End-to-End Quality of Service Coordination Models for Mobile Networks

TEODORA GUENKOVA-LUY
Distributed Systems Department
Faculty of Computer Science, University of Ulm
O-27, Oberer Eselsberg, 89069 Ulm, Germany
guenkova@vs.informatik.uni-ulm.de

ANDREAS KASSLER
SCE, Nanyang Technological University
Academic Complex North N4-02a-32
Nanyang Avenue, Singapore 639798
kassler@ieee.org

Abstract: - Providing End-to-End QoS in 4G heterogeneous networks is a challenge due to the involvement of several components like network reservation, service level agreements, charging, contract management and security constraints. A flexible yet efficient protocol framework is required, which can integrate all this functionality. Recently, we proposed the End-to-End Negotiation Protocol (E2ENP) for negotiating and orchestrating QoS on an end-to-end basis both at application and network layer. Based on an flexible XML model extending SDPng concepts, the protocol enables the negotiation of system capabilities. In this paper, we show how third-party service-providers can effectively influence the negotiation process. We also discuss authentication and security features of E2ENP.

Key-Words: - QoS, E2ENP, XML, SDPng, SIP, A4C, Security

1 Introduction

Current protocols and mechanisms for supporting end-to-end QoS cover predominantly single facets of the global QoS management. For example, the Real-Time Protocol suite (RTP/RTCP) [1] is used for multimedia data transfer and QoS feedback, Resource Reservation Protocol (RSVP) [2] allows to reserve network resources, Common Open Policy Service (COPS) [3] manages policy exchange between routers on the network, RADIUS [4] and DIAMETER [5] enable authentication of QoS-aware services, etc. However, QoS is a system aspect that crosses all components and layers of a distributed multimedia system. Hence, QoS-management models should incorporate possibilities to support: multiple QoS-representations; mapping between different QoS classes and types of QoS parameters; and orchestration between various system facets.

In wireless environment, QoS management should be handled effectively when session and device mobility is involved, because handover or resource-availability changes may result in QoS violations. Signaling of QoS-relevant and mobility events must be very efficient to minimize service disruption in such scenarios.

This paper discusses presentation/orchestration features of the End-to-End Negotiation Protocol (E2ENP) for capabilities and QoS [6][7][8][9][10]. We introduce the E2ENP concepts for global multimedia application/session management, including features like authentication and security. E2ENP uses Session Initiation Protocol (SIP) [11] to transfer control data. A description model based on Extensible Markup Language (XML) [12] is applied to specify system characteristics and QoS parameters based on enhancements for Session Description Protocol new generation (SDPng) [13]. We discuss problems of managing QoS within different types of networks and present models for applying end-to-end QoS and system-resource coordination, in consideration with authentication and security.

2 Application Management Models for QoS – An Overview

We introduce a reference model (Fig. 1) to distinguish between two important features: packet forwarding and service provisioning. This model is used to analyze different roles concerning QoS management within a global network architecture and serves as a basis for defining a strict differentiation between QoS-relevant system facets. The roles of the different devices (e.g. terminal, router, security/authentication services, etc.) are characterized via a separation into two domains, namely Service Domain and Transport Domain.

Transport Domain is mainly responsible for QoS-aware packet forwarding and contains all management functions that provide QoS for IP packets. Examples are resource reservation via RSVP [2], packet marking and filtering in DiffServ domains [14], etc. The Service Domain manages the service aspects of multimedia service like negotiation of session descriptions and capabilities for session setup; Authentication, Authorization, Accounting, Auditing, Charging (A4C); security features of a service, etc. We introduce also a User Domain in order to complete the picture of the QoS-relevant system parameters.

Within this reference model (Fig. 1) a service-abstraction for the application is provided by the end-system resource management, which is typically implemented within terminal's operating system (OS) or middleware and is denoted here as End-system QoS-

Management (EQM). EQM informs the application about available resources and capabilities to handle multimedia data (type of audio/video hardware devices, multimedia codecs, etc.) and for controlling network transport (type of network devices, currently available access networks and signal quality thereof, etc.). The application can flexibly (re-)configure itself in accordance with the available resources (both locally and on the network) by using monitoring data provided by the resource management (i.e. the monitoring is performed within the components for Network Management, NM, and is delivered to the local management EQM). NM nodes are also responsible for controlling the admission to network resources in accordance with: provider policies (originally defined over the Service Management (SM) components of the system), results of monitoring of network resources, packet-routing rules (which may also be specified from SM), etc. SM and NM represent the implementation of the management separation in terms of service management (i.e. SM) and packet forwarding (i.e. NM). Within the network SM can participate actively or passively in negotiation of end-systems when establishing or adapting a multimedia session by listening in and/or actively participating in the negotiation process. SM can instruct the NM entities how to apply provider policies by uploading such policies in the NM(s), e.g. SM can reserve network resources on behalf of those end-system which cannot or by definition are not allowed to reserve resources on the network or in a DiffServ domain [14] SM can instruct the bandwidth broker (i.e. a concrete implementation of a NM) how to perform packet-marking and -filtering.

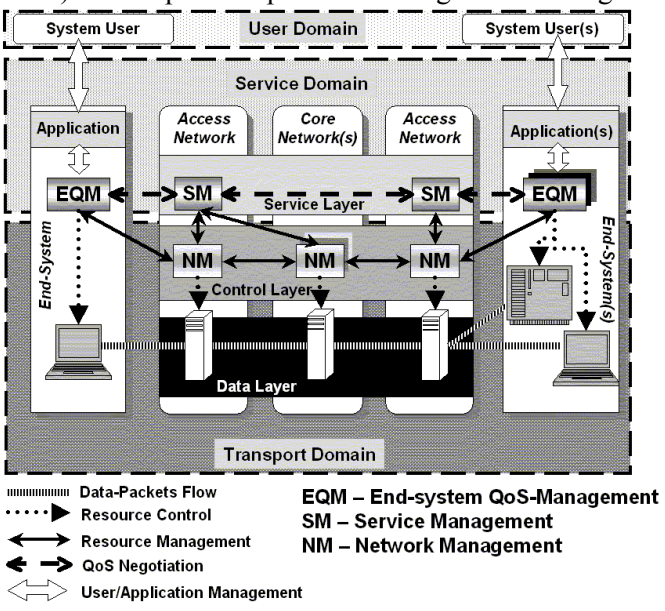

Fig. 1 *A Model for End-to-End Resource Management*

Thus within this model several abstraction levels of QoS-parameters are identified (see [15] and [16] for comparison):

End-to-End perceivable QoS parameters – This set of parameters corresponds to user's perception of the performance of the application. Such parameters, used within the User Domain, guide the users to specify QoS in their natural understanding, e.g. via a Graphical User Interface (GUI). Translation of perceivable QoS characteristics in more technical terms is typically implemented inside the Service Domain of the system.

Application QoS parameters are used to describe end-to-end application performance in accordance with software and hardware resources of end-systems/services. Compliance of Application QoS-parameter with Service Contract specifications can be evaluated within the network by a *Service Layer* within the Service Domain (see Fig. 1). Here, a service abstraction with respect to system access-management, policy management, definition of service level agreements (SLA) can be used for service-level admission control.

Transport QoS parameters are used to describe end-to-end requirements with respect to network resources like bandwidth, delay, jitter, etc. The definition and evaluation of Transport QoS parameters should be in-line with the monitoring data and the specific transport-handling mechanisms provided by the Transport Domain. These mechanisms within the Transport Domain are grouped into two functional layers – *Data Layer* (responsible for QoS-aware packet forwarding, traffic shaping/policing/classification, buffer management, etc.) and *Control Layer* (responsible for router access, reservation mechanisms and resource-admission control).

In addition to all the parameters described above, end-systems must specify connection points for communication (e.g. IP-addresses and ports) and other presentation capabilities for the media (e.g. media packetization rules) in order to establish a valid end-to-end multimedia session.

Within the IST-MIND project [7] theoretical considerations and scenarios were developed involving the management model (Fig. 1), A4C and security system-features. The application of E2ENP for such cases was discussed in consideration with four interaction scenarios. In Section 4.2 we present a generalized model that supports and extends the scenarios developed in [7].

3 General E2ENP Features

In order to enable optimized QoS negotiations for different QoS abstraction levels and for various types of networks a well-structured description model for QoS is necessary. Here, we introduce the E2ENP concept for optimizing the session description format and the negotiations of multimedia session parameters.

3.1 Hierarchical QoS Specification

Mobile multimedia applications typically manage multiple media types (e.g. audio, video, data). Corresponding media streams can be logically grouped based on various criteria and dependencies. Hence, the overall QoS specification can be modeled in a hierarchical manner therewith capturing time synchronization, QoS correlation, and resource-constraints aspects among streams.

A hierarchical model (Fig. 2) allows designing alternative QoS specifications, which the application needs for automatically and efficiently adapting its resource usage with respect to the current resource availability and user's expectations. The E2ENP model consists of a tree of QoS specifications. The leaves of the tree define the adaptation alternatives in form of single QoS specifications, termed "QoS Contract"-s; a parent node defines the adaptation behavior of a single stream; a further ancestor node specifies the abstraction of stream-association adaptation behavior (termed "QoS Context"), and so on.

Any given sub-tree originating from a specific branch node of the hierarchical QoS specification is associated with an adaptation-rule predicate. The resolving of this predicate selects a child node and hence instructs the system to enforce the QoS specification associated with that child. For instance (see Fig. 2): "if Video Parameters V11 are not longer enforceable, switch to Video Parameters V12" or "if Stream Association 1" (Video and Audio) is not supportable (e.g. due to handover and thus lower data-rate availability) switch to "Stream Association 2" (only Audio)". It is up to the adaptive application and its business logic to determine when to adapt, how to adapt and to what extent to adapt. E2ENP provides only a description model for such adaptation events and a signaling mechanism so that peers can agree on adaptation conditions and prepare resources correspondingly. Further details on the E2ENP hierarchical QoS specification can be found in [7][9].

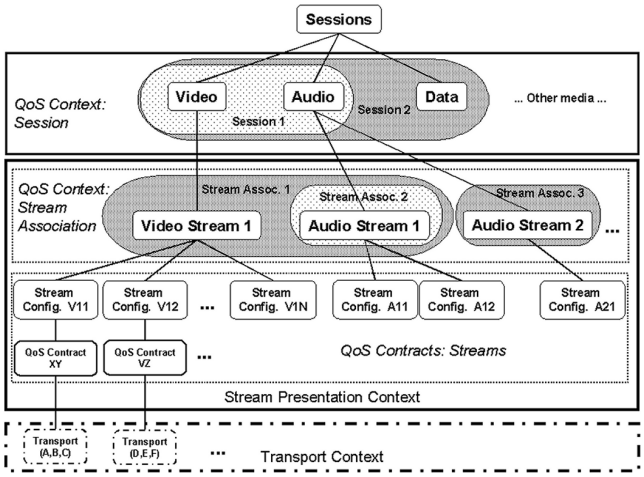

Fig. 2 *Example of Hierarchical QoS Specification*

3.2 Referencing Mechanism for Control Data

In current systems, a complete set of session-configuration information is exchanged every time a session is negotiated or adapted [13][17][18]. This scenario is not efficient as the minimization of signaling traffic is crucial, especially within the area of access networks, where the signaling and the data traffic share bandwidth. Though E2ENP applies ideas stemming from [13][17][18], E2ENP defines in addition an optional referencing mechanism to minimize the amount of control signaling, using the well-structured model of the hierarchical QoS specification (Section 3.1). The E2ENP QoS-coordination process is separated in three different phases corresponding system configuration (*pre-negotiation*), session establishment (*negotiation*) and session adaptation (*re-negotiation*) [6][7]. These phases are used to successively convey information for building a QoS-coordination document (see Section 4.1). The E2ENP design considers the fact that some of the QoS configurations are valid for multiple multimedia sessions and some of the parameters are applied only to a single session. E2ENP associates the multiply applicable information bits with identifier-keys in order to be able to reference every such bit at later time. E2ENP allows the enhancement or overwriting of the QoS-coordination document, whenever configuration changes in the system occur (e.g. application of new codecs due to a software download). Consequently, we differentiate between *short* negotiation/re-negotiation, which use keys to previous E2ENP phases, and *full* negotiation/re-negotiation, which exchange complete configuration descriptions, irrespective of the previous E2ENP phases [6][7]. The application of *short* negotiation-procedures can significantly reduce the signaling load (see [9] for measurement examples). E2ENP defines as its default carrier the Session Initiation Protocol (SIP) [11] and the negotiation procedures of E2ENP based on SIP can be found in [7] together with the E2ENP payload-structure.

4 Application Model for E2ENP

4.1 Generation, Validation, Authentication and Security Models for E2ENP

Heterogeneous end-systems negotiate QoS in form of QoS configurations to agree upon common set of supportable parameters for session establishment and adaptation. Any QoS configuration used in the end-to-end QoS coordination process should be validated against user's QoS preferences, end-system resources and provider-specific restrictions (like user contracts or resource availability in providers transport domain). E2ENP introduces a validation procedure (Fig. 3) as a requirement to the application/middleware using E2ENP

to guarantee generation of correct QoS-management information. Within this procedure the application maps Application QoS parameters to local-system capabilities, building one or several alternative QoS configurations, which can also be associated with Transport QoS parameters in accordance with the ability of the end-systems to reserve network-resources on their own. The QoS configurations are then validated against provider policies either at the local system (if policies are downloaded during registration), or inside the network within the QoS negotiation process (see [10] for a detailed example). This depends on the trust models between end-systems and access-network providers, and on the ability of the user-terminals to perform network reservations [7]. E2ENP requires [6] that policy validation does not change the contents of the QoS configurations (in contrast to some SIP-based specifications, where stateful proxies can remove certain configurations [19]). E2ENP explicitly includes descriptions of provider constraints associated with the QoS configurations (see Fig. 4), as a result of this valid QoS configurations are built. The valid QoS configurations describe a complete technically supportable QoS range, thus the end-systems are able to change the access-network technology and/or provider during an ongoing session by planning ahead appropriate management actions [6][7].

QoS configurations in E2ENP are modeled in XML [12]. Every QoS configuration or some parts thereof are associated with a key (see XML-link technology [20]). All relevant QoS configurations exchanged by applications build a QoS-coordination document, used by the communicating end-systems to establish and manage multimedia sessions. During ongoing multimedia sessions, applications can refer to the appropriate QoS-coordination document exchanging only negotiated ahead keys to the information in the QoS-coordination document [7][9].

In order to provide authentication for the E2ENP information and to prove that the end-to-end exchanged information is not being changed along the way between two end-systems, an end-system can secure the information using digests (Fig. 4). In this case the information is open and readable by the access provider, but if the provider or some malicious third party tries to intrude with the E2ENP information the digest would become invalid. Additionally, if the provider knows the key for generating the end-system digests, the provider can prove the authenticity of the E2ENP information. The exchange of keys for identifying the E2ENP information can be done by some standard key-management software and services on the Internet. In order to constrain the end-to-end exchanged control-information the provider references the QoS-information using specific E2ENP identifiers (see also Section 3.2), thus the provider entities do not interfere with the E2ENP information, but only enhance it without invalidating the end-system digest. The providers can also secure their information with digests, in a similar to the end-systems way. Thus, it is possible to apply provider-to-end-system authentication and provider-to-provider authentication, i.e. the end-systems can also verify the provider digest to recognize malicious third-party intruders or providers can prove the authenticity of their adjacent provider on the boundaries between two provider-domains, like between access-network and core-network providers. The end-systems and provider entities can exchange digest keys, e.g. in a registration procedure. Furthermore, the usage of certificate keys for authenticating control information can also be applied within the resource management process on the network. The provider and the end-systems can exchange keys within the signaling process, which are used for the reservation. At registration/negotiation the end-systems may receive reservation certificates from the provider and attach them to the reservation payloads to authenticate them (e.g. within RSVP).

As E2ENP is a meta-protocol carried by another session protocol, E2ENP can conveniently apply the security techniques and mechanisms provided by its carrier, e.g. SIP security [11].

4.2 End-system and Provider Interactions

QoS-aware services are to some extent also related to A4C services [21][22], as usually one has to pay more for a premium service. Hence, service/session/QoS signaling can interact with and be integrated in the A4C and security signaling. The interactions with A4C and security services within a QoS-aware architecture occurs at two levels (see also Section 2):

- Transport Domain interactions – These interactions correspond to network authentication using AAA protocols like RADIUS [4], or DIAMETER [5] and give the user-terminals the possibility to join a provider-owned network.
- Service Domain interactions – These interactions correspond to access management at service level with service authentication via SIP DIAMETER [23], SIP authentication and security management [11], etc., which give the user the possibility to request personalization and differentiation for his/her service presentation in comparison with other system users.

We distinguish between two roles within the QoS negotiation and coordination process (see also Fig. 1):

End-system – can interpret user's QoS requirements. It can represent and enforce these requirements in terms of Application QoS parameters for every application on the terminal,

Access Provider Network (APN) – validates Transport and/or Application QoS parameters in an admission-

control procedure (during end-system registration and/or during session establishment) and calculates provider QoS restrictions for the application, using provider policies. These restrictions are applied within the validation process (Section 4.1). The APN incorporate both the Service and Network Management of the distributed multimedia system (Section 2). The Service Management in APN corresponds to actions like registration, service negotiation, service authentication and authorization, etc. The Network Management in APN includes functionalities like network reservation, network-packet management, router configuration, accounting and charging for network flows, etc.

Fig. 5 depicts a general view of a negotiation procedure with E2ENP in consideration with internal network services (Only the logical meaning of the interactions is shown. The exact negotiation with SIP-specific methods in accordance with different negotiation scenarios and alternative reservation possibilities can be found in [7]). During a registration procedure (Fig. 5 – Actions 1 - 6), APN(s) could provide end-systems with validation policies so that end-systems can validate QoS configurations by themselves. In this example, however, the APN(s) validates later. The registration procedure can be applied in two steps – first: registration to gain network access over pure DIAMETER services [5] and second: registration to gain details on the service support over SIP REGISTER [11] or over a combination of hybrid approaches (e.g. DIAMETER/SIP [23]). The end-systems exchange QoS relevant information in a negotiation procedure (Messages 7 - 17). The respective QoS-configuration requests and responses are validated and enhanced with transport QoS (e.g. bandwidth, delay, etc.) and policy parameters (e.g. max bandwidth, max number of streams per applications, etc.) by the corresponding APN(s) (Actions 8, 10, 14 and 16). When the response is received, the initiator of the signaling phase can calculate the final QoS configuration (Action 18), and it generates a general system configuration (during the pre-negotiation phase) or a particular session configuration (during negotiation phase). During session setup (negotiation) or adaptation (re-negotiation) the notification of the final QoS Configuration (Message 19) can also serve as a trigger for network resource-reservation (e.g. via RSVP). End-systems need to explicitly coordinate their resource reservations as different underlying reservation mechanisms might be in use (e.g. RSVP [2] might be terminated at a network gateway within APN1 and/or APN2 might be pure DiffServ [14]). Therefore, E2ENP informs the peers about the state of network resource-reservation (Messages 19 - 28). During this resource-coordination process both end-systems and the APN(s) save the final QoS configuration in order to understand the keys to the QoS descriptions that the end-systems exchange during the next E2ENP phases. Thus the APN(s) are also able to assist the end-systems in their network reservations. When listening for the Resource-Reservation-Coordination message and interpreting this message, the APN(s) can reserve resources on the network on behalf of those end-systems, which cannot do this by themselves. When the answer to the Resource-Reservation-Coordination message is interpreted, the APN(s) can start the billing process for the reserved resources within the network.

Due to the flexibility of the E2ENP description model, it is possible to define and control different signaling scenarios, e.g. with end-system reservation signaling, with provider-initiated reservations, with specific scenarios for initiating metering and charging processes, etc. (see also [7][9][10]).

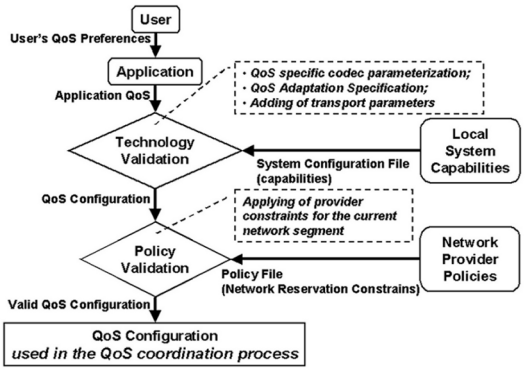

Fig. 3 *Generation and Validation of Relevant QoS Configurations*

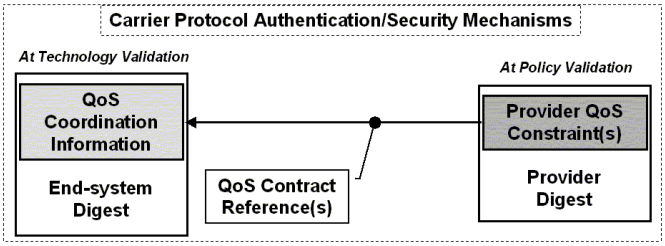

Fig. 4 *Authentication and Security for End-to-End QoS Configurations*

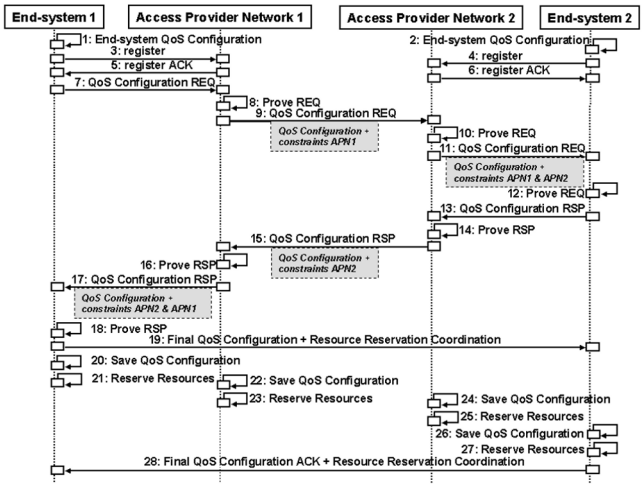

Fig. 5 *End-system and APN(s) interactions*

5 Conclusions

Within this paper, we introduced a novel approach for end-to-end QoS coordination, which allows end-systems and access-network providers to cooperate in mobile networks and to flexibly configure multimedia services, including protocol authentication and security features. The usage of E2ENP and its information referencing mechanism enables end-systems and APN-proxies to speed up the procedures for multimedia-session establishment and adaptation by efficiently applying known (from previous signaling sessions) system configurations. The current work on E2ENP includes further definitions of relevant E2ENP elements for security and authentication using XML. Additionally, further studies are carried out to enable the E2ENP application not only for device mobility, but also for session mobility between different devices. This would require the introduction of third-party call-control functions within E2ENP. Considering the E2ENP well-structured description model and its negotiation flexibility, we expect that the E2ENP concept will smoothly combine with legacy applications, along the path towards the future of QoS-aware multimedia services.

References:

[1] H. Schulzrinne et. al., "RTP: A Transport Protocol for Real-Time Applications", IETF RFC 1889, January 1996

[2] A. Mankin et. al., "Resource ReSerVation Protocol (RSVP)", IETF RFC 2208, September 1997

[3] D. Durham et. al., "The COPS (Common Open Policy Service) Protocol", IETF RFC 2748, January 2000

[4] C. Rigney et al., "Remote Authentication Dial In User Service (RADIUS)", IETF RFC 2865, June 2000

[5] Pat R. Calhoun et al., "Diameter Base Protocol ", IETF Internet-Draft: draft-ietf-aaa-diameter-16.txt, December 2002

[6] T. Guenkova-Luy, A. Kassler, J. Eisl, D. Mandato, "Efficient End-to-End QoS Signaling - concepts and features", IETF Internet-Draft (draft-guenkova-mmusic-e2enp-sdpng-00), March 2002

[7] MIND Project, "Top-level architecture for providing seamless QoS, security, accounting and mobility to applications and services", Deliverable D1.2, November 2002

[8] P. Ruiz, J. Sanchez, E. Garcia, A. Gomez-Skarmeta, J. Botia, A. Kassler, T. Guenkova-Luy, "Adaptive Multimedia Multi-party Communication in Ad Hoc Environments", HICSS-37, Software Technology Track, January 2004

[9] T. Guenkova-Luy, A. Kassler, D. Mandato, "End-to-End Quality of Service Coordination for Mobile Multimedia Applications", to appear in IEEE Journal on Selected Areas in Communications, issue on Advanced Mobility Management and QoS Protocols for Wireless Internet, 2004

[10] T. Guenkova-Luy, A. Kassler, "End-To-End Quality of Service Coordination for Multimedia Applications in Heterogeneous, Mobile Networks", to appear in proceedings of IEEE International Conference on Communications, June 2004

[11] J. Rosenberg et. al., "SIP: Session Initiation Protocol", IETF RFC 3261, June 2002

[12] W3C, "Extensible Markup Language (XML) 1.0 (Second Edition)"

[13] D. Kutscher et. al., "Session Description and Capability Negotiation", IETF Internet-Drafts (draft-ietf-mmusic-sdpng-06 and draft-ietf-mmusic-sdpng-07), March/October 2003

[14] S. Blake et. al., "An Architecture for Differentiated Services", IETF RFC 2475, December 1998

[15] Xiaohui Gu, K. Nahrstedt et. al., "An XML-based Quality of Service Enabling Language for the Web", Project Report - National Science Foundation, 2001

[16] H. Lu, I. Faynberg, "An Architectural Framework for Support of Quality of Service in Packet Networks", IEEE Communications Magazine, pp. 98-105, June 2003

[17] J. Rosenberg, H. Schulzrinne, "An Offer/Answer Model with SDP", IETF RFC 3264, June 2002

[18] G. Camarillo et. al., "Integration of Resource Management and Session Initiation Protocol (SIP)", IETF RFC 3312, October 2002

[19] 3GPP TS 24.228 V5.1.0, "Signalling flows for the IP multimedia call control based on SIP and SDP", Technical Specification, June 2002

[20] W3C, "XML Linking Language (XLink) – Recommendation Version 1", June 2001

[21] G. Huston, "Next Steps for the IP QoS Architecture", IETF RFC 2990, November 2000

[22] C. Mills et al., "Internet Accounting: Background", IETF RFC 1272, November 1991

[23] M. Garcia-Martin et al., "Diameter Session Initiation Protocol (SIP) Application", IETF Internet-Draft: draft-ietf-aaa-diameter-sip-app-00.txt, October 2003