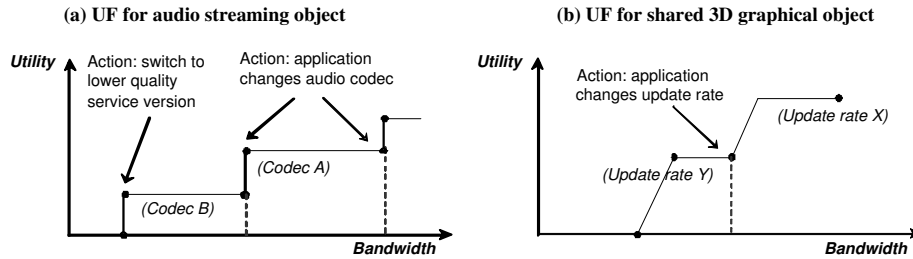


Fig. 3. Example Utility Functions for VE objects.

rations (versions) of services. Multiple application configurations may be feasible in order to address issues of heterogeneity stemming from diverse end user capabilities and preferences. For example, an application may be implemented in two ways: with media streaming (configuration 1), and without media streaming (configuration 2). The repository may be located on the application server hosting the actual service content, or on a different server in the network. We specify a generic service profile as shown in Table 2.

By considering the NVR service to be comprised of a collection of virtual world objects (e.g. 3D graphic objects, sound, video, etc.) integrated into a virtual environment (VE), communication requirements may be expressed as a combination of the requirements of particular objects mapped down to transport level QoS parameters. A detailed discussion of the parameters and a proposed mapping to QoS parameters and classes for UMTS, is given in [11]. In addition, factors such as a user's level of interest in an object inside the VE may be considered. Hence, we specify particular object requirements by combining object utility functions (UFs) and the notion of their relative importance. An example of UFs defined as functions of bandwidth for an audio streaming object (a) and a shared 3D graphical object (b) within a VE is shown in Fig. 3. Minimum requirements are indicated by those values at which utility drops to zero, while maximum requirements correspond to the marginal utility of utility value approaching one. It should also be noted that UFs may vary over time and across different configurations. Assuming that the application can adapt its

Table 2. Generic service profile

<i>Service Profile</i> parameters			
Metadata	Network Requirements	Processing Requirements	Adaptation Policy
- Service designation	- General	- Display size	- User interest adaptation policy
- Service description	- downlink (bandwidth, delay, loss, jitter, BER)	- Polygons	
- Service identification	- uplink (bandwidth, delay, loss, jitter, BER)	- Textures	- Network resource adaptation policy
- Service format	- Object i profile	- Sound	- Frame rate adaptation
	- name, object url, media type, formats	- Video	
	- downlink: UF(s)/loss/delay, jitter, weight factor	- Lighting	
	- uplink: UF(s)/loss/delay, jitter, weight factor	- Text	
	- media quality preference mapping to traffic class	- Data	
	[... profiles for other objects ...]	- File size	
		- Software support	

configuration to available bandwidth, Fig. 3 also illustrates actions to be taken when certain bandwidth thresholds are reached. Application adaptation based on multi-dimensional UFs has been addressed previously in literature [13].

In Table 2, the *Network Requirements* parameters refer to E2E QoS requirements as perceived by a user. The first subset (labelled *General*) corresponds to the overall minimum network requirements. The second subset contains particular object requirements (labelled *Object n profile*), specified in the form of UFs. The relative importance of objects (media components) to the user is taken into account by multiplying utility values with weight factors (WF) ranging from 0 to 1. One way of determining object WFs is by specifying user perceived Level of Interest (LoI) as *high*, *medium*, and *low*. A user indicating that audio is more important than video can lead to an LoI value of *high* for audio (e.g. a WF of 1 for the audio object UF), and an LoI value of *low* for video (e.g. a WF of 0.25 for the video object UF). The *Adaptation Policy* options specify the actions to be taken as a result of change in a) User interest (LoI); b) Network resources; and, c) Frame rate. For example, in response to decrease in bandwidth, a lower bitrate codec may be selected; or, to maintain a stable frame rate, the order in which to degrade the quality of objects may be specified. Parameters of the service profile relating to network requirements are updated dynamically over the course of the service lifetime (e.g. media streaming object is started/stopped; a new user joins a shared 3D environment; certain WFs have changed). Based on the given service profile and the particular client profile, the matching process is conducted by the PM to determine *feasible* service configurations, such that: (1) A client's terminal capabilities can support the service processing requirements; and (2) The client's access network can support the minimum requirements for all (required) VE objects; and (3) The client's preferences in terms of acceptable media components and maximum acceptable download time can be met.

After the matching process, the PM extracts a set of potential session parameters (may include multiple potential media formats, codec types, etc.) from those service configurations that are feasible. These parameters are passed to the QNC module, which sends a session offer to the *Client* with a set of session description parameters. The *Client* may then accept/deny/modify offered parameters. A message is returned indicating the subset of offered parameters agreed to by the *Client*. Network entities authorize resources based on the agreed parameter subset. The returned session parameter subset and authorization is passed back to the PM, which then orders the feasible service configurations based on achievable user perceived quality into a so-called *degradation path* from the highest to the lowest quality configuration. Establishment of a degradation path is determined by user preferences (e.g. a user considers audio to be more valuable than video). This is used when service degradation or upgrading is necessary. Finally, the service profile corresponding to the highest quality feasible configuration is passed on to the *QoS Optimization Process* (QOP) module.

QoS Optimization Process. The *QOP* is responsible for calculating the optimal resource allocation and application operating point. The goal is to maximize user perceived quality, by combining the user's notion of relative VE object

“importance” and UF based adaptation, and taking into account the following constraints: (1) Requested network resources must be less than or equal to authorized resources; (2) Allocated resources must fall within threshold values for service requirements across all VE objects; (3) Total price of allocated resources must be less than or equal to specified user budget; (4) The configured operating point must satisfy negotiated parameters (terminal capabilities). The function to be maximized is a linear combination of UFs multiplied by WFs defined across all VE objects. A detailed description of the objective function and constraint formulation is given in [12]. After calculation, the *QOP* passes the final profile (via the *QNC* module) specifying the calculated application operating point and required resources to the *QNC* module of the *Application Server*.

3.2 Application Server

The *Application Server* is responsible for retrieving and adapting stored VR service content based on the negotiation and calculation conducted by the *Access and Control* entities. The *QNC* module receives the final profile and initiates resource reservation mechanisms according to it. The final service profile is then passed to the *VR service processor*, which retrieves the service content from the VR service repository (possibly modified by the *VR service transcoder*), and forwards the content to the end user.

A user’s interest in VE objects is subject to dynamic changes. A change in user’s LoI, a user interaction, or a change detected by the application cause an event to be passed from the application to the *Event Manager* (EM). The *EM* then determines whether the change is “significant” enough to require recalculation of the optimal application operating point and reallocation of network resources to meet new QoS requirements. It maps the event to an LoI for a given VE object, and sends to the *QOP*, which consults the *user interest adaptation policy* (of the service profile) and updates object profiles.

3.3 Session establishment and modification

The process of initial session establishment is shown in Fig. 4. At the top of the diagram we see the modules of the proposed model involved in the process. For multimedia services, the model is not tied to any particular session setup protocol, but they could be mapped, for example, to SIP messages such as INVITE, SESSION PROGRESS, PRACK, OK, ACK, etc. We outline three general scenarios supported by the proposed model involving optimization and renegotiation procedures invoked during the course of the service lifetime, due to changes in: (1) resource availability/cost; (2) client profile (preferences, terminal capabilities, access network); and (3) service requirements (object profiles).

Scenario 1: Change in resource availability/cost

Procedures related to a change occurring in resource availability are shown in Fig. 5. When available bandwidth decreases (or cost increases), the *QOP* runs the optimization process and searches for an optimal operating point and resource

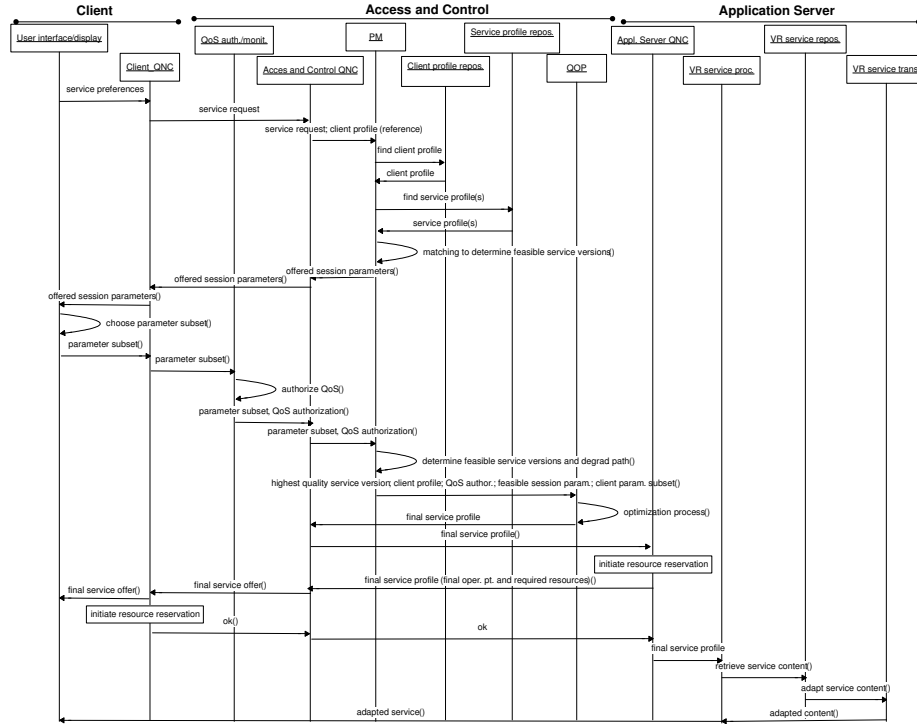


Fig. 4. Initial VR session establishment.

allocation (arrow 3). As long as no solution is found within given constraints, service degradation is requested, based on the previously defined degradation path, and the optimization process is run (arrows 4–6). Once a solution is found, the new final service profile is sent to the QNC module. A session description update is sent to the client indicating newly determined session parameters, and the final service profile is passed to the Application Server responsible for updating session parameters accordingly. If no solution is found, the client is informed of session establishment failure. In case of increased resource availability (or, cost decrease), the QOP sets the current version to the highest quality version in the degradation path and then runs the optimization process again.

Scenario 2: Change in client profile

The client profile may change due to dynamic changes in the access network, in software configurations (e.g. new codec), in terminal resource usage, in user preferences, as well as in the terminal itself. In addition, a service or network provider may update client profile parameters such as those relating to service subscription data. The signaling related to changes in client profile is shown in Fig. 6. The PM conducts a new matching process to determine the feasible session parameters (arrow 3). If the new client profile parameters and the old client profile parameters yield the same matching results, session data is updated

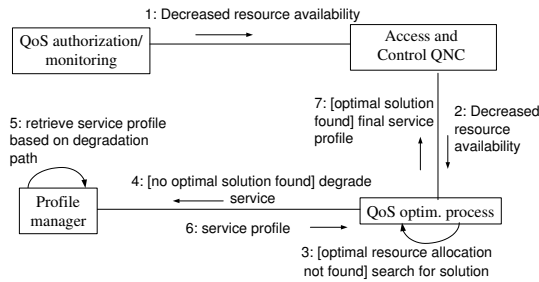


Fig. 5. Change in resource availability.

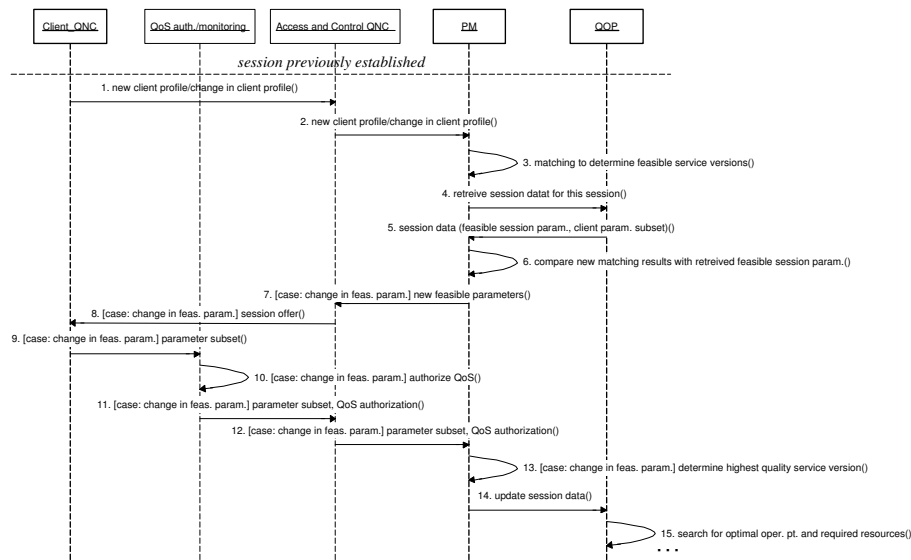


Fig. 6. Change in client profile.

with the new client profile and the optimization process is invoked (arrows 14–15). Procedures following this process correspond to those already shown in Fig. 4. On the other hand, if the result of the new matching process is not the same as the one obtained from the previously conducted matching, this means that a change has occurred in the media parameters that are considered feasible for this session (For example, due to decreased bandwidth in the client access network, video streaming may no longer be supported.) In this case, negotiation procedures are invoked (arrows 7–13).

Scenario 3: Change in service requirements

Changes in service requirements, such as change of an object's UF (LoI), or changes in an object's assigned WF, are signaled to the QOP by the EM module of the Application Server, as shown in Fig. 7 (arrows 2–6). The QOP updates

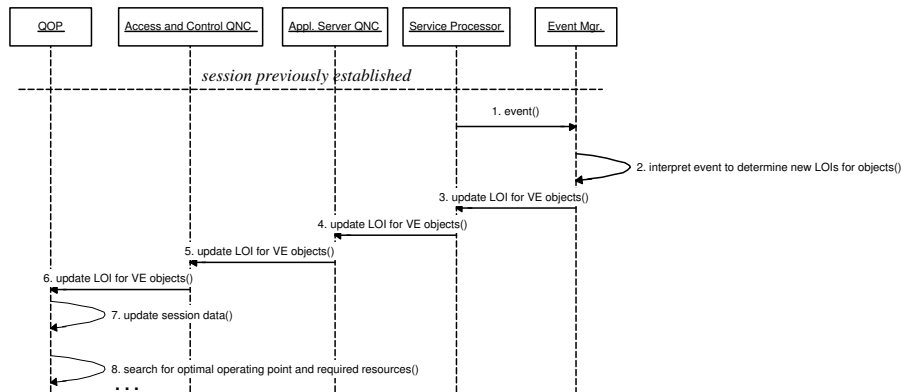


Fig. 7. Change in service requirements.

the necessary object profiles in the corresponding session data (arrow 7), and re-formulates the objective function based on the current UF states and WFs (arrow 8). Finally, it should be noted that re-calculation of the optimal operating point at each and every change in constraining factors is not desired (nor viable). Rather, thresholds may be established indicating modifications “important enough” to do so. Determining such thresholds, and effects thereof onto user perceived value, requires extended study and is considered for future work.

4 Case study: mapping to IP Multimedia Subsystem

While the discussed model is independent of the particular network, we conduct a case study to determine model applicability in UMTS networks by mapping it onto the IMS, as shown in Fig. 8. The IMS session establishment begins with an IMS terminal, shown as User Equipment (UE) (e.g. mobile phone, PC), obtaining access to the IMS through an access network (e.g. GPRS). The next steps involve the allocation of a Proxy-Call Session Control Function (P-CSCF) to serve as an outbound/inbound SIP proxy, and SIP application level registration to the IMS network. The P-CSCF also interfaces with a Policy Decision Function (PDF) which authorizes the use of bearer and QoS resources within the access network [2]. The central node of the signaling plane that is responsible for session establishment, modification, and release is the Serving-CSCF (S-CSCF), located in the user’s home network. The S-CSCF interrogates the Home Subscriber Server (HSS) to access user profile information, fetch subscription data, and for authentication/authorization/accounting purposes. Additionally, the S-CSCF plays an important role in service provision by invoking one or more Application Servers (AS). The (IMS) Application and Services domain hosts both application/content, and reusable common functions that provide value-added services (e.g. messaging, presence). Various other ASs (e.g. gaming server) may

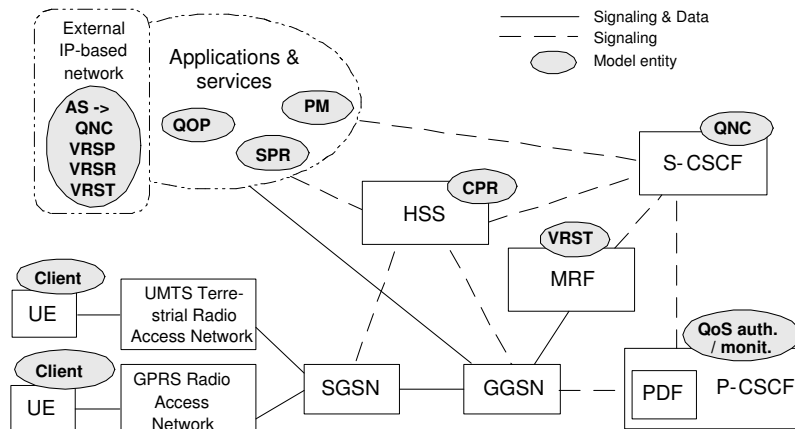


Fig. 8. Mapping of model entities to the IMS architecture.

be invoked along the signaling path. Having briefly described the IMS, we now present the mapping of the proposed model functionality onto it.

The *Client* is located in the *User Equipment* (UE). We map the functionality of the *Access and Control QoS authorization/monitoring* entity to the P-CSCF and the PDF. The *Access and Control QNC* entity is mapped to the S-CSCF. The *client profile repository* (CPR) storing user profile information is considered to be a part of the HSS. We place the *VR Service Profile Repository* (SPR) in the Application and Services domain, where it is interrogated by the S-CSCF. The *PM*, responsible for matching client profile parameters with service requirements, and the *QOP*, which calculates the optimal service operating point and resource allocation, are both mapped to a SIP AS in the Application and Services domain. With the specification of generic client and service profiles, the implementation of these functions is independent of the actual application content being delivered to the end user. From an IMS perspective, introducing such enhanced QoS support in the network, as a generic “reusable service”, could lead to quicker time-to-market for new services, with service providers only being required to provide the network with a service profile stating service requirements. The network would then take care of determining the QoS parameters that would maximize user perceived quality. The *Application Server* (AS) is composed of four modules: *QNC*, *VR service processor* (VRSP), *VR service repository* (VRSR), and *VR service transcoder* (VRST). These modules may reside in the IMS network or on a server in an External IP Based Network (ExIPBN). Transcoding functionality of the *VRST* module may be mapped to the IMS *Media Resource Function* (MRF), responsible for the manipulation of multimedia streams. The proposed mapping indicates possible upgrades to some mentioned IMS entities, but also in most cases identifies existing functionality that should be used. The main enhancement to IMS would be the AS providing advanced QoS support in a reusable way. In addition, our approach stresses the need for standardization of QoS parameter specification in the form of client and service profiles.

5 Conclusions and Future Work

In this work, we presented a proposed QoS control model supporting the signaling, negotiation, and adaptation of QoS for multimedia services, intended for use in NGNs, and mapped onto the 3GPP IMS. Two issues may be identified as future work. The foremost is the question of scalability, considering the overhead resulting from QoS related signaling, both in terms of cost and time. The second important issue for further study is security. With regards to scalability issues relating to model deployment, it is clear that for a large number of users, running the optimization procedure separately for each user and when dynamic changes occur is definitely time consuming and costly. Besides the establishment of re-calculation thresholds, a solution may be to offer a set of discrete solutions calculated in advance for particular combinations of constraints. Investigation of such issues is considered for future work. A key requirement is for dynamic service adaptation and QoS renegotiation to occur with minimal user perceived disruptions.

References

1. –, 3GPP TS 23.228: IP Multimedia Subsystem (IMS); Stage 2. June (2005)
2. –, 3GPP TS 23.803: Evolution of Policy Control and Charging. September (2005)
3. –, ITU-T Recommendation Y.2001: General Overview of NGN. December (2004)
4. –, ITU-T Recommendation H.360: An architecture for end-to-end QoS control and signalling. March (2004)
5. Araniti, G., De Meo, P., Iera, A., Ursino, D.: Adaptively Controlling the QoS of Multimedia Wireless Applications through ‘User Profiling’ Techniques. *IEEE J. on Selected Areas in Communications*, Vol. 21, No. 10, Dec. (2003) 1546-1556
6. Gao, X., Wu, G., Miki, T.: End-to-end QoS Provisioning in Mobile Heterogeneous Networks. *IEEE Wireless Communications Magazine*, June (2004) 24-34
7. Handley, M., Jacobson, V.: SDP: Session Description Protocol. IETF RFC 2327 (1998)
8. Guenkova-Luy, T., Kessler, A. J., Mandato, D.: End-to-End Quality-of-Service Coordination for Mobile Multimedia Applications. *IEEE J. on Selected Areas in Communications*, Vol. 22, No. 5, June (2004) 889-903
9. Prior, R., Sargento, S., Gomes, D., Aguiar, R. L.: Heterogeneous Signaling Framework for End-to-end QoS support in Next Generation Networks. Proc. of *38th Hawaii International Conf. on Systems Sciences* (HICSS-38 2005), CD-ROM/ Abstracts Proceedings, Big Island, HI, USA January (2005)
10. Rosenberg, J., et al.: SIP: Session Initiation Protocol. IETF RFC 3261 (2002)
11. Skorin-Kapov, L., Mikic, D., Vilendecic, D., Huljenic, D.: Analysis of end-to-end QoS for networked virtual reality services in UMTS. *IEEE Communications Magazine*, Vol. 42 No. 4, April (2004) 49-55
12. Skorin-Kapov, L., Matijasevic, M.: Dynamic QoS Negotiation and Adaptation for Networked Virtual Reality Services Proc. of *IEEE WoWMoM 2005*, Taormina, Italy, June (2005) 344-351
13. Wang, X., Schulzrinne, H.: An Integrated Resource Negotiation, Pricing, and QoS adaptation framework for multimedia applications *IEEE J. on Selected Areas in Communications*, Vol. 18, No. 12, Dec. (2000) 2514-2529