2012

Quality of Service optimisation framework for Next Generation Networks

Weber, Frank Gerd

http://hdl.handle.net/10026.1/1199

http://dx.doi.org/10.24382/4661
University of Plymouth

Copyright Statement

Quality of Service optimisation framework for Next Generation Networks

by

Frank Gerd Weber

A thesis submitted to University of Plymouth
in partial fulfilment for the degree of

Doctor of Philosophy

School of Computing and Mathematics

In collaboration with
Darmstadt Node of the CSCAN Network

September 2012

Quality of Service optimisation framework for Next Generation Networks

Frank Gerd Weber

Abstract

Within recent years, the concept of Next Generation Networks (NGN) has become widely accepted within the telecommunication area, in parallel with the migration of telecommunication networks from traditional circuit-switched technologies such as ISDN (Integrated Services Digital Network) towards packet-switched NGN. In this context, SIP (Session Initiation Protocol), originally developed for Internet use only, has emerged as the major signalling protocol for multimedia sessions in IP (Internet Protocol) based NGN.

One of the traditional limitations of IP when faced with the challenges of real-time communications is the lack of quality support at the network layer. In line with NGN specification work, international standardisation bodies have defined a sophisticated QoS (Quality of Service) architecture for NGN, controlling IP transport resources and conventional IP QoS mechanisms through centralised higher layer network elements via cross-layer signalling.

Being able to centrally control QoS conditions for any media session in NGN without the imperative of a cross-layer approach would result in a feasible and less complex NGN architecture. Especially the demand for additional network elements would be decreased, resulting in the reduction of system and operational costs in both, service and transport infrastructure.

This thesis proposes a novel framework for QoS optimisation for media sessions in SIP-based NGN without the need for cross-layer signalling. One key contribution of the framework is the approach to identify and logically group media sessions that encounter similar QoS conditions, which is performed by applying pattern recognition and clustering techniques. Based on this novel methodology, the framework provides functions and mechanisms for comprehensive resource-saving QoS estimation, adaptation of QoS conditions, and support of Call Admission Control. The framework can be integrated with any arbitrary SIP-IP-based real-time communication infrastructure, since it does not require access to any particular QoS control or monitoring functionalities provided within the IP transport network.

The proposed framework concept has been deployed and validated in a prototypical simulation environment. Simulation results show MOS (Mean Opinion Score) improvement rates between 53 and 66 percent without any active control of transport network resources.

Overall, the proposed framework comes as an effective concept for central controlled QoS optimisation in NGN without the need for cross-layer signalling. As such, by either being run stand-alone or combined with conventional QoS control mechanisms, the framework provides a comprehensive basis for both the reduction of complexity and mitigation of issues coming along with QoS provision in NGN.

Contents

List of Figures

List of Tables

Acknowledgements

In the first place I wish to express my sincere thanks to my supervisors Prof. Woldemar Fuhrmann, Prof. Udo Bleimann, and Dr. Bogdan Ghita for their kind and helpful support and guidance throughout this research.

I wish to express my deep thanks to my supervisor Prof. Ulrich Trick for the wide range of support offered to me throughout my time as a member and friend of the Research Group for Telecommunication Networks at FH Frankfurt a. M., University of Applied Sciences. Prof. Trick has not only aroused my interest in academic research, but also has provided me with the chance for development and advancement in various professional aspects. This research would probably never have been performed without his encouragement.

Many valuable ideas regarding this research arose from innumerable discussions with my colleagues and friends at the Research Group for Telecommunication Networks at FH Frankfurt a. M. I wish to thank all current and former members of this group for the great inspiration that I have experienced, and for their all-round support, which included, but was not limited to assisting in prototype bug-fixing.

Warm thanks go to the members of both the graduate school and the CSCAN Network at Plymouth University, and special thanks go to the members of the CSCAN Darmstadt node for their experienced support in academic aspects.

I wish to thank my family and friends for their liberating non-technical support offered throughout the rare times that I could spend with them.

Finally, my biggest thanks belong to my loving wife Heike for her great support, understanding, and patience throughout the entire demanding time of this research.

Author’s Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee.

Relevant scientific seminars and conferences were regularly attended at which work was often presented, and several papers were prepared for publication.

Word count of main body of thesis: 52,266

Signed ________________________

Date ________________________

1 Introduction

Within the last years, the concept of Next Generation Networks (NGN) has become widely accepted within the field of both fixed and mobile telecommunications. The conversion of telecommunication networks from traditional circuit-switched technologies such as ISDN (Integrated Services Digital Network) towards packet-switched NGN currently takes place. According to (iLocus, 2010) about 22 percent of all wireline voice telecommunication subscribers worldwide are assumed to have already been migrated from PSTN (Public Switched Telephone Network) to NGN until 2010.

From a technical point of view, packet-switched telecommunication networks, which nowadays are typically based on Internet Protocol (IP) transport, provide crucial and future-proof benefits such as the flexibility to carry a wide range of user services including Internet access, video streaming, and real-time conversational communications such as voice and video telephony. Providing different modes of communication and involving a variety of different media, each service comes with individual requirements such as a defined mean and peak bandwidth, consistency of network performance, network synchronity, or transport reliability. One decisive reason for the broad applicability of NGN is the separation of transport and service functions, which is one of its key characteristics according to (ITU-T Y.2001, 2004) and (ETSI TR 180 000, 2006). From a service provider point of view, a connection-oriented communication service such as a voice or video telephone call only requires the general functionality of user localisation and signalling exchange, independent of the type and individual characteristics of the media (such as voice, video, chat, or desktop sharing) which the users have agreed to be the subject of the respective session. Once a session has been initiated, the media data are exchanged directly via the packet-switched transport network. The NGN service infrastructure typically based on the Session Initiation Protocol (SIP) is not required to be involved in the media data transfer and, hence, its dimensioning and layout is completely media-independent, which is a strong benefit compared to PSTN.

Beside multi-service applicability and separation of service and transport, Quality of Service (QoS) is another main key characteristic of NGN as defined in (ITU-T Y.2001, 2004) and (ETSI TR 180 000, 2006). According to (European Regulators Group, 2008a) the term Quality of Service describes a broad concept, covering all aspects that have an influence on the perception of a user of a service. Beside the transport network performance, which is defined by objective performance parameters such as packet delay, jitter, and loss, QoS also comprises other factors such as the terminal equipment, media codecs, and customer support. According to (European Regulators Group, 2008b), the European Telecommunications Network Operators' Association (ETNO) identified QoS as one of the core factors that will influence the success of NGN.

A crucial challenge faced by QoS management is the control of the network performance, which immediately affects media data transport and, therefore has a strong influence on the QoS perceived by the users, especially regarding real-time conversational media such as voice and video telephony. However, the separation of NGN service and transport infrastructures does not allow for the direct control of transport network resources (and, hence their performance) through the service infrastructure that is aware of communication sessions. Therefore, in order to facilitate service provider-driven QoS control, an intermediary signalling infrastructure, known as the NGN QoS architecture, is typically deployed to interconnect service and transport infrastructures for QoS management. It has been adopted by major telecommunication standardisation bodies in (ETSI TS 185 001, 2005) and (ITU-T Y.2111, 2006). This sophisticated architecture, while reliably fulfilling strict QoS requirements on the one hand, does not only come by the cost of a high volume of additional signalling, but also introduces a high level of complexity and potential scalability issues on the other hand. Although being addressed by several studies within the last years, no satisfactory solution has been described to solve this major issue, which also results in a discrepancy within the concept of NGN itself, challenging three of its main key features: separation of service and transport infrastructures, scalability, and Quality of Service.

This research work has been dedicated to find and describe a novel approach to simplify and alleviate the provision of Quality of Service in NGN. The aims and objectives of this research are presented in section 1.1, followed by an outline on the thesis structure in section 1.2.

1.1 Aims and objectives

The aim of this research is to propose an advanced framework that allows for the optimisation of QoS in SIP-based NGN. This has been achieved through combining methodologies from different fields of research. As a result this research proposes a novel approach for the dynamic adaptation of the utilisation of available network resources. The resulting framework comes as an integrating approach for simplified

QoS provision in NGN while making use of its functional architecture and utilised protocols.

The main objectives of this research can be outlined as follows.

1. To analyse the conventional NGN QoS provision architecture regarding optimisation opportunities, such as the reduction of signalling burden. Subsequently, to define requirements to be fulfilled by alternative NGN QoS provision approaches, and to design a simplifying, though advanced framework for the optimisation of QoS conditions in SIP-based NGN.
2. To develop a technique for the efficient recognition and logical grouping of media sessions affected by similar QoS conditions as a fundamental basis for the comprehensive estimation and optimisation of QoS conditions.
3. To investigate non-invasive techniques for both efficient online QoS estimation and optimisation.
4. To derive mechanisms for the realisation and implementation of the identified techniques, and to integrate those mechanisms into the proposed framework, utilising functions provided by the NGN service infrastructure (such as Call Admission Control).
5. To implement and test a prototype comprising the most relevant functions of the defined framework to demonstrate their feasibilities.

The order of objectives declared above corresponds to the general structure of this thesis as presented within the following section.

Parts of the work carried out within this project have been contributed to the QoSSIP research project (Abu Salah *et al.*, 2008) funded by the German Federal Ministry of Education and Research.

1.2 Thesis structure

Chapter 2 provides the general background in the area of NGN. After introducing the NGN concept, the section provides a review of the main functionalities of SIP as a protocol and how SIP is typically mapped to the general architecture of NGN.

Chapter 3 addresses the fundamental problem of providing QoS in SIP-based NGN, illustrating the importance of QoS for real-time communications and showing existing approaches for the provision of QoS in IP networks. Furthermore the conventional architecture designed for QoS provision in NGN is introduced and analysed, and its issues and resulting requirements for the outcome of this project are pointed out.

In chapter 4 a framework for comprehensive QoS optimisation in SIP-based NGN is developed and introduced as a main outcome of the research work. The requirements resulting from the investigations discussed in chapter 3 are mapped to concrete tasks of an alternative NGN QoS architecture. The chapter describes the overall framework as well as its components.

One of the most crucial functions of the framework introduced in chapter 4 is termed *QoS profiling*, a concept that defines the classification of media streams into logical groups by the similarity of their QoS conditions. QoS profiling is considered in chapter 5 with all of its relevant aspects, such as the application of an unsupervised learning Artificial Neural Network, and mechanisms for both, accuracy self-scaling and bootstrap of the QoS profiling procedure.

Chapter 6 discusses the real-time estimation of the QoS experienced by any ongoing media session at any time. A resource-saving monitoring methodology is introduced, based on the profiling approach introduced in chapter 5. Furthermore the application of the ITU-T E model for real-time QoS estimation is discussed.

Chapter 7 proposes an approach for the optimisation of QoS conditions through the efficient passive control of network utilisation. For this purpose a feedback control system has been developed, which is described within this chapter. In order to identify sessions to be considered for required optimisation action, methods and algorithms have been developed, which are introduced. Furthermore the evaluation of the introduced QoS optimisation approach is discussed.

Chapter 8 focuses on both the proof of concept research prototype and on the evaluation of the proposed framework. The scope and general layout of the prototype is introduced, and its functional architecture and operation is described. The prototype has been used to accomplish several evaluation tests that demonstrate the benefits of the proposed framework. While the prototype is introduced in section 8, the performed tests and underlying scenarios are described in the respective related sections of the thesis. The framework evaluation described in chapter 8 is accomplished based on both, the research prototype that has been developed throughout this research work as well as on quantitative comparisons with the standardised NGN QoS architecture.

Chapter 9 presents the main conclusions of this research. Its key achievements are highlighted, and limitations are named. Finally, potential areas of future research and development are proposed.

This thesis is provided with a number of appendices in support of the general discussion, including a number of published papers.

2 SIP-based NGN

This chapter introduces the general environment considered by this research. After outlining the background of the NGN concept, the general NGN architecture is presented in parallel with definitions and key features specified by international standardisation bodies (section 2.1). The Session Initiation Protocol (SIP) and its provided basic and advanced functionalities are briefly described (section 2.2). Finally, the utilisation of SIP in NGN is depicted (section 2.3).

2.1 The concept of NGN

The history of the NGN (Next Generation Networks) concept starts in the mid-1990s when the term NGN was introduced describing a new era of telecommunication networks, particularly allowing the implementation of advanced personal communications for mobile networks (Ohuchi *et al.*, 1994). In the following years the term NGN increasingly became popular to face the comprehensive future trends emerging in the fixed and mobile telecommunications industry, such as the deregulation of the market (followed by the open and international competition among network operators), the increase of Internet utilisation (and hence the increase of data traffic), and the demands from users for new multimedia services and general mobility (Cochennec, 2002). In the year 2000 the ITU-T (International Telecommunication Union – Telecommunication Standardization Sector) started its research-based pre-standardisation work concerning NGN. Shortly before ITU-T, ETSI (European Telecommunications Standards Institute) had started its research work in the field of NGN. Nowadays, NGN research and standardisation work performed by both ETSI and ITU-T has been synchronised regarding all relevant aspects.

Coexisting and interacting with both industrial and academic research activities, ITU-T released its definition of NGN in 2004. In (ITU-T Y.2001, 2004) an NGN has been defined as

“a packet-based network able to provide telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies and in which service-related functions are independent from underlying transport-related technologies. It enables unfettered access for users to networks and to competing service providers and/or services of their choice. It supports generalized mobility which will allow consistent and ubiquitous provision of services to users.”

Hence, according to (ITU-T Y.2001, 2004), (ETSI TR 180 000, 2006), and (Trick and Weber, 2004), the term NGN stands for a telecommunication network concept that can be characterised by the following key features.

- Packet-based data transport
- Quality of Service support
- Applicability for arbitrary services
- Separation of call/service control and media data transport
- Capable for the integration of any existing, important telecommunication network, especially access networks
- Application Server support
- Support for multimedia services
- High bit rates
- Overall unified network management

- Mobility support
- Integrated security functions
- Service-appropriate charging
- Scalability
- Unrestricted access for users to different networks and service providers
- Consideration of obligatory legal requirements (such as lawful interception and emergency calling features)

As shown in Figure 2.1, the NGN core consists of a packet-switched network, supporting security and QoS functionalities.

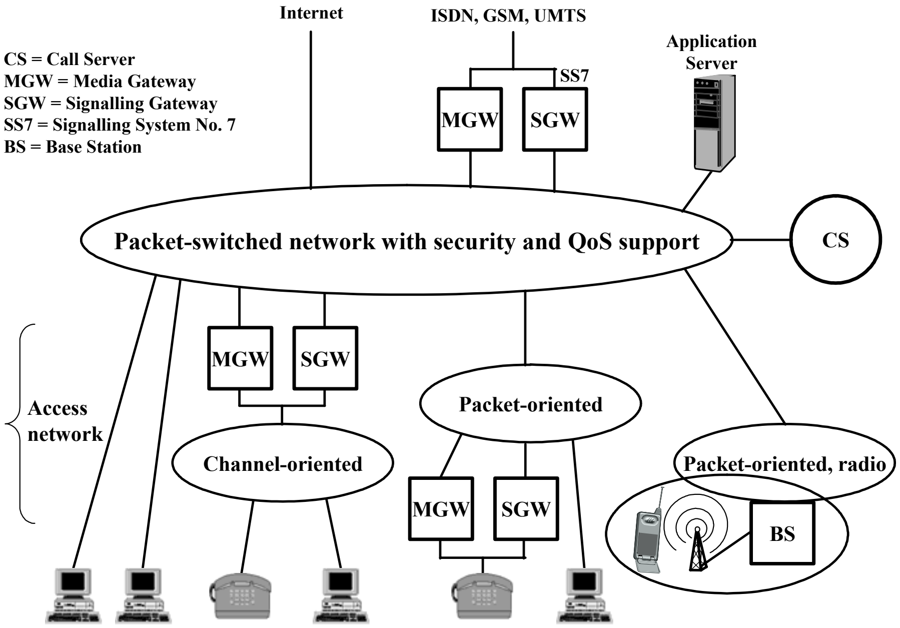


Figure 2.1: Principle structure of an NGN (Trick and Weber, 2009)

The user end systems, such as telephones, can be connected to the NGN either directly or via arbitrary access technologies, be it a channel- or packet-oriented, fixed or mobile access network, respectively. For this purpose, if required, Media and Signalling Gateways can be applied. Service requests are served by a Call Server (CS) and, if required, Application Servers can be involved in order to provide advanced services (such as value-added telecommunication services). The NGN generally offers access to other networks such as the Internet or, via gateways, to both circuit-switched and packet-switched telecommunciation networks such as the ISDN, GSM (Global System for Mobile communications), or UMTS (Universal Mobile Telecommunications System).

Regarding its functional architecture, according to (ITU-T Y.2011, 2004), an NGN can generally be divided into two strata (or layers, respectively), the service stratum and the transport stratum (see Figure 2.2). While the service layer is in charge of the control and management of user services (such as telephony sessions or video services) the transport layer provides user connectivity as well as data transport for both signalling and media transmission.

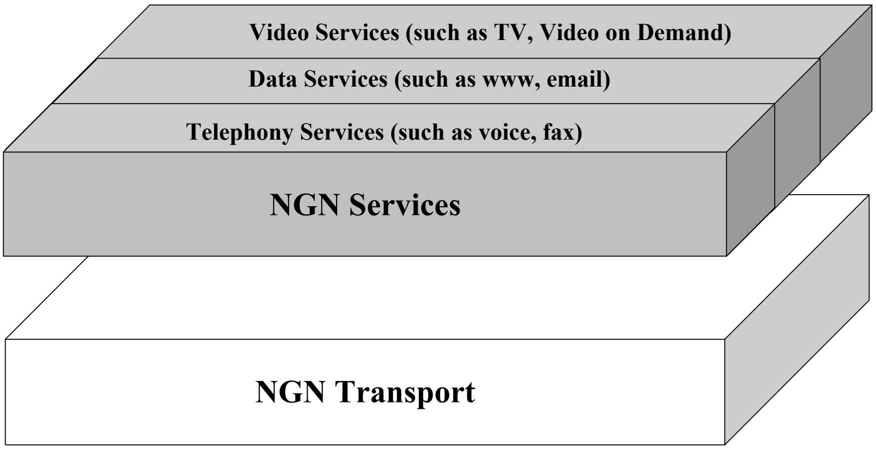

Video Services (such as TV, Video on Demand)
Data Services (such as www, email)
Telephony Services (such as voice, fax)
NGN Services
NGN Transport

Figure 2.2: NGN transport and service strata (ITU-T Y.2011, 2004)

Although the NGN concept is defined completely independent of specific technologies or protocols, currently NGN typically use IP networks (Internet Protocol) in order to serve the first mentioned key feature, packet-based data transport (Trick and Weber, 2009); (ITU-T Y.2011, 2004).

2.2 SIP (Session Initiation Protocol) and its related architecture

The Session Initiation Protocol (SIP) (IETF RFC 3261, 2002) has been defined by IETF (Internet Engineering Task Force) to provide connection-oriented services in connectionless networks such as IP-based networks. In NGN the main purpose of SIP is to allow for the initiation and management of connection-oriented communication sessions such as a VoIP call (Voice over IP) in IP-based networks (see section 2.2.1). Additionally, further basic services such as Event Notification (state monitoring) can be provided by the use of SIP (see section 2.2.2).

Within the proposed framework resulting from this research, SIP is utilised for both the control of communication sessions and to collect information on perceived QoS conditions.

2.2.1 SIP basic architecture and functionality

This section introduces the basic functionality of SIP as the protocol used for the initiation and control of connection-oriented sessions in NGN within the scope of this research.

In order to provide connection-oriented and reliable service functionality, SIP includes handshake, retry, and timeout mechanisms. SIP messages are carried over IP networks by the use of transport layer protocols such as UDP (User Datagram Protocol), TCP (Transmission Control Protocol), or SCTP (Stream Control Transmission Protocol), respectively.

In a SIP-based telecommunication infrastructure, user end systems (such as a VoIP telephone) are called SIP User Agents. Like any SIP entity, a SIP User Agent is a logical function running on an IP host and, therefore, has to be provided with an IP address. Because IP addresses of hosts might change, and because IP addresses do not allow for a comfortable user addressing, every SIP subscriber is provided with a permanent SIP URI (such as sip:B@Domain.com) by its respective telephony service provider. The service provider runs a SIP service infrastructure that consists of several logical server entities (such as SIP Proxy Servers, SIP Registrar Servers, and Location Servers). Several SIP server entities can be integrated in one IP host or, alternatively, can be implemented distributed as stand-alone entities.

Figure 2.3 gives an overview of the basic functionality of SIP and its main logical entities by presenting a session establishment scenario.

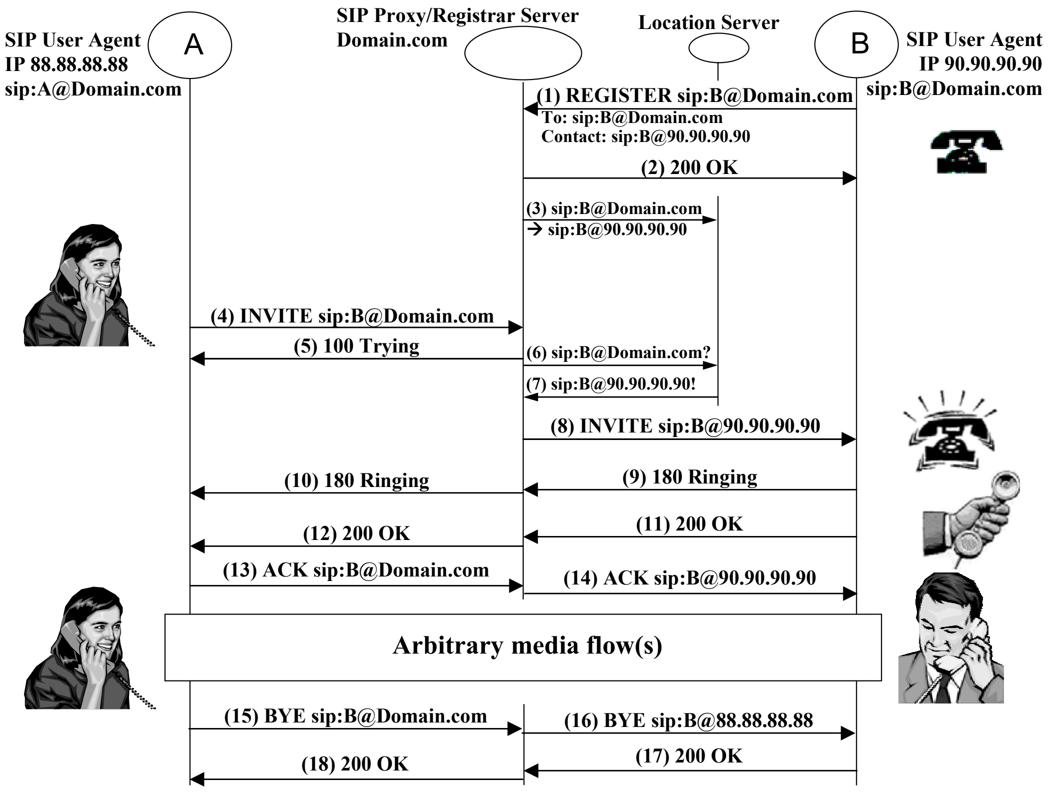


Figure 2.3: Basic SIP session initiation scenario

If a subscriber activates his SIP end system, the User Agent registers with a preset SIP infrastructure by sending a SIP REGISTER request to the SIP Registrar Server within the respective SIP service infrastructure (see Figure 2.3, step 1). The REGISTER request includes both the temporary SIP URI of the User Agent and the permanent SIP URI of the subscriber. While a temporary SIP URI such as *sip:B@90.90.90.90* addresses a defined user on a terminal with a dedicated IP address, a permanent SIP URI such as *sip:B@Domain.com* addresses a defined user as a costumer of a dedicated service provider that serves the denoted domain. After accepting the registration (see step 2) the SIP Registrar Server entity passes permanent and temporary SIP URIs to an entity referred to as Location Server (see step 3) that is usually represented by a database function. The Location Server holds the correlations of permanent and temporary SIP URIs for all registered subscribers of the respective service provider. Note that a Location Server by definition is not a SIP entity because information exchange with this entity is usually not based on SIP, but other appropriate protocols such as LDAP (Lightweight Directory Access Protocol). Also note that the Location Server functionality might be integrated with the SIP Registrar Server.

If a subscriber A wants to initiate a session (such as a VoIP call) with subscriber B, User Agent A sends a SIP INVITE request (see step 4) to the SIP Proxy Server entity operated by its respective telephony service provider (note that a SIP Proxy Server is usually integrated together with the Registrar Server entity on one IP host). After confirming the receipt of the request (see step 5), the Proxy Server prompts the Location Server for the temporary SIP URI associated with the permanent SIP URI that the request is addressed to (see step 6). After the Location Server has passed the requested information to the Proxy Server (see step 7) the request is forwarded to the respective SIP User Agent, now addressed by its temporary SIP URI (see step 8). User Agent B might respond to the request with several provisional SIP responses (such as 180 Ringing (see step 9)). When subscriber B has accepted the session, his User Agent at last responds to the INVITE request with a final SIP response (200 OK, see step 11). Note that all relevant SIP responses are transferred back to SIP User Agent A by the Proxy Server (see steps 10 and 12). User Agent A finally has to reconfirm its readiness by sending a SIP ACK request to User Agent B (see step 13). In managed SIP environments the ACK request is also forwarded through the Proxy Server (see step 14). After the session initiation is completed, a logical connection-oriented communication state (referred to as a SIP dialog) has been established between the involved User Agents. The end systems are now ready to exchange media data of arbitrary nature (such as VoIP and/or video data flows) by making use of any appropriate transport protocol, such as RTP (Real-time Transport Protocol (IETF RFC 3550, 2003)) for voice and video flows. Note that the media sessions and their characteristics (such as the choice of media, and the algorithms used for media encoding) are negotiated between the User Agents in line with session establishment by exchanging SDP media descriptions (Session Description Protocol, (IETF RFC 4566, 2006)) carried within the SIP messages. This negotiation is performed mutually by adopting the SDP offer/answer model specified in (IETF RFC 3264, 2002). Also note that SIP Proxy Servers are generally not involved in the media exchange between SIP User Agents. The packets containing the media data are forwarded in a peer-to-peer manner over the IP network between the User Agents.

If a subscriber decides to terminate the session, his SIP User Agent sends a SIP BYE request (see step 15). This request is typically forwarded through the SIP Proxy Server to the User Agent of the communication partner (see step 16). This User Agent has to confirm the receipt of this request by sending a SIP 200 OK response (see step 17) that is also forwarded back to the terminating User Agent by the Proxy Server (see step 18).

3GPP (Third Generation Partnership Project) has chosen SIP in order to provide multimedia communications within the UMTS as of Release 5 by defining an abstract architecture called IMS (IP Multimedia Subsystem) (3GPP TS 23.228, 2006). An IMS consists of a collection of logical functional entities, which, as a whole, represent an extended SIP service infrastructure with well-defined interfaces providing, amongst others, capabilities for the connection of service delivery platforms, or to interconnect with other telecommunication networks). Compared to a basic SIP environment, the design of IMS generally allows for the provision of QoS, charging, and customisation (Al-Begain *et al.*, 2009). Both IMS and SIP are also suggested for service and call control in NGN standardisation, respectively (see section 3.3.2).

2.2.2 Advanced SIP functionality

Beside the basic address resolution and session management functionalities of SIP illustrated in section 2.2.1, several further functionalities are provided by SIP and its architecture. Two of these extended SIP mechanisms are utilised by the proposed framework for the collection of QoS information and for the centralised control of session parameters. Within the following these mechanisms are generally introduced, showing their originally intended functionalities.

- The event notification framework as specified in (IETF RFC 3265, 2002) enables SIP entities to subscribe to specific events in other SIP entities, such as a change of a specific status. The most common use for this framework is the subscription to the presence state, which refers to "the willingness and ability to communicate with other users" according to (IETF RFC 3856, 2004). Because event notification formats are event-specific by nature, every event to which a SIP entity can subscribe is defined by an individual event package. The SIP requests SUBSCRIBE and NOTIFY (IETF RFC 3265, 2002) are the most relevant SIP messages used for event subscription and notification, respectively. Figure 2.4 shows an example of how the SIP event notification framework is typically applied for presence.

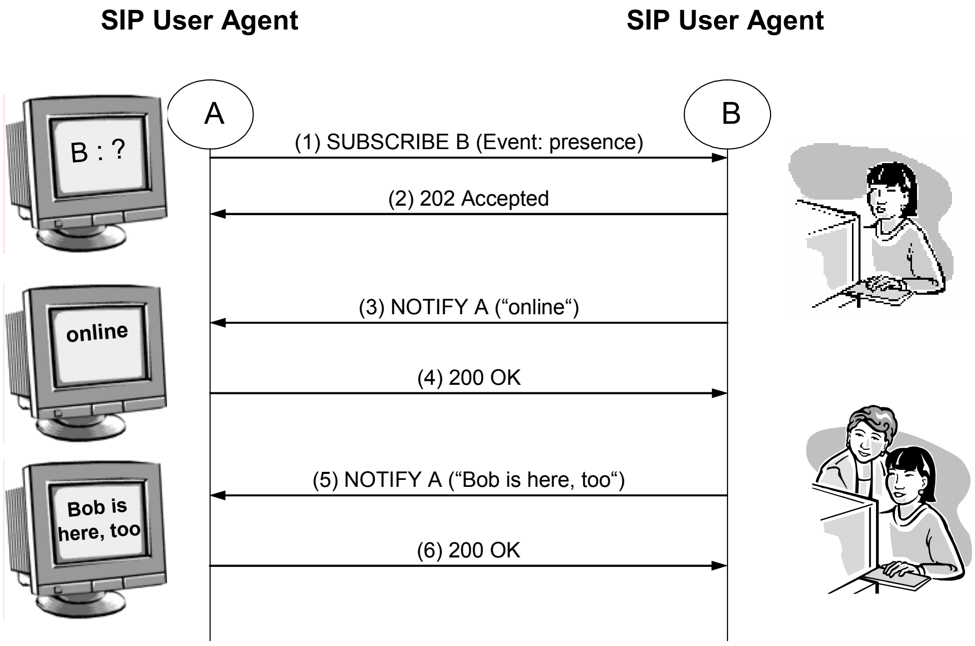


Figure 2.4: SIP presence event notification example

Within the event notification example in Figure 2.4, user A wants to subscribe to the presence state of user B. Therefore his terminal queries the terminal of user B for her presence state, using the SIP request SUBSCRIBE (step (1)). User Agent B accepts this request (step (2)). Given that B agrees to the subscription of user A, her User Agent provides presence information to A, carried within the message body of the SIP request NOTIFY (step (3)). The receipt of this request is confirmed by User Agent A with a 200 OK SIP response (step (4)). Note that, beside simple online/offline state information, the NOTIFY request may contain further details such as current location and activity of the user, or information regarding communication limitations, such as being available for text-based communication only. Whenever the presence state of user B changes, the updated information is provided to user A within a subsequent NOTIFY request (see steps (5) and (6)).

In (IETF RFC 6035, 2010), a SIP event package for voice quality reporting has been defined. This event package assumes that voice quality metrics (such as information on chosen codecs, packet loss rate, round trip delay, jitter, and estimated values expressing the voice quality as experienced by the users, such as MOS (Mean Opinion Score)) are exchanged between the parties of a call by the use of RTCP XR as defined in (IETF RFC 3611, 2003) (RTP Control Protocol eXtended Reports). In a further step, at least one party provides the voice quality information to a central server entity. According to (IETF RFC 6035, 2010) event package-based voice quality reporting can either be performed after the end of a respective media session or as an alerting mechanism in case of decreasing QoS during ongoing media sessions.

Within the proposed framework, SIP-specific event notification is used to subscribe to QoS-related information provided by user devices (see sections 4.4 and 4.5).

- The SIP method REFER (IETF RFC 3515, 2003) can be used to control another SIP entity regarding the generation of SIP requests addressing a third SIP instance. Amongst others, this method is used to manage call transfer procedures. After sending a REFER request, the controlling instance becomes informed about the status of the respective controlled action within SIP NOTIFY requests. Figure 2.5 shows a message flow for an unattended call transfer scenario.

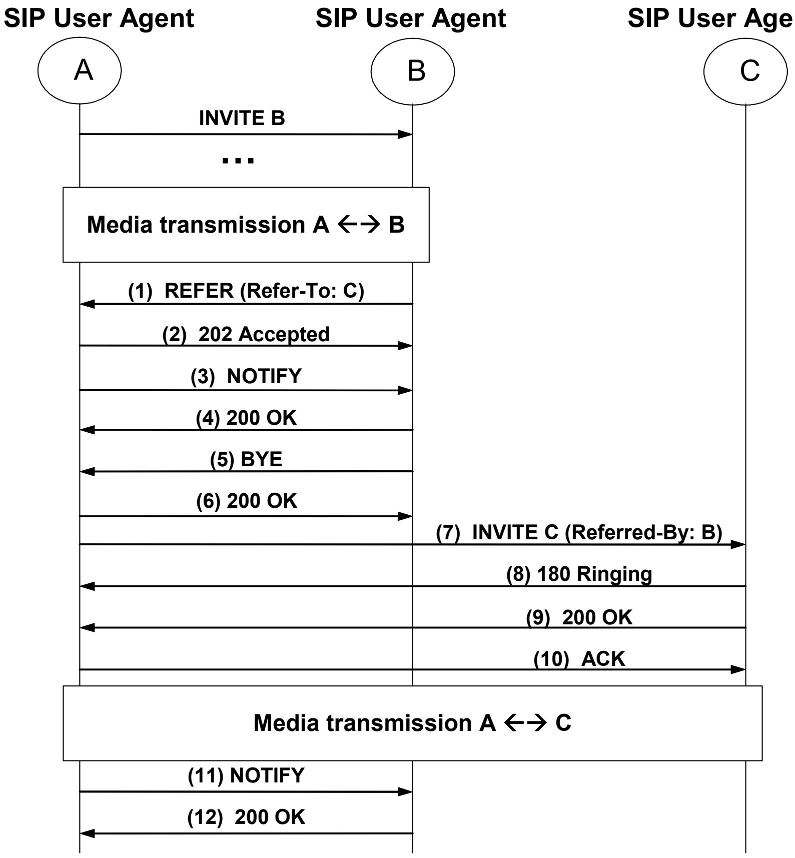


Figure 2.5: Application of the SIP REFER method for unattended call transfer (Trick and Weber, 2009)

Within the scenario shown in Figure 2.5, an already existing session between User Agents A and B is transferred by the initiative of B so that, as a consequence, A is going to be connected to a third subscriber C. Therefore User Agent B sends the SIP REFER request to User Agent A (step (1)). This request contains in the Refer-To header field the SIP URI of the destination to which the session is to be transferred. User Agent A accepts the REFER request (step (2)) and uses the SIP NOTIFY request (step (3)) to inform B that the transfer action is going to be performed subsequently. As a consequence, after confirming the receipt of the NOTIFY (step (4)), User Agent B terminates the SIP session with A (steps (5) and (6)). User Agent A now starts the session initiation with the new contact C. The INVITE request (step (7)) includes a Referred-By header field to inform C that this session is initiated on behalf of B. After successful session initiation (steps (7) to (10)), User Agent A informs User Agent B by the use of a NOTIFY request that the transfer has succeeded (steps (11), (12)).

Within the proposed framework the REFER method is used to control the renegotiation of media characteristics (Re-INVITE) and to trigger SIP session termination (see section 4.6).

Note that a brief informative overview on further SIP extensions and mechanisms are provided in Appendix B.

2.3 Utilisation of SIP in NGN

As denoted in section 2.2, SIP is a powerful protocol for the control and management of arbitrary user-related communications in IP infrastructures. That is why, in today's telecommunication business, SIP has become widely accepted as the protocol of choice for communication control in NGN. A SIP-based NGN, generally matching the NGN concept introduced in section 2.1, comprises an adequate IP transport infrastructure (logically situated within the transport stratum) and a suitable SIP service infrastructure (logically arranged within the service stratum). Note that in a SIP-based NGN infrastructure all network elements are linked to an IP transport environment.

Figure 2.6 shows the principle structure of an NGN based on SIP.

An NGN's centralised SIP service infrastructure is operated and managed by a respective SIP service provider. Within this infrastructure centralised Call Server (CS) functionality is provided by SIP Proxy/Registrar Servers, relying on Location Servers for the correlation of permanent and temporary SIP URIs (see section 2.2.1).

To provide value-added services (such as a multi-party conference service) to the subscribers both, Application Servers and Media Servers are be implemented.

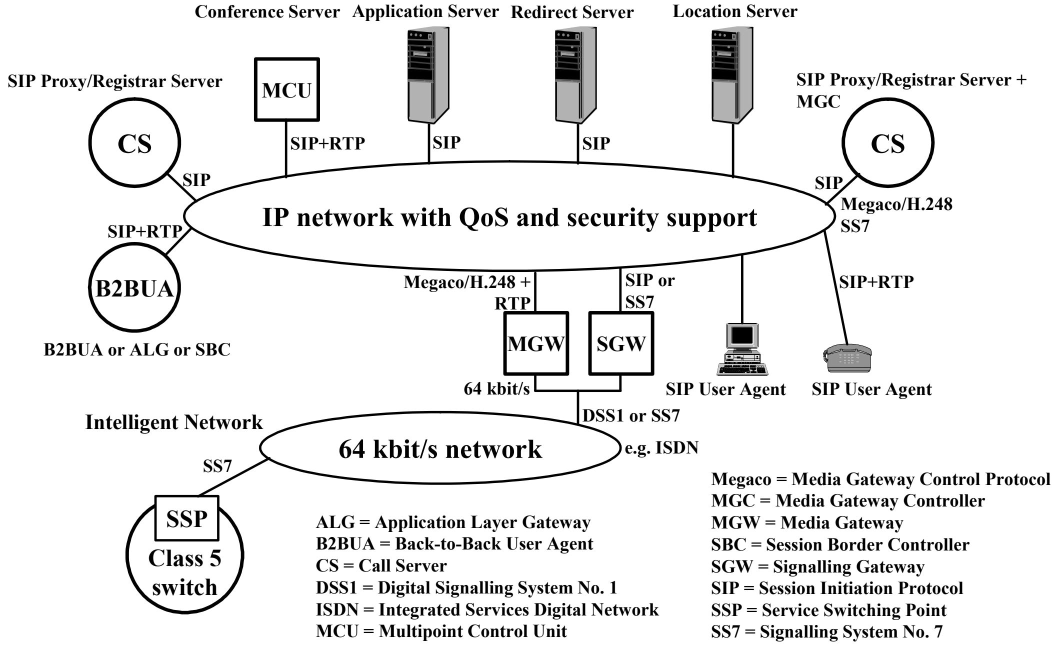

Conference Server
Application Server
Redirect Server
Location Server
SIP Proxy/Registrar Server
MCU
SIP Proxy/Registrar Server + MGC
CS
SIP+RTP
SIP
SIP
CS
SIP
SIP
IP network with QoS and security support
Megaco/H.248
SS7
SIP+RTP
B2BUA
Megaco/H.248 + RTP
SIP or SS7
SIP+RTP
MGW
SGW
B2BUA or ALG or SBC
64 kbit/s
SIP User Agent
SIP User Agent
Intelligent Network
DSS1 or SS7
64 kbit/s network
e.g. ISDN
SS7
SSP
Class 5 switch
Megaco = Media Gateway Control Protocol
MGC = Media Gateway Controller
ALG = Application Layer Gateway
MGW = Media Gateway
B2BUA = Back-to-Back User Agent
SBC = Session Border Controller
CS = Call Server
SGW = Signalling Gateway
DSS1 = Digital Signalling System No. 1
SIP = Session Initiation Protocol
ISDN = Integrated Services Digital Network
SSP = Service Switching Point
MCU = Multipoint Control Unit
SS7 = Signalling System No. 7

Figure 2.6: Principle structure of a SIP-based NGN (Trick and Weber, 2009)

By the use of Media and Signalling Gateways access to traditional circuit-switched networks, such as the PSTN (Public Switched Telephone Network), is offered. Subscriber end systems implement SIP User Agent functionality.

Note that in SIP-based NGN, IP transport infrastructure and SIP service infrastructure are operationally independent and, hence, are not necessarily operated by the same provider.

If required SIP signalling and media streams can be forced to be routed in parallel via intermediate service layer network elements trusted by the SIP service provider. This is potentially useful for the consideration of certain legal requirements such as lawful interception, for the interconnection with other providers' NGN, and for NAPT (Network Address and Port Translation) traversal. SIP Back-to-Back User Agents

(B2BUA) are used for this purpose, implemented in network elements coming in different flavours such as Session Border Controllers (SBC) or Application Layer Gateways (ALG). Note that the deflection of both signalling and media streams to certain network elements potentially counteracts the NGN key feature *Separation of call/service control and media data transport* (see section 2.1) but, on the other hand, potentially facilitates other NGN key features such as *Consideration of obligatory legal requirements*.

2.4 Summary

Within this chapter, the general environment of this research has been introduced. The emergence of the demand for packet-based telecommunication networks has been formulated, and the NGN concept has been described as defined by the standardisation bodies ITU-T and ETSI. The general partitioning of transport and service stratum has been depicted.

The protocol SIP has been chosen by 3GPP as the multimedia signalling protocol for IMS in UMTS Rel. 5 and higher, and has emerged being the de facto standard for session signalling in NGN. Therefore within this research, SIP has been considered as the protocol of choice for session-related NGN signalling. The general architecture of a SIP-based telecommunication infrastructure has been introduced, discussing the basic functionality of this protocol for session signalling and management, based on a message sequence example. Furthermore, a selection of more advanced SIP functionalities have been presented, comprising amongst others protocol frameworks for event notification and Third Party Call Control. Completing this section, the architecture of Next Generation Networks based on SIP has been outlined, providing an overall image of the environment of this research.

3 The challenge of Quality of Service (QoS) in SIP-based NGN

As denoted in chapter 2, in a SIP-based NGN, the control and management of services (such as VoIP sessions) is performed by the use of the application layer protocol SIP on top of IP for the exchange of service control information between SIP-aware network elements. Media data flows and their characteristics, such as medium and codec choice, are typically defined and negotiated between the user end systems by the use of SDP, carried within SIP messages. On the other hand, if a SIP session has been established, media data, such as an RTP (Real-time Transport Protocol (IETF RFC 3550, 2003)) packet flow delivering a voice stream, are exchanged between the user end systems in a peer-to-peer manner over the IP transport infrastructure. Hence, the characteristics (and their dynamic variances) of the involved IP transport infrastructure directly affect the delivery conditions of the media data flow, and thus, its real-time characteristics.

However, the nature of both SIP as a protocol and its related architecture does not allow for the direct control of packet transport characteristics in IP infrastructures. Beside this the connectionless and packet-oriented character of IP networks entails additional issues.

It is a challenge to provide QoS in SIP-based NGN. The fundamentals of this challenge are presented within the following subsections. After introducing the impact of QoS on real-time communication and how QoS is typically evaluated (section 3.1), approaches for QoS provision in IP networks are discussed (section

3.2). The remainder of this chapter (section 3.3) focuses on QoS provision in NGN, introducing and analysing both a reference NGN QoS architecture which follows a standardised layout, and several alternative approaches emanating from various research. Identified issues in NGN QoS provision are depicted, and requirements for an optimised approach are derived.

3.1 QoS and its importance for real-time communications

Quality of Service (QoS) is a general expression whose definition depends on the context of use. For telecommunication networks the term QoS is defined as the "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" (ITU-T E.800, 2008). According to (ITU-T Y.1291, 2004) QoS provision for services provided over packet-based networks is strongly related to the network performance provided on the IP layer, which, as defined in (ITU-T E.800, 2008), represents the "detailed technical part of the QoS planned and offered to the customer by the service provider".

As presented in (Gozdecki *et al.*, 2003), assuming that the bandwidth required by the data flow associated with a respective service is generally available between the involved data sources and destinations, network performance can be characterised by the following parameters:

- IP Packet Loss Ratio: In IP networks IP packets can become lost due to network traffic overload, link failures, or competing packets dropped. The packet loss ratio is typically defined as the ratio of the number of lost packets to the number of originally sent packets (Hagsand, 2004).

- IP Packet Transfer Delay: The transfer of an IP packet through an IP transport infrastructure inevitably entails a time delay. Beside the data transit time on involved transmission systems (depending, amongst others, on the overall length of links between sending and receiving host) the IP packet transfer delay is exceedingly influenced by the overall processing time required by IP routers, which are typically based on packet queuing and sequential processing.
- IP Packet Delay Variation (Jitter): The packet forwarding process in IP networks and hence the resulting the transfer delay experienced by IP packets is potentially influenced by variable factors. These include the processing times required by IP routers to forward a packet, amongst others depending on the utilisations of the involved routers at every specific point in time, and on the prioritisation rules of the queues a particular packet is classified for. This variation in the transfer delay directly affects the interarrival time of consecutive packets at the receiver side of a data stream.

The value ranges of these QoS-relevant parameters effective on a respective routing path between a sending and a receiving IP host significantly depend on the network dimension, topology, and configuration. In addition, various events, such as network congestion due to traffic overload, or misconfiguration of routing paths, and unforeseeable incidents, such as failures of network elements or transmission paths, within an IP transport infrastructure immediately have an effect on the dimensions of delay, jitter, and packet loss rate (White *et al.*, 2004).

The general levels of delay, jitter and packet loss ratio effective within an IP network can be influenced to a certain degree by applying appropriate network optimisation techniques. Amongst others, the following actions can potentially reduce the effects of delay, jitter, and packet loss.

- Reduction of the number of transit systems (such as routers) per routing path
- Appropriate dimensioning of the overall bandwidth available per routing path, considering the mean and maximum traffic requirements typically effective on the respective path
- Predefinition and appropriate dimensioning of alternative routes becoming effective in case of traffic overload or link failure

Note that, however, none of the above-mentioned QoS influences can be completely eliminated in any deployed IP transport infrastructure.

Note also that, in contrast to QoS, the term *Quality of Experience* (QoE) refers to "The overall acceptability of an application or service, as perceived subjectively by the end-user", which "may be influenced by user expectations and context" (ITU-T P.10/G.100 Amd. 2, 2008). Within the remainder of this thesis, the term QoS is explicitly used as defined within this section.

Section 3.1.1 introduces network performance requirements for real-time services such as VoIP, followed by a discussion of methods to evaluate QoS in real-time communications in section 3.1.2.

3.1.1 Network performance impact on real-time communications

In order to provide real-time communication services (such as a telephony service) over IP networks (referred to as VoIP) the respective media flows typically consist of a continuous sequence of IP packets generated and sent equidistantly in time, carrying encoded media data (such as segments of a continuous audio signal). Hence, conversational real-time services provided over IP are potentially vulnerable to increased delay and jitter values as well as an increased packet loss rate, as acknowledged by a number of papers. (Zheng *et al.*, 2001); (Kos *et al.*, 2002), (Borella et *al.*, 1998); (White *et al.*, 2004)

Figure 3.1 illustrates the potential impact of the transport network performance on a media stream, which consists of a series of consecutive packets. It is obvious that delay, jitter, and packet loss all potentially affect the real-time playout quality of the media stream received by B. Note that in most real-time communication forms, media streams are bidirectionally exchanged between the communicating parties, and that both streams can be potentially affected by insufficient network performance.

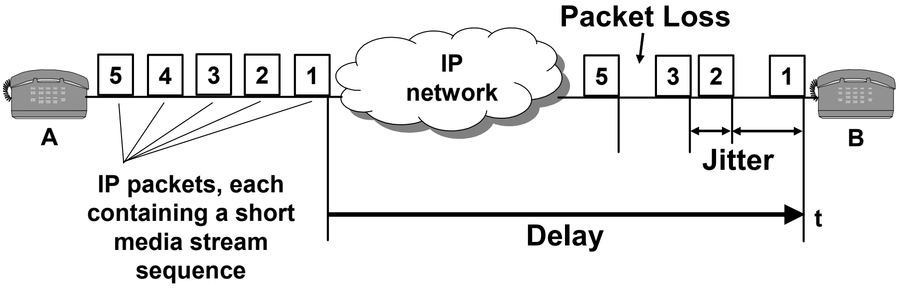


Figure 3.1: Impact of packet network performance on a real-time media stream

If an IP flow providing a real-time service is affected by **packet loss** the flow sequence incoming at the receiver is disrupted and hence, the media replay shows gaps. While the infrequent loss of single packets of a VoIP data stream is usually tolerated (if even noticeable) by a listener a significant rate of packet loss (and, in particular, packet loss bursts) will considerably affect the QoS experienced by the user according to (Borella *et al.*, 1998). For real-time communications such as VoIP a packet loss ratio not exceeding 10^{-3} is recommended according to (ITU-T Y.1541, 2006).

A constant **transfer delay**, which affects all packets of a real-time IP data flow causes a constant time shift in replay of the media stream at the receiver side. For unidirectional real-time services, such as Video on Demand (VoD), this effect typically has virtually no impact on the QoS experienced by users. Regarding conversational real-time services such as VoIP, an excessive delay in one or both communication directions directly affects the interaction within the communication experience. For instance, in a voice conversation a speaking person potentially expects the dialogue partner to be able to immediately react verbally within the context of the conversation. If an expected instant verbal reaction is considerably retarded due to round-trip packet delay effects, the communication experience can be seriously affected. Hence, the packet delay has an impact on the QoS. According to (ITU-T G.114, 2003) good QoS can be experienced for voice communications if the speech delay from the mouth of a speaker to the ear of a listener is less than 200 ms. However, most users would still be satisfied with a mouth-to-ear delay up to 400 ms. For real-time communications over IP networks (ITU-T Y.1541, 2006) recommends a mean delay of 100 ms between the user-to-network interfaces of the speaking and the listening users' end systems in case of constrained routing and distance, or 400 ms, respectively, in case of less constrained routing and distances.

A **transfer delay variation (jitter)** affecting packets of a real-time data flow results in time gaps or overlaps of data carried in consecutive packets of a data stream. Hence packet transfer delay variation results in a data phase shift experienced by the receiver of a real-time data flow such as a VoIP data stream. If the data were processed and replayed without further sequencing, this would result in dropout and crossover effects within the replayed signal. To desensitise real-time media replay for delay variation the use of jitter buffers applied to the end systems (such as VoIP terminals) is recommended (Zheng *et al.*, 2001). A jitter buffer further delays the replay of the media flow to allow for the correct time sequencing of data carried within consecutive packets. Note that jitter buffers come by the cost of a general replay delay that can be perceived by users. Hence, the dimensioning of a jitter buffer represents a trade-off between an additional delay and the neutralisation of jitter effects up to a certain extend. In order to always meet the best trade-off for effective network conditions advanced jitter buffers can be applied that adapt their size in case of changing network conditions. However, if the delay variation appearing in a transport infrastructure exceeds the maximum buffer time of the jitter buffer, the concerned packets are dropped. The resulting packet loss effect has an impact on the QoS (Kos *et al.*, 2002).

For real-time communications such as VoIP, the transfer delay variation affecting a respective IP data path should not exceed 50 ms according to (ITU-T Y.1541, 2006).

3.1.2 Evaluating QoS in real-time communications

As mentioned in section 3.1.1, real-time communication services such as voice telephony are relatively vulnerable to variations in network performance. On the other hand, a real-time communication service has to provide a satisfying QoS, since telecommunication subscribers do typically not accept unreliable service quality. In order to provide evidence of adequate QoS, and to identify potential issues affecting the service quality, several different methods exist to identify and characterise the quality of real-time communication services in general, and, particularly, for those services provided over IP networks. Within this section, the background of QoS evaluation is summarised, mainly considering narrow-band real-time voice communication as the most common telecommunication service provided. Note that similar extensions and equivalent methods generally exist for other types of services and media, respectively, such as wide-band voice communication or video conferencing.

In (ITU-T P.800, 1996), the ITU-T has specified the Mean Opinion Score (MOS) as a single measure to evaluate the transmission quality as subjectively experienced by persons listening to transmitted speech, or communicating verbally over a bidirectional voice transmission path. The MOS scale has been defined in a value range from 1 (“bad”) over 2 (“poor”), 3 (“fair”), 4 (“good”) to 5 (“excellent”).

In standardisation work, several different methods and models have been specified for the technical evaluation of MOS values for speech transmission and voice communication. The application of such mechanisms can be described as QoS measurement, generating objective result scores. Furthermore, beside QoS measurement, calculation models can be applied to estimate the QoS a user will typically experience under specific conditions (such as given performance characteristics of a particular transmission path). Hence, in order to distinguish among different notions of MOS obtainment, and to further differentiate whether a particular MOS is related to listening, talking or conversational QoS, different flavours of MOS have been defined as stated in Table 3.1.

Table 3.1: Flavours of Mean Opinion Scores for QoS assessment, simplified (ITU-T P.800.1, 2006)

	Listening-only (LQ)	**Conversational (CQ)**	**Talking (TQ)**
Subjective (S)	MOS-LQS	MOS-CQS	MOS-TQS
Objective (O)	MOS-LQO	MOS-CQO	MOS-TQO
Estimated (E)	MOS-LQE	MOS-CQE	MOS-TQE

Regarding the measurement of QoS and performance in telecommunication environments, several methods for testing can be classified being either active or passive, and either intrusive or non-intrusive. According to (ITU-T P.10/G.100 Amd. 2, 2008), in **active testing** a dedicated "channel" (or a session, respectively) is set up for the transmission of data to be assessed (such as an audio stream). In contrast, in **passive testing** an existing communication path is utilised. While in **intrusive testing** a dedicated signal (such as a predefined audio stream) is injected into the testing environment, in **non-intrusive testing** available signals (such as the payload of an ongoing telephony conversation) are analysed. Typically, methods for objective QoS measurement can be either described as combining active and intrusive testing, such as **PESQ** (Perceptual Evaluation of Speech Quality) (ITU-T P.862, 2001), or combining passive and non-intrusive testing, such as **3SQM** (Single Sided Speech Quality Measure) (ITU-T P.563, 2004), respectively.

In (Raja *et al.*, 2007) an alternative non-intrusive approach for speech quality prediction (listening-only) is introduced, that, according to its authors, in contrast to PESQ or 3SQM, is explicitly capable to perform under real-time conditions. A Genetic Programming (GP)-based symbolic regression approach is applied, taking into account the codec-specific bit rate, packetisation interval, frame duration, mean packet loss rate, and the mean burst length. PESQ is utilised to generate target output values for the training of the GP algorithm prior to its application.

As stated in section 3.1, the QoS experienced by a user of a real-time communication service (such as voice or video telephony) provided over IP mainly depends on the performance of the involved networks. Hence, as identified in research work such as (Zheng *et al.*, 2001) and (Calyam *et al.*, 2004), regarding the QoS experienced by users, conclusions can be drawn from the measurable network performance to the QoS provided to human beings involved in a conversation. The E model is, by far, the most prevalent approach for the estimation of the effective QoS in VoIP communications from measured network performance; the method has been first adopted by ITU-T in the year 1998 and has been revised several times until the current version published as (ITU-T G.107, 2009).

The E model, originally designed as a network planning tool, can be described as a collection of equations for the calculatory estimation of the influences of, amongst others, the network performance on the QoS experienced by users of a voice telephony service. If applied for QoS estimation, the E model provides a pseudo-percentage value called **R factor** (transmission rating factor; see Equation 3.1; (ITU-T G.107, 2009)), composed as a sum of the following components.

$$R = Ro - Is - Id - Ie\text{-}eff + A \qquad (3.1)$$

- *Ro* = overall signal-to-noise ratio
- *Is* = Combination of all impairments occurring simultaneously with the voice signal
- *Id* = Impairments resulting from delay

- *Ie-eff* = Impairments resulting from the data compression performed by low bitrate audio codecs; strongly related to packet loss influences
- *A* = Advantage factor, allowing for the consideration of potential advantages for the user, such as personal mobility when using a mobile device for the access to a telecommunication service

Note that *Ro*, *Is*, *Id*, and *Ie-eff* come as discrete equations, involving several parameters defined in (ITU-T G.107, 2009), each coming with a permitted value range and a default value. Assuming the default values for those parameters not related to network performance or codec characteristics, the basic signal-to-noise ratio *Ro* is replaced by a constant value, and the Advantage factor is not applied. Hence, solely considering network performance and codec properties, the transmission rating factor *R* can be expressed as a function of the following components (see equation 3.2).

$$R = f\,(Is;\ Id;\ Ie\text{-}eff) \qquad (3.2)$$

If itemised, the remaining components of *R* can be expressed as functions of parameters expressing both, network performance and codec characteristics, as shown in equations 3.3 to 3.5. The symbolic abbreviations used in these equations were chosen according to (ITU-T G.107, 2009) and refer to the following parameters. Their respective units are given in square brackets. Dimensionless values are labelled [-].

- *T* = Mean one-way delay [ms]
- *Tr* = Round-trip delay [ms]; typically *Tr* = 2 *T*
- *Ta* = Absolute (one-way) delay [ms]; typically *Ta* = *T*
- *Ie* = Equipment impairment factor (codec-specific) [-]

- *Bbl* = Packet-loss robustness factor (codec-specific) [-]
- *Ppl* = Random packet-loss probability [%]
- *BurstR* = Packet-loss burst ratio [-]

$$Is = f\,(T) \tag{3.3}$$

$$Id = f\,(T,\ Tr,\ Ta) \tag{3.4}$$

$$Ie\text{-}eff = f\,(Ie,\ Bbl,\ Ppl,\ BurstR) \tag{3.5}$$

Note that for the most common audio codecs, default values of *Ie* and *Bbl* can be obtained from the respective latest version of (ITU-T G.113, 2007). If a novel audio codec has to be adopted, in order to apply the E model for quality rating, values for *Ie* and *Bpl* have to be determined beforehand by subjective testing. (AL-Akhras *et al.*, 2009) proposed a method to determine the dependency of *Ie-eff* from *Ppl* and *BurstR* by objective testing, based on the PESQ algorithm. Within this method, the codec-specific components of *Ie-eff*, namely *Ie* and *Bbl*, are substituted by weights of an Artificial Neural Network (ANN). Hence, after having trained the ANN for a respective codec, according to (AL-Akhras *et al.*, 2009) this method can be applied for QoS estimation combined with the conventional E model, involving any arbitrary voice codec without explicit knowledge of its related *Ie* and *Bpl* values.

Also note that symmetric delay conditions in both communication directions are typically assumed, which allows for the simplification *Tr* = 2 *T* as mentioned above. If, however, unsymmetric delay conditions have to be faced, *Tr* has to be composed as the sum of the mean one-way delays of both communication directions.

Generally, from the R factor, once calculated according to (ITU-T G.107, 2009), a MOS can be derived, referring to the estimated quality experienced by a user having a bidirectional conversation ($\mathrm{MOS}_{\mathrm{CQE}}$) (Conversational Quality Estimated). As defined in (ITU-T G.107, 2009) the conversion from R to MOS_{CQE} is performed according equation 3.6.

$$\text{For } 0 < R < 100\text{: } MOS_{CQE} = 1 + 0.035R + R(R-60)(100-R)7*10^{-6} \qquad (3.6)$$

Within this research work, the E model has been extended and applied for real-time QoS estimation (see section 6.2).

3.2 Providing QoS for IP real-time services

Both IP as a protocol and as a network layer technology were originally designed for non-guaranteed, best effort delivery. Hence specific issues exist regarding the provision of QoS for conversational real-time communication services, such as VoIP, within basic IP environments. Several auxiliary protocols and mechanisms, referred to as IP QoS mechanisms (see section 3.2.2), had to be specified in order to allow for QoS provision over IP networks. However, all these active mechanisms come with their individual issues and deficiencies, respectively.

As a matter of fact no reliable IP QoS mechanism is applied to the public segments of today’s Internet. That is the main reason why several alternative, rather passive methods for QoS provision in IP environments (in particular aiming on real-time communications) have been devised (see section 3.2.1).

3.2.1 Passive IP QoS provision approaches

Several research studies focused on QoS provision for real-time services over basic IP transport environments. A subset of these proposed methods, which assume that communication end points such as VoIP terminals may partially control the provision of QoS to the user, are reviewed in this section.

One well-researched general approach for QoS provision in IP environments is referred to as Endpoint Admission Control (EAC) or Self-Admission Control. The common goal of research within this field is the end system-driven verification of effective network conditions (hence QoS) by which an already established or intended communication session is or will be affected. Based on the verification results an admission decision is made, which mainly aims on the denial or cancellation of sessions which would not gain sufficient QoS. Some alternatives within this field recommend a preceding phase in which special probing packets are exchanged before the initiation of an intended communication session (Bianchi *et al.*, 2002). In other studies, the verification is performed by the use of media payload packets of an already established communication (Hagsand *et al.*, 2004), or by the analysis of dedicated QoS determination traffic which is exchanged simultaneously with the media payload (Mase and Toyama, 2002). In all cases, either one or both of the endpoints collect and analyse network performance parameters such as jitter (Bianchi *et al.*, 2002) or packet loss (Mase and Toyama, 2002), (Hagsand *et al.*, 2004), resulting in an admission decision for the intended or ongoing communication session, respectively. (Senthilkumar and Sankaranarayanan, 2006) compares various EAC approaches, concluding that network performance verification accomplished by communication endpoints analysing packet inter-arrival delay (jitter) should be preferred over packet loss analysis.

Another approach related to EAC (Qiu *et al.*, 2001) applies an end-to-end based verification of network conditions, but assumes that admission control functionalities are separated from the end systems. However, the effective improvement or manipulation of QoS conditions is not provided sufficiently by any known EAC approach.

Beside EAC, further approaches aim to improve QoS in basic IP environments. For instance in (Jyoti *et al.*, 2006) a routing diversity approach is proposed, aiming on the utilisation of uncontrolled load balancing effects. It was found that routing diversity can improve the QoS of VoIP calls over long distance links.

An NGN typically includes a centralised service control infrastructure, operated by a service provider (see section 2.1). The provision of services within an NGN environment is based on the interaction between the service provider infrastructure and the subscriber end systems. Hence, in order to provide reliable services to NGN subscribers, the service control infrastructure must be involved in decisions and actions subject to the QoS of a service provided. Thus, note that the approaches introduced within this subsection are not applicable in SIP-based NGN without additional arrangements for the involvement of the centralised service control infrastructure regarding Admission Control.

3.2.2 Standardised IP QoS mechanisms

In order to allow for the active control of QoS provided by IP transport networks several mechanisms have been designed and specified. Nowadays IntServ/RSVP (Integrated Services / Resource Reservation Protocol) (IETF RFC 1633, 1994); (IETF RFC 2205, 1997), DiffServ (Differentiated Services) (IETF RFC 2474, 1998); (IETF RFC 2475, 1998), and MPLS (Multi-Protocol Label Switching) (IETF RFC 3031, 2001) are the most well-known techniques for QoS provision in generic IP transport infrastructures. Because of their acceptance within the IP community and because of the lack of adequate alternatives these mechanisms are also designated for QoS provision within IP-based NGN (ITU-T Y.1291, 2004).

However, all these mechanisms come with their own modes of operation and potential issues that can cause IP transport characteristics to become inefficient or insufficient, respectively, depending on the overall design and dimension of a respective IP transport infrastructure.

The IntServ approach, usually implemented by the use of RSVP, is able to provide absolute QoS to IP data flows by supporting precise end-to-end per-flow service provisioning (Simmonds and Nanda, 2002), (Welzl and Mühlhäuser, 2003). It works as a session-based, path-coupled resource reservation service within the IP transport network, providing certain QoS-relevant attributes (such as a certain minimum bandwidth required for a respective IP flow; and a certain maximum processing delay). The reservation of paths is initialized before the actual user data flow becomes active and usually has to be refreshed during the user data flow session. Applied on an NGNs' IP transport infrastructure the amount of resource management traffic generated by IntServ/RSVP is not efficiently controllable by the NGN provider because the amount of resource reservation traffic does not only depend on the number of active NGN subscribers but also on their session behaviour (such as the number of session requests per time per user, and the average duration of each session) (Weber *et al.*, 2007). Summarising it can be said that IntServ/RSVP is described to lack scalability by several research work (Simmonds and Nanda, 2002), (Welzl and Mühlhäuser, 2003), (Zhou *et al.*, 2006). Furthermore, according to (Giordano *et al.*, 2003) IntServ/RSVP potentially introduces security issues, and, as mentioned in (Bohnert *et al.*, 2007), should not be considered as an adequate solution for the use in complex scenarios.

DiffServ provides the opportunity to mark every IP packet of a respective data flow to be prioritised in a certain way, and so, belonging to a certain class of service. Because no additional signalling is required, DiffServ is considered to be more scalable than IntServ/RSVP. With DiffServ versatile classes of service can be distinguished to provide differing QoS-relevant properties, such as certain rates of maximum packet loss, and a certain maximum processing delay. Several DiffServ classes for the use in NGN, referred to as QoS classes, have been defined according to (ITU-T Y.1541, 2006). Note that DiffServ on its own, by definition, provides only relative QoS (Zhou *et al.*, 2006), since all IP packets that belong to a certain class are treated in the same way. If the capacity of an IP transport link becomes insufficient due to abrupt traffic increase, the shortage of resources also affects packets marked by DiffServ (Trick and Weber, 2009). Hence, DiffServ is not efficient in dealing with network overload and is considered impractical for the use in networks that have to deal with a relative high volume of high priority traffic (Menth, 2006).

Although being stated to be potentially beneficial for QoS provision in NGN, according to (Bohnert *et al.*, 2007), also the application of MPLS is limited related to scalability and efficiency. Strictly speaking, MPLS is not an IP layer QoS mechanism, but acts in between OSI layers 2 and 3. More specifically, an MPLS network provides the service of IP packet transport, but therefore a special MPLS-aware transport infrastructure is required (note that MPLS is typically not supported by cheap IP routing standard equipment). Initially, paths (each providing certain obligatory QoS-relevant properties, such as a certain amount of bandwidth) within the MPLS network are set up. All IP packets belonging to the same data flow are provided with the same MPLS label by the network border entity and, subsequently, are switched through the network on the same path. Within an MPLS-based IP transport infrastructure packets can be processed and forwarded faster in comparison to a network that forwards IP packets on the basis of IP routing.

Although the combined application of at least two of the above-mentioned mechanisms can result in a trade-off between scalability and efficiency, their individual issues potentially still affect the overall performance of an IP transport infrastructure and, hence, the throughput conditions for IP data flows (Welzl and Mühlhäuser, 2003), (Giordano *et al.*, 2003).

Note that the specified IP QoS mechanisms are, by definition, not aware of the communication sessions, such as a VoIP session, established by higher layer protocols such as SIP.

3.3 QoS provision in SIP-based NGN

DiffServ, IntServ, and MPLS, although coming with their particular functional restrictions (see section 3.2.2), generally allow the control of packet transport conditions within IP infrastructures. These mechanisms are also considered by (ITU-T Y.1291, 2004) as exemplary QoS mechanisms within IP-based transport networks of NGN.

In most cases, an IP-based NGN transport infrastructure consists of a core IP network and one or more access networks. The shared goal for both network types is to provide the functionality of IP data transport to and from the subscriber end systems.

According to (ITU-T Y.2012, 2010), several types of access networks, including wired, such as x-DSL (generic Digital Subscriber Line), wireless, such as WLAN (Wireless Local Area Network) and cellular, such as UTRAN (Universal Terrestrial Radio Access Network), should be potentially supported by NGN. In general, an NGN access network supports policy enforcement which means it supports, amongst others, QoS traffic conditioning. Currently, access networks are usually implemented in a multi-layered fashion, such as IP over ATM (Asynchronous Transfer Mode) over DSL (Digital Subscriber Line). Beside the access technologies already mentioned, other access technologies arise, which come with their own specified QoS mechanisms such as DOCSIS over HFC-based access networks (Data-Over-Cable Service Interface Specifications; Hybrid Fiber Coax) (DOCSIS 1.1, 2005).

The core transport network consists of functions that ensure the transport of user traffic throughout the core network. According to (ITU-T Y.2012, 2010) those functions support QoS as one of their main features, applying mechanisms related to QoS provision such as buffer management, queuing and scheduling, traffic classification, marking, and policing. Note that within the core transport network, typically several independent protocols appear on top of each other, each of which may provide its own QoS and traffic control mechanisms. For instance, IP may be carried over MPLS over Ethernet.

In order to provide QoS interworking for IP transport between an access network and an NGN core network, mapping between different QoS mechanisms (such as a mapping between DiffServ and MPLS) has to be accomplished.

3.3.1 Bridging the service and transport strata of NGN

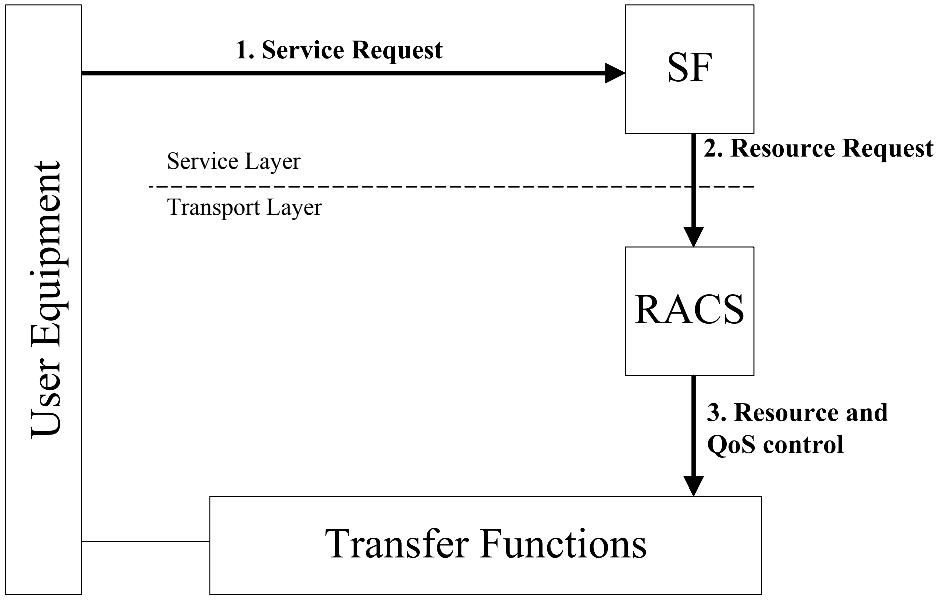
As mentioned in section 2.1, according to standardisation work, the NGN architecture is generally divided into a service stratum (in case of a SIP-based NGN mainly providing SIP service request processing and service control) and a transport stratum (mainly providing IP connectivity and data transport). Within a SIP service infrastructure (which is logically arranged within the service stratum of an NGN), the IP QoS conditions required for a respective media flow can generally be derived from the media and codec negotiation information exchanged via SIP/SDP (see section 2.2.1). However, QoS provision in IP transport infrastructures (logically situated within the transport stratum of an NGN) technically cannot be controlled directly by the protocol SIP or its infrastructure as-is. Therefore, it seems self-evident and reasonable to let the SIP service infrastructure control IP QoS mechanisms by the use of an intermediate functionality, situated between the NGN's service and transport strata. Giving more details, once an NGN subscriber requests a service, a top-down resource handling is accomplished in order to provide QoS for IP packet flow transport within the context of the respective service (such as a VoIP session). In terms of SIP, if a SIP session initiation request originating from a SIP User Agent is received by the SIP service infrastructure (such as an IMS) situated within the NGN service stratum, resource and QoS requirements for the media data transport in the session scope are identified and handed down into the transport stratum where network resources for QoS provision are authorised, allocated, and reserved within the IP transport infrastructure. This kind of cross-stratum mechanism has become the prevalent approach for NGN QoS provision. It has been mentioned in NGN research work such as (Vautier *et al.*, 2002), and has been adopted by ITU-T in (ITU-T Y.1291, 2004); (ITU-T Y.2012, 2010), and by ETSI in (ETSI TS 185 001, 2005); (ETSI TR 182 022, 2007) for both general and specific definitions of the standardised NGN QoS architecture.

Note that appropriate interfaces and protocols, such as Diameter (IETF RFC 3588, 2003), have been assigned by standardisation bodies to allow for the required cross-stratum signalling. Also note that, because additional signalling is required, this NGN QoS provision approach inevitably comes with additional traffic load generated in order to allow for the cross-stratum communication between an NGN's service and transport stratum.

3.3.2 Introducing the standardised NGN QoS architecture

Within NGN standardisation work performed by both ITU-T and ETSI attention has been and still is given to the challenge of QoS provision in NGN. Within this section the ETSI TISPAN NGN QoS architecture according to (ETSI TS 185 001, 2005) is briefly introduced as a typical example for the application of the cross-stratum QoS provision approach denoted in section 3.3.1.

Figure 3.2 shows the generalised signalling path of the related QoS provision scenario. Beside the involved subscriber end system (referred to as User Equipment), the IP transport infrastructure (referred to as Transfer Functions), and the SIP service infrastructure (referred to as Service Call Control Functions, mainly fulfilling the functionality of a Call Server) an additional entity (referred to as RACS (Resource and Admission Control Subsystem)) is involved. The latter entity accomplishes the required communication between the SIP service environment and the IP transport infrastructure, exchanging signalling with both entities.



RACS = Resource and Admission Control Subsystem
SF = Service and Call Control Functions (including IMS)

Figure 3.2: NGN QoS provision with cross-stratum approach (push mode) (ETSI TS 185 001, 2005)

A User Equipment requests a service (such as the initiation of a SIP-based voice session) by sending a service request (such as a SIP INVITE request; see Figure 3.2, step 1) to the Service and Call Control Functions of the NGN it is connected to. The Service and Call Control Functions identify the required resource conditions (such as a certain minimum bandwidth, and the QoS conditions required) for the respective service and, in step 2, send a resource allocation request to the RACS. Depending on the respective NGN's typical procedures the RACS might authorise the subscriber to use a specific amount of resources and specific QoS conditions (depending on the user policy of the subscriber) and check whether the requested resources are available within the NGN transport infrastructure. If the resource request can be fulfilled, the RACS triggers the resource and QoS allocation (step 3) in the transfer functions of the respective IP transport network. Hence, the RACS is in charge of the resource and QoS settings of several network elements (such as IP routers and switches) within the transport network that are involved in the resource and QoS processing for the respective media data flow. Appropriate signalling is required between RACS and the transport network in order to manage resource and QoS conditions. If a subscriber terminates a service used (for instance, if a VoIP session is closed) the RACS is in charge of resource release and QoS reset in the transfer functions of the respective IP transport network.

Note that, in order to allow for the integration of the abstract NGN QoS architecture into NGN IP infrastructures, both RACS and Transfer Functions comprise several sub-entities labelled as follows (ETSI ES 282 003, 2008).

- **RACS (Resource and Admission Control Subsystem)**
    - **SPDF** (Service-based Policy Decision Function): single point of contact to the Service and Call Control Functions. Distributes information to several x-RACFs if required.
    - **x-RACF** (Generic Resource and Admission Control Function): in charge of resource and admission control. An x-RACF comes either as A-RACF (Access-RACF) or C-RACF (Core RACF), each situated within the respective part of network. Note that several x-RACFs can be served within the context of one RACS.

Note that one RACS typically comprises one SPDF but, possibly several instances of x-RACF, distributed among different network parts (such as a core network and several access networks).

- **Transfer Functions**
    - **RCEF** (Resource Control Enforcement Function): performs policy enforcement functions, controlled by the x-RACF in charge of the respective part of network.
    - **BGF** (Border Gateway Function): packet-to-packet gateway for user plane media traffic. Controlled by the RACS's SPDF.
    - **BTF** (Basic Transport Function): performs packet forwarding based on rules that can be controlled by the use of path-coupled signalling protocols such as RSVP.

Note that all sub-entities of one Transfer Function are assumed to be typically situated within the same physical node.

Also note that, considering the fact that the overall functionality of an RACS typically is distributed among different physical nodes, it is understood that the signalling required for resource and QoS control as shown in Figure 3.2 is drastically simplified. Implementing the conventional NGN QoS architecture in fact results in a remarkable amount of resource and QoS control traffic (see section 3.3.3).

3.3.3 Analysis of the standardised NGN QoS architecture

This section analyses the mode of operation of the conventional NGN QoS architecture, based on a concrete exemplary communication scenario which uses a specific reference system introduced in (ETSI ES 282 003, 2008). This reference system is also assumed for the discussion throughout the remainder of the thesis.

Figure 3.3 provides an overview of the related reference network structure.

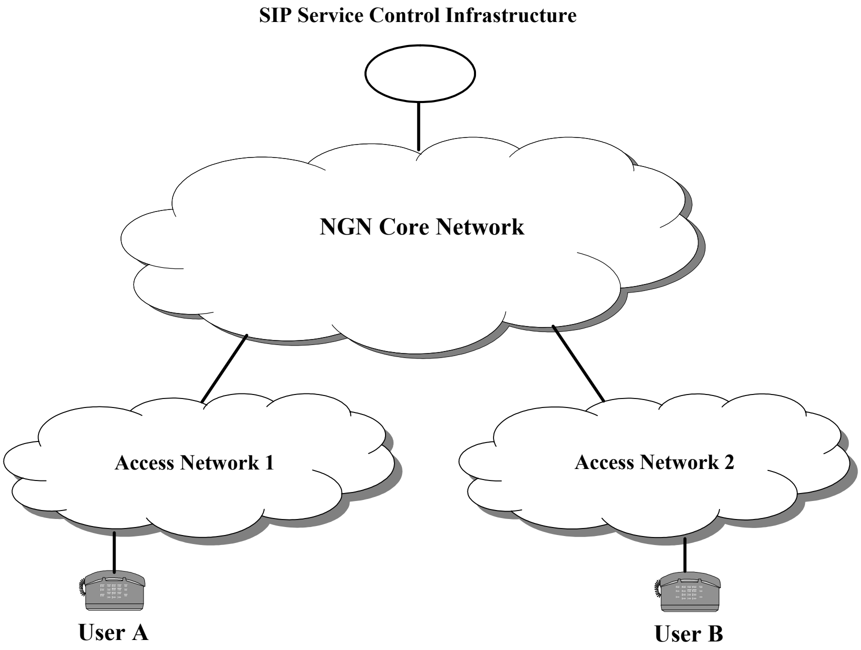


Figure 3.3: NGN reference network structure

User A, subscriber of a respective NGN and, as such, connected to an Access Network 1 intends to establish a communication session with user B, connected to another Access Network 2 as a subscriber of the same NGN. Both access networks are interconnected via the NGN Core Network.

In case the conventional NGN QoS architecture as introduced in section 3.3.2 is applied several relevant sub-entities have to be considered in each network part, that is, in both access networks and in the core network. Figure 3.4 shows the integration of the NGN QoS architecture into this example’s general network structure.

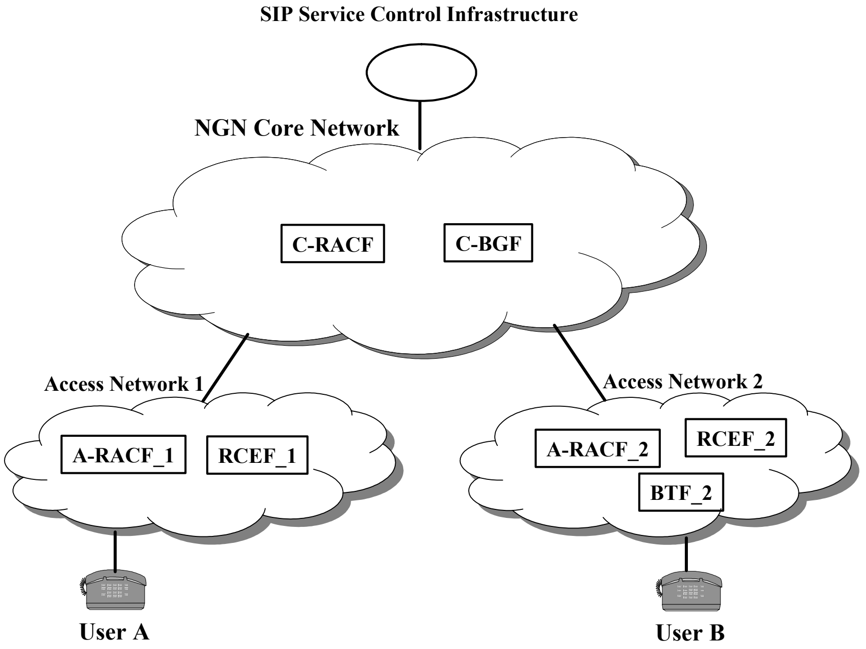


Figure 3.4: Reference network structure including QoS-related entities

Figure 3.5 shows the basic signalling flow related to resource (and hence, QoS) and admission control within the exemplary scenario.

A subscriber end system (referred to as CND (Customer Network Device)) sends a session initiation request (step (1)) to the SIP service infrastructure (referred to as AF (Application Functions)). The AF requests resource and admission control at the SPDF as the centralised part of the RACS (step (2)). The SPDF triggers A-RACF_1 (part of the RACS, outsourced into the related access network) (step (3)) to perform access admission control (step (4)) and to determine policy enforcement (step (5)). Subsequently the A-RACF_1 communicates the policies to RCEF_1 (step (6)) as part of the transfer functions of Access Network 1. The policies become enforced by the RCEF (step (7)) which responds the successful enforcement to A-RACF_1 (step (8)), which in turn reports the success to the SPDF (step (9)). The SPDF now triggers the RACF in charge of the core network (C-RACF) (step (10)). The C-RACF instructs the C-BGF to initiate path-coupled signalling (which typically refers to IntServ/RSVP resource reservation procedures) in order to reserve network resources between the core network and Access Network 2 (step (11)).

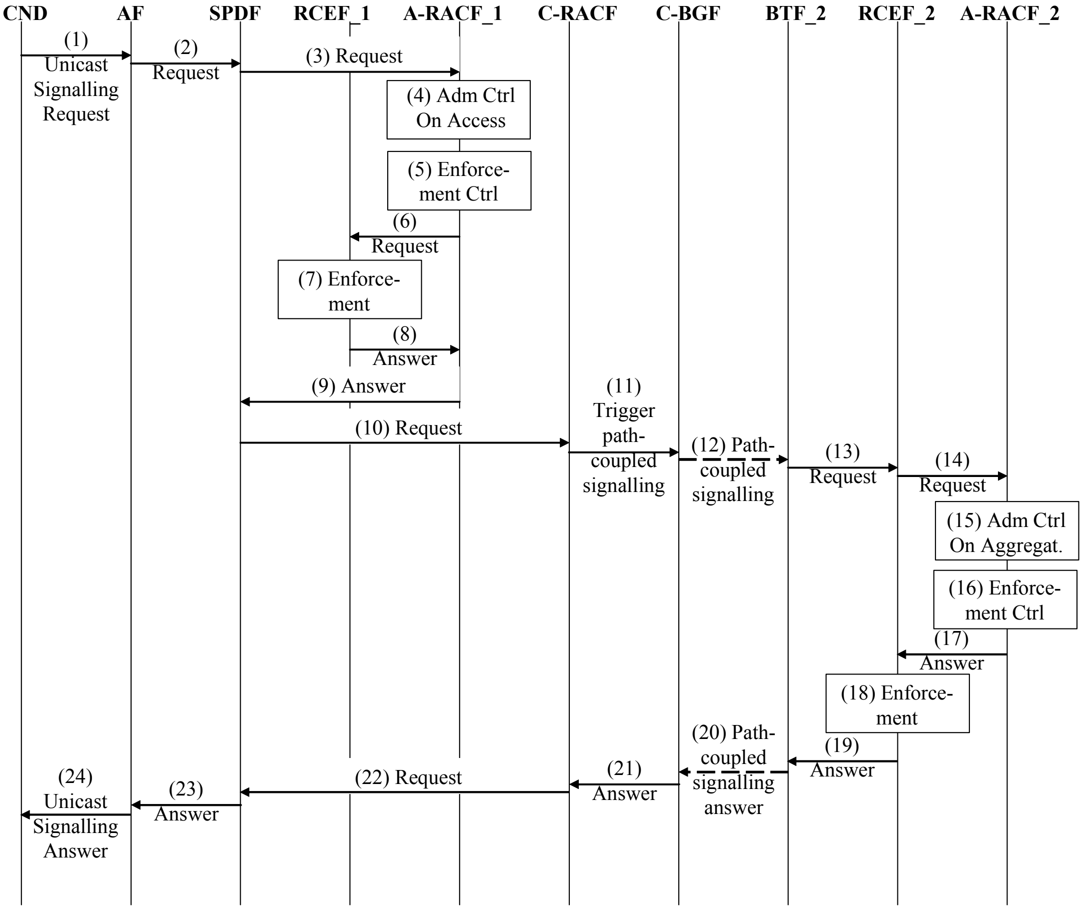


Figure 3.5: Resource and Admission Control signalling flow for the given example scenario (ETSI ES 282 003, 2008)

The C-BGF starts path-coupled resource reservation signalling with the BTF_2 of Access Network 2 (step (12)). Upon receipt of the resource reservation request the BTF_2 requests resources at RCEF_2 (which is assumed to be situated within the same physical node as BTF_2) (step (13)). RCEF_2 contacts A-RACF_2 as the RACS's sub-entity in charge of resource and admission control for Access Network

2 (step (14)). Both admission control and policy enforcement control (steps (15) and (16)) are performed by A-RACF_2, responding to RCEF_2 (step (17)). RCEF_2 installs the policies provided by A-RACF_2 (action (18)) and answers BTF_2 internally (step (19)). In turn BTF_2 provides the path-coupled signalling answer (step (20)) to the request received before (step (12)) from the C-BGF. The C-BGF answers to the C-RACF (step (21)) which, in turn, forwards the answer to the SPDF (step (22)). The SPDF subsequently provides the final answer (step (23)) to the AF. The AF now is able to respond appropriately (step (24)) on the SIP service request originally received by the user end system.

Note that, except the initial SIP service request (step (1)) and the response of the service control infrastructure (step (24)), no service-related signalling is shown in Figure 3.5. In order to accomplish session establishment between subscribers A and B the AF additionally is in charge of forwarding the SIP service request to subscriber B’s end system and, upon receipt of an answer, possibly react appropriately.

Assessment of QoS-relevant signalling traffic effort

In order to assess the effort that comes with the application of the conventional NGN QoS architecture the amount of signalling traffic is estimated. This is performed assuming the following general assumptions in step with actual practice.

- Cross-Stratum signalling (signalling between the SIP service infrastructure (referred to as AF) and RACS components) and signalling between RACS components and Transfer functions, as well as inter-RACS signalling between SPDF and x-RACF, are based on the Diameter protocol (IETF RFC 3588, 2003).
- Both Diameter requests and responses each show a packet length of 300 Byte.

- The signalling flow shown in Figure 3.5 is in principle appropriate for both the reservation and the release of network resources.
- Path-coupled resource reservation signalling is based on RSVP (IETF RFC 2205, 1997). RSVP is used for path reservation in an unidirectional manner. That is, one path has to be reserved for communication from end system A to B and another path has to be reserved in the reverse direction.
- RSVP Path messages show a packet length of 150 Byte. RSVP Resv messages show a length of 300 Byte. Because of the symmetric appearance of RSVP the same amount of Bytes has to be transmitted in either direction (for the data path from end system A to B and vice versa) in order to perform bidirectional path reservation.
- The mean refresh time for each RSVP path is 20 s.
- Both SIP signalling traffic and media data traffic are not considered.
- Based on the assumptions that all Diameter messages come with a similar length and that path-coupled resource reservation traffic is performed symmetrically in either communication direction the same amount of traffic occurs in both directions on every signalling path.

Based on the signalling flow chart provided in Figure 3.5 and the assumed network structure shown in Figure 3.4 the following allocation of signalling messages to network parts has been identified (see Table 3.2). Note that Diameter information exchange is generally based on Request-Response-pairs, each referred to by the number of the related messages as shown in Figure 3.5.

Table 3.2: Identification of network parts involved in QoS-relevant signalling

Type of Traffic	**Affected part of network**		
	Access 1	**Core**	**Access 2**
RSVP		x	x
Diameter message pair 2/23		x	
Diameter message pair 3/9	x	x	
Diameter message pair 6/8	x		
Diameter message pair 10/22		x	
Diameter message pair 11/21		x	
Diameter message pair 14/17			x

In order to prepare for the calculation of the mean bandwidth required for QoS-relevant signalling on each involved network part, in a first step, the packet length values stated above for Diameter messages are converted from Byte to bit (see Table 3.3). The values are added up for each network part under consideration of the respective message pairs affecting the respective network part. For Diameter message pairs all packet length values have to be doubled because it is assumed that the same message sequence is required for both resource reservation and release. For path-coupled resource reservation with RSVP a refresh time of 20 seconds is assumed. Note that, in order to release the respective resources after the media session has been terminated, an additional RSVP signalling cycle is considered.

Table 3.3: Traffic effort required for QoS control

Type of Traffic	**bit per session per network part**		
	Access 1	**Core**	**Access 2**
RSVP		(3600 bit per 20 s) + (3600 bit)	(3600 bit per 20 s) + (3600 bit)
Diameter message pair 2/23		2x(2400 bit)	
Diameter message pair 3/9	2x(2400 bit)	2x(2400 bit)	
Diameter message pair 6/8	2x(2400 bit)		
Diameter message pair 10/22		2x(2400 bit)	
Diameter message pair 11/21		2x(2400 bit)	
Diameter message pair 14/17			2x(2400 bit)

In order to calculate the mean bandwidth incurred for Diameter signalling for a respective media session the number of bits have to be divided by the duration of the respective media session considered. Regarding the path-coupled RSVP signalling, the refresh cycles have to be taken into account.

The results presented in Table 3.4 assume a media session duration of 120 seconds (which is known as a typical mean approximate value for telephone calls).

Table 3.4: QoS control traffic bitrate (session duration: 120 s)

Type of Traffic	bit/s per session per network part (session duration: 120 s)		
	Access 1	Core	Access 2
RSVP		210	210
Diameter message pair 2-23		40	
Diameter message pair 3-9	40	40	
Diameter message pair 6-8	40		
Diameter message pair 10-22		40	
Diameter message pair 11-21		40	
Diameter message pair 14-17			40
Total (bit/s)	**80**	**370**	**250**

Note that, assuming a shorter media session duration, the mean bandwidth required for QoS signalling increases. Substantially this is due to the fact that, considering the Diameter signalling, the same number of bits for both resource allocation and release are transmitted within a shorter period of time. Table 3.5 presents the calculation for a media session lasting for 30 seconds.

Table 3.5: QoS control traffic bitrate (session duration: 30 s)

Type of Traffic	bit/s per session per network part (session duration: 30 s)		
	Access 1	Core	Access 2
RSVP		360	360
Diameter message pair 2-23		160	
Diameter message pair 3-9	160	160	
Diameter message pair 6-8	160		
Diameter message pair 10-22		160	
Diameter message pair 11-21		160	
Diameter message pair 14-17			160
Total (bit/s)	**320**	**1000**	**520**

Figure 3.6 shows the trend for the mean bandwidth required for QoS signalling within this example's core network for 10000 parallel sessions, subject to the mean session duration. This graph is based on the message length values provided in Table 3.3, showing the number of bits required for QoS signalling within the core network for one session, including both, initiation and termination.

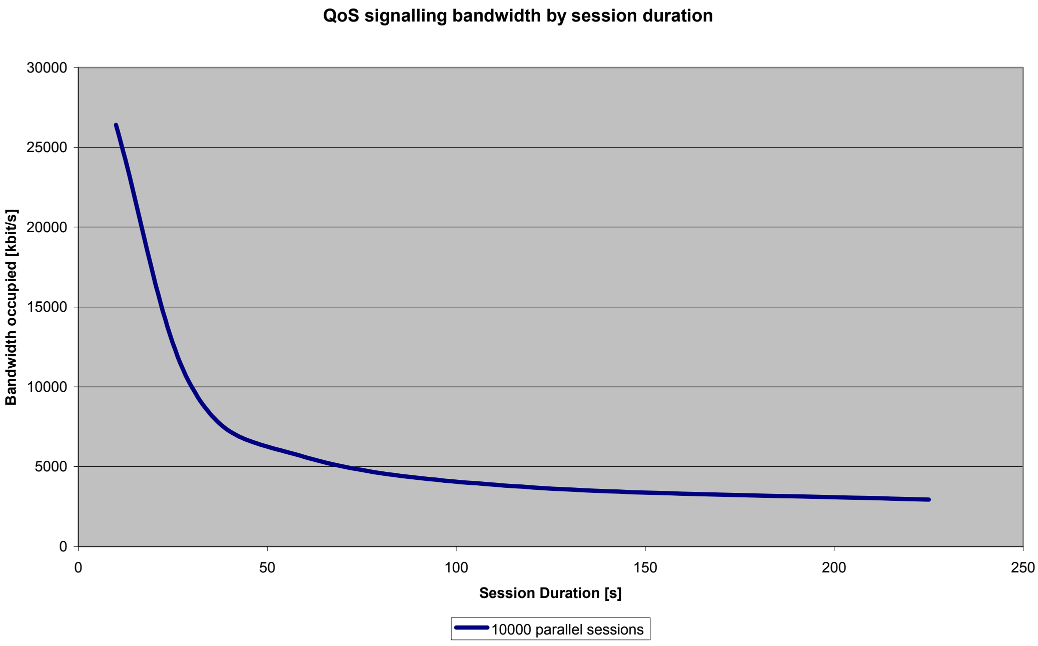


Figure 3.6: QoS signalling traffic effort over mean session duration

It is apparent that the amount of bandwidth required for QoS signalling increases remarkably with decreasing session duration.

Figure 3.7 shows the amount of burst bandwidth required for QoS signalling in the core network during session initiation by the number of sessions initiated within a timeframe of one second. The values are derived from Table 3.3, considering only the amount of bandwidth required during session initiation (four single Diameter messages plus one RSVP path reservation). As a contrast, the horizontal lines show the gross bandwidths required for voice transmissions per direction, considering VoIP streams coded with G.711 (dashed line; 90.4 kbit/s) and iLBC 13.33 (dotted line; 30.9 kbit/s) codec, respectively. It appears that, under the assumptions considered for the above-described example, the burst bandwidth utilised for QoS signalling of almost seven sessions initiated within one second would be sufficient for one G.711-coded VoIP stream. The QoS signalling for three sessions initiated within one second requires more burst bandwidth than the voice data traffic of one iLBC-coded VoIP stream.

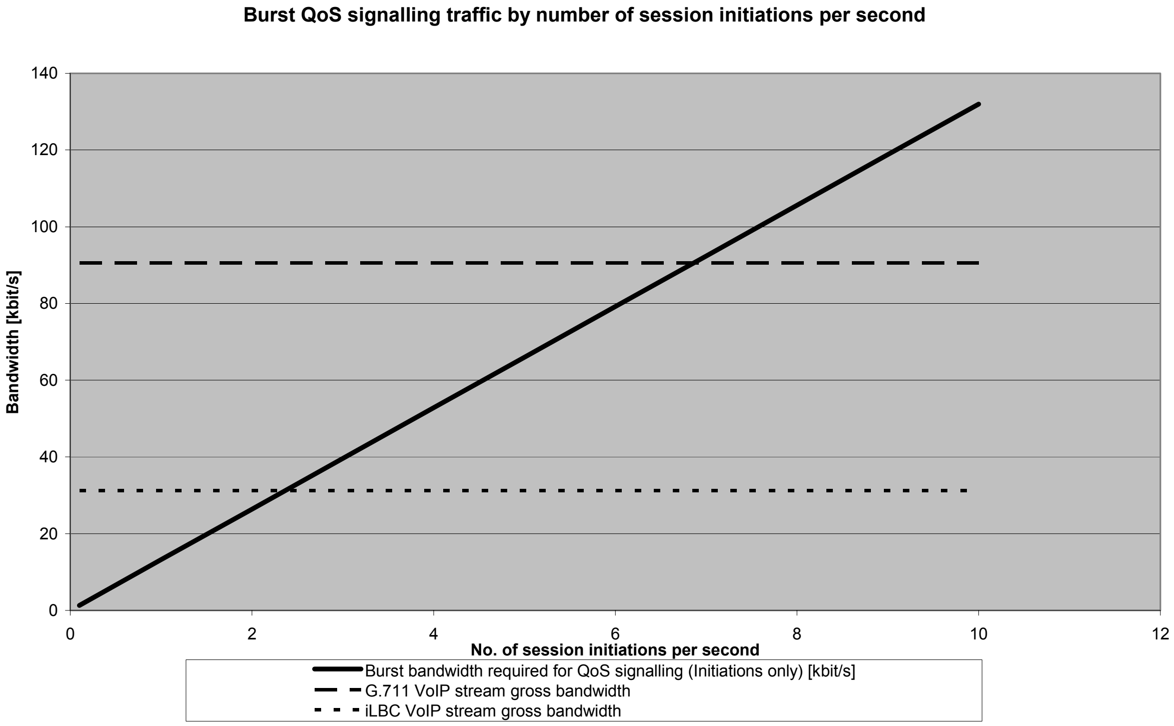


Figure 3.7: QoS signalling burst bandwidth required for session initiations

Summarising it can be pointed out that the application of the assumed reference NGN QoS architecture generally results in remarkable signalling traffic required for QoS and resource management. The considered reference NGN QoS architecture is not scalable with the statistical duration of media sessions established. Note that the amount of burst bandwidth that has to be considered for QoS signalling traffic is significant.

Also note that the amount of QoS signalling traffic generally depends on the network topology and architecture of the respective NGN. In a scenario with a larger number of network nodes involved, even more bandwidth becomes occupied by QoS-relevant signalling traffic.

3.3.4 Alternatives to the conventional NGN QoS architecture

Several research studies focused on the field of QoS provision in SIP-based NGN, and two main research directions became apparent. The majority of work published since the beginning of standardisation by ITU-T and ETSI deals with the investigation, optimisation and/or extension of the NGN QoS architecture as introduced in section 3.3.2. On the other hand, a few independent approaches exist, dealing with arbitrary NGN based on open IP network environments, yet following the NGN concept as introduced in section 2.1. The former approaches rely on a well-defined architecture that provides standardised interfaces and is designed for interoperability from scratch. In contrast, the latter approaches are free to be deployed and extended in an arbitrary manner.

Representing the former research direction, (Vidal *et al.*, 2007) directly refer to the NGN architecture standardised by ETSI TISPAN in a previous version of (ETSI ES 282 001, 2008), proposing additional functionalities for Residential Gateways (RGWs) that connect an end user network infrastructure (such as a home network) to an NGN's access network. The RGW proposed by (Vidal *et al.*, 2007), although initially appearing transparent to the end user equipment and the NGN service and transport infrastructure, is aware of both SIP/SDP signalling and the transport resource utilisation of the respective end user network. As such, the proposed RGW is able to match the transport resource conditions given within the end user network with the conditions required for a requested media session. In case a media session is requested the proposed RGW is in charge of resource reservation and allocation for the data path between the end user network and the NGN access network. Furthermore, the RGW provides an interface for interaction with an RACS (see section 3.3.2) regarding QoS policy exchange. Summarising, the approach proposed by (Vidal *et al.*, 2007) allows for QoS provision and consideration of network resource conditions beyond the control scope of the centralised resource control of NGN as standardised in (ETSI ES 185 001, 2005). However, the reduction of both complexity and scalability issues appearing within the context of the conventional NGN QoS architecture (see section 3.3.2) is not addressed by this approach.

The approaches independently proposed by (Mani and Crespi, 2005) and (Cho *et al.*, 2006) both result in the provision of reliable QoS conditions for every media session on its respective entire data flow path under consideration of several sub-networks involved. While the approach by (Cho *et al.*, 2006) deals with arbitrary NGN providing IP QoS by combining IntServ (for access networks) and DiffServ (for the core network) the approach by (Mani and Crespi, 2005) particularly considers telecommunication networks based on the UMTS IMS (3GPP TS 23.228, 2006) as a service platform (Note that the use of the IMS for SIP service control is recommended for NGN by (ETSI TISPAN ES 282 001, 2008). Hence, this approach is applicable in NGN equipped with a QoS architecture as described in section 3.3.2). Both teams propose the application of specific policy control elements for every involved network segment (for instance, one policy control element per access network connected). This element is in charge of resource and policy control, actuating the resource management of the IP transport elements within the respective network segment. In order to provide QoS on a per-session basis, both teams assume that each policy control element is generally aware of all media sessions (and their respective QoS requirements) affecting the respective network segment. In the approach by (Mani and Crespi, 2005) this is accomplished by integrating a SIP/SDP monitoring entity into the policy control element. (Cho *et al.*, 2006) recommend that the required information is extracted by modified SIP Proxy Servers involved in the session initiation and provided to the policy control element by the use of additional signalling. Both teams assume the application of extensions to SDP for improved capabilities regarding QoS negotiation (such as the nomination of media-dependent QoS-levels and media flow IP packet size). While the implementation of these extensions in SIP end systems is mandatory according to (Mani and Crespi, 2005) the approach of (Cho *et al.*, 2006) also considers SIP User Agents using standard SDP (in this case a translation between standard and extended SDP has to be performed within the modified SIP Proxy Server before forwarding the respective SIP message).

However, in both approaches the policy control elements of each sub-network are in charge of providing QoS to all media data flows on a per-session basis by applying appropriate methods (such as initiating resource reservation procedures or triggering packet marking) specific to the respective IP transport network segments. This is performed by the active control of QoS and resource conditions within the IP transport elements that typically cause additional signalling effort.

In order to avoid the complexity and scalability issues that generally result from the utilisation of DiffServ and IntServ/RSVP in NGN (Park and Kang, 2005) propose the categorical differentiation of only two classes of IP traffic (QoS and Best Effort (BE)). The standardised NGN QoS architecture is not assumed. Rather, for the NGN IP transport network, the application of modified IP routing elements is proposed, providing specific functionalities regarding resource management and routing behaviour. The classification of IP packets into QoS and BE traffic is performed by a new introduced entity (QoS handler) which is located between the end host and the NGN transport network. The QoS handler monitors the resource reservation process initiated by a respective user end system before the initiation of the media session (note that RSVP-like resource reservation is mandatory for this approach, which potentially results in unscalable reservation traffic). If the bandwidth required for a requested session cannot be provided on one or more of the involved IP segments the respective IP routing elements simply deny the resource reservation, and hence the end system requesting the resources is informed of the resource lack within the resource reservation process. Within this approach, the end system is in charge of preventing the initiation of sessions that can not be provided with sufficient bandwidth (and hence, QoS) conditions within the transport network. Note that, because no interaction between service control infrastructure and transport infrastructure is scheduled, this approach does not allow for the centralised control of and reaction on changing QoS conditions. Once a media session has been initiated continuous RSVP signalling is required in order to maintain the reservation state of the set up resource path. Also note that, while making use of lightweight protocol mechanisms to reduce the complexity and, to a limited extend, improve the scalability of QoS provision in NGN, the approach by (Park and Kang, 2005) requires specific IP forwarding elements (which might be problematic regarding their feasibility for real-world NGN IP transport network operators). Also (Mohajerzadeh *et al.*, 2010) consider QoS provision in NGN from a routing network point of view by proposing a novel queue management mechanism for routers, explicitly aiming on congestion avoidance and control in NGN.

As the early standardisation work on NGN mainly focussed on QoS provision supported by access networks, a certain trend was observable of research work addressing the provision of QoS also in the metro and core parts of NGN transport networks, respectively. (Saika *et al.*, 2011) endorse the utilisation of MPLS in transport networks for IMS-based NGN (such as described in (ETSI TISPAN ES 282 001, 2008)). Also explicitly aiming on the standardised NGN QoS architecture, (Cho and Okamura, 2009) propose an extended resource and admission control architecture which makes use of a novel, centralised and hierarchical approach for traffic engineering in MPLS-based core networks of NGN. (Martini *et al.*, 2009) introduce a novel approach for the utilisation of RACF (see section 3.3.2) to configure MPLS network nodes in metro core networks. Therefore, the application of DS-TE (DiffServ-aware MPLS Traffic Engineering; (IETF RFC 4124, 2005)) and NETCONF (NETwork CONFiguration protocol; latest version (IETF RFC 6241, 2011)) is proposed. (David *et al.*, 2010) introduce a novel QoS model for the application in general NGN. This model facilitates QoS provision in MPLS-based networks, and is additionally capable to consider QoS even in the MAC layer (Medium Access Control). Also the NGN QoS optimisation approach from (Matsumoto *et al.*, 2009) aims on the consideration of OSI layer 2 for QoS provision, explicitly focussing on both, access and metro networks. Regarding the interworking between QoS functionalities of the data link layer and the NGN QoS architecture, a SOA-based middleware (Service-Oriented Architecture) is proposed. (Ghazel and Saïdane, 2009) propose an efficient resource-based CAC (Call Admission Control) approach to be applied in NGN. It explicitly focuses on IP/MPLS-based NGN transport infrastructures, making use of the standardised NGN QoS architecture by utilising the RACS for QoS control. According to its authors, the introduced approach improves the scalability and resilience issues of the conventional standardised NGN QoS architecture, resulting in gains in performance and improved QoS conditions.

(Skorin-Kapov and Matijasevicy, 2009) propose a framework extending the standardised NGN QoS architecture, especially focussing on the negotiation and adaptation of end-to-end QoS conditions. Therefore, the general requirement for QoS in a particular communication situation is comprehensively considered from different viewpoints, taking into account the involved users and end systems (such as user preferences and terminal capabilities), application-specific data (such as performance requirements), information regarding both, service and transport provider properties (such as policy rules and network capabilities), as well as regulatory requirements (such QoS-related requirements in emergency calls). Considering these and further preconditions, according to the authors, results in optimised QoS conditions for a considered service used, as well as in a more scalable QoS management behaviour of the respective NGN, compared to NGN equipped with the conventional NGN QoS architecture.

In their paper, (Callejo-Rodríguez *et al.*, 2008) introduce the QoS architecture developed within the IST project EuQoS (End-to-end Quality of Service support over heterogeneous networks), illustrating an alternative approach for QoS provision in IP-based telecommunication networks. A flexible and open framework is specified, capable of meeting QoS requirements end-to-end, even for inter-domain end-to-end QoS provision. This is accomplished by defining a novel QoS service plane, coming independent of the intrinsic provider service plane (such as the SIP infrastructure operated by a SIP service provider). Hence, QoS demands are received and processed separately from service requests by using SOAP (formerly known as Simple Object Access Protocol; (W3C SOAP, 2007)), and hence, are independent of any arbitrary service or service provider, respectively. The QoS service plane interacts with a novel-defined QoS control plane situated below via a specified interface based on the NSIS protocol (Next Steps In Signalling; (IETF RFC 4080, 2005)). The QoS control plane provides a universal interface to the respectively served IP transport infrastructure. Hence, arbitrary IP transport infrastructures, providing arbitrary QoS mechanisms are supported.

Summarising this compilation of research regarding alternatives to and amendments of the specified NGN QoS architecture as introduced in section 3.3.2, any of the considered approaches is described by at least one of the following categories.

- Extension of QoS-related NGN architecture by additional logical entities in order to expand the QoS control range to further network segments (Vidal *et al.*, 2007), (Mani and Crespi, 2005), and (Cho *et al.*, 2006), (Saika *et al.*, 2011), (Cho and Okamura, 2009), (Martini *et al.*, 2009), (David *et al.*, 2010), (Ghazel and Saïdane, 2009)
- Modification or re-modelling of basic IP transport functions (such as IP routers) in order to improve QoS support on the IP transport level (Park and Kang, 2005), (Mohajerzadeh *et al.*, 2010)
- Introduction of traffic engineering in the NGN transport stratum, dynamically controlled by the NGN service stratum (Saika et al., 2011), (Cho and Okamura, 2009), (Martini *et al.*, 2009), (David *et al.*, 2010), (Ghazel and Saïdane, 2009)

- Accomplish NGN-controlled utilisation of QoS mechanisms situated below the network layer (such as layer 2 QoS support) (David *et al.*, 2010), (Matsumoto *et al.*, 2009)
- Optimisation of QoS provision by applying improved methods for network resource requisition, QoS management, or Call Admission Control (Ghazel and Saïdane, 2009), (Skorin-Kapov and Matijasevicy, 2009)
- Alternative approach for both, end-to-end QoS negotiation and provision over arbitrary IP environments (capable of any QoS mechanism), independent of particular applications, specified service or transport network infrastructures, or protocols (Callejo-Rodríguez *et al.*, 2008)

As a conclusion it can be pointed out that, apart from the approaches published by (Park and Kang, 2005) and (Mohajerzadeh *et al.*, 2010), all other considered concepts are based on cross-layer signalling in order to manage resources in the involved transport networks. Hence, regarding their general functional principles, these concepts can be considered similar to the standardised NGN QoS approach introduced in section 3.3.2, potentially coming with similar issues.

While (Park and Kang, 2005) make use of a proprietary (rather lightweight) IntServ approach for resource handling, the approach of (Mohajerzadeh *et al.*, 2010) only aims on the application of an optimised IP routing algorithm. Hence, in order to apply these approaches within an NGN environment, in both cases, the IP transport infrastructure had to become adapted to support the respective proprietary functionalities, signalling and algorithms. Due to the potential costs of a customised IP transport infrastructure, these approaches might not be applicable in real-world NGN.

3.3.5 Issues of common QoS provision approaches

As depicted in sections 3.3.2 and 3.3.3, the application of the assumed reference NGN QoS architecture as specified by ITU-T and ETSI requires the deployment of numerous additional network elements in both, the service and the transport stratum of NGN. These additional entities are mainly required to allow for the control of transport network resources dynamically triggered by network elements situated within the service stratum, being aware of ongoing and requested media sessions, and of their respective QoS requirements. This so-called cross-layer / cross-stratum approach potentially results in an remarkable increase of complexity in both, the network architecture and the signalling traffic effort. In section 3.3.3 the cross-layer / cross-stratum signalling has been demonstrated.

The proposition for the imperative control of transport network resources in order to ensure satisfying QoS conditions for any requested media session is assumed to result from the general awareness of the incoherency of the following characteristics generally given in SIP-based NGN.

- Connectionless packet-based media data transport networks (such as IP networks) showing potentially unsteady capacity utilisation and, as a result, unsteady performance
- Session-based real-time media flows (such as RTP VoIP or Video over IP flows), each requiring constant transport network performance conditions within certain (media- and codec-dependent) parameter limits for the duration of the respective session
- Protocol for call/service control (such as SIP) unaware of and unable to directly control resource and QoS conditions effective within the media data transport network

To avoid potential lack of QoS for media sessions resulting from the discrepancies among the above-mentioned points, in the standardised NGN QoS architecture, common IP QoS mechanisms (such as MPLS or IntServ, see section 3.2.2) are typically applied within the transport network. Those mechanisms are controlled based on decisions made within the service stratum of the respective NGN, which therefore have to be linked to the respective functions within the transport network. This linking is performed by signalling between those strata (cross-stratum signalling; see section 3.3.1), typically utilising DIAMETER as a signalling protocol.

In section 3.3.4, further approaches for QoS provision in NGN have been discussed, coming as alternatives or enhancements to the conventional NGN QoS architecture, respectively. Except of one case, all approaches involve cross-layer / cross-stratum signalling, resulting in the drawbacks stated above (rather, the exceptional approach utilises IntServ, and special IP routing equipment is required for media stream transport). Hence, although considering several different issues, none of the approaches considered eliminates the major problem of complexity and unscalability generally concerning QoS provision in NGN.

3.3.6 Requirements for alternative QoS provision in SIP-based NGN

Generally, an alternative approach for QoS provision in NGN must consider the fundamental requirement that an NGN service provider is in charge of the QoS provided to subscribers involved in media sessions. In order to satisfy this requirement, the following side conditions have to be fulfilled by the proposed solution framework.

- As a crucial prerequisite, in order to be able to meet the conditions declared in the following, the operator of an NGN service platform must be aware of the QoS experienced by any of its customers involved in an ongoing media session at any time.
- The operator of an NGN service platform must ensure that only those sessions are granted which presumably will experience adequate QoS throughout their lifetime.
- Once having granted a session, the operator of an NGN service platform must be able to manage (and, if required, recover) the QoS provided to the customers involved in this session.

This research has identified the following main requirements for an alternative approach for QoS provision in SIP-based NGN, as published in an earlier form as (Weber *et al.*, 2007).

- The new approach should result in a radically simplified QoS architecture.
- Functions and mechanisms are required that result in satisfactory QoS for any established session and, at the same time, do not require cross-signalling between service and transport stratum of the respective NGN.
- Simple and resource saving mechanisms for QoS optimisation should be preferred over complex architectures coming with traffic-intensive mechanisms. If possible, approaches should rely on already standardised protocols (such as SIP) and architectures (such as the general NGN architecture according to (ITU-T Y.2012, 2010) and (ETSI ES 282 001, 2008)). Standard network components should be applicable in both, transport and service stratum.
- NGN QoS optimisation should be aware of non-session traffic (such as TCP web traffic).
- The QoS provision in NGN should be independent of underlying transport technologies and respective QoS mechanisms, such as IntServ or DiffServ on

the IP level, or MPLS, ATM, and VLAN in lower levels. Hence, arbitrary IP network architectures should be supported, regardless their specific QoS mechanisms. However, the auxiliary applicability of such mechanisms, if available, should not be excluded.

- The additional traffic effort for QoS provision should be generally predictable, and should not exceed the amount of traffic that would be generated in a comparable scenario applied to the assumed reference NGN QoS architecture.

3.4 Summary

Within this chapter, the term Quality of Service has been introduced and defined in the sense of telecommunication services. Within this research, the bidirectional real-time speech transmission service (voice telephony) has been chosen as a reference service due to its comprehensive occurence and its fundamental role as the most common telecommunication base service. The characteristics of this real-time service regarding its specific requirements for QoS have been outlined within this chapter, introducing the three major (packet) network performance parameters delay, jitter, and packet loss. Limiting values for these parameters have been declared. Regarding the measurement, evaluation, and estimation of the quality of a speech service, essentials have been discussed such as the Mean Opinion Score (MOS), and testing / measurement modes such as PESQ, 3SQM and further alternatives were outlined. The ITU-T E model has been introduced, which can be utilised for QoS estimation. Its output consists of a rating factor R, which can be converted to a MOS value and, as such, provides a basis for the estimation of the quality provided for a voice service as perceived by the user, based on the performance of the respective network.

Furthermore, a broad overview has been presented regarding methods and mechanisms to facilitate acceptable QoS for real-time services provided especially over IP networks. After considering network independent functions for QoS improvement such as Endpoint Admission Control, benefits and limitations of specific IP QoS mechanisms such as DiffServ, IntServ, and MPLS have been discussed. Focussing on SIP-based NGN, the standardised NGN QoS architecture has been introduced, as specified by ITU-T and ETSI. A reference scenario has been assumed and analysed regarding scalability and traffic volume efforts, discovering potential disadvantages of the standardised QoS architecture, which comes along with a sophisticated infrastructure, potentially prone to scalability issues and deficiencies.

Also within this chapter, a broad selection of alternative approaches for reliable QoS provision in SIP-based NGN has been compiled and discussed, most of which have emanated from the academic community.

It has been outlined that an inevitable incoherency exists regarding characteristics generally given in SIP-based NGN. Amongst others, as one main achievement of the research documented within this chapter, the substantial overhead required for cross-layer QoS provision approaches has been identified. The main outcome of this chapter results in a set of requirements for alternative approaches for QoS provision avoiding cross-layer signalling, which provides the basis for the framework introduced in chapter 4.

4 Proposed QoS optimisation framework concept

This thesis proposes a novel QoS optimisation framework that fulfils the requirements stated in section 3.3.6, and hence solves the identified issues of conventional QoS provision in NGN, which were exposed in section 3.3.3. This chapter begins by defining the preconditions and tasks to be considered by the framework (section 4.1), followed by the introduction of both the functional principle of the framework and the underlying concepts (section 4.2). Subsequently, the framework components are presented (section 4.3), followed by descriptions of the protocol procedures that come with the framework application, including the communication required for logical grouping of media sessions and CAC support (section 4.4), QoS estimation (section 4.5), and the optimisation of QoS conditions (section 4.6).

As a simplification, NGN components not required for and not directly affected by framework functionalities are not considered within the following framework descriptions.

4.1 Framework preconditions and tasks

This subsection specifies preconditions for and tasks to be fulfilled by the integrated framework for comprehensive QoS optimisation in SIP-based NGN.

Generally the proposed framework has to fulfil the requirements for optimised NGN QoS provision stated in section 3.3.6. The main prerequisite to satisfy these requirements is to minimise the level of complexity required for QoS management and control in SIP-based NGN. Therefore cross-stratum signalling (see section 3.3.1) is explicitly avoided. Hence, the control scope of the QoS optimisation framework is limited to the service stratum, which is referred to as **passive QoS control** throughout the thesis. This results in the provision of **soft QoS conditions**, which means that the framework aims at defined quality levels for all sessions but, however, cannot guarantee for the targeted level. The **active QoS control** of transport network resources, which would result in the adherence to **strict QoS conditions**, as offered by the standardised NGN QoS architecture introduced in section 3.3.2, is explicitely not considered within the scope of the proposed framework.

The proposed framework should provide the following tasks for QoS optimisation.

- Task 1: Non-invasive mechanism for the continuous estimation of the QoS experienced by any ongoing media session in order to detect insufficient QoS.
- Task 2: Integration with Call Admission Control
- Task 3: Non-invasive mechanism for the optimisation of the QoS experienced by users involved in media sessions.

Within the following sections, an overview on the proposed solution framework is given, and its components and functions are introduced. It is shown that the framework for comprehensive QoS optimisation in SIP-based NGN fulfils the above-mentioned tasks and, hence, meets the requirements and conditions defined in previous sections of this thesis.

4.2 Framework functional principle and underlying concepts

Assuming idealised user terminals providing approximate perfect real-time media recording, processing, and play-out, it is expected that coexistent media sessions can be said to experience similar QoS conditions if they are forwarded in parallel over those paths of a transport network which, as a whole, cause the essential portion of the network performance similarly affecting the QoS conditions encountered by any of these sessions. The proposed framework utilises this effect

- for the determination of QoS conditions that similarly affect a number of media sessions through the retrieval of QoS information from only a few selected media receivers,
- for the optimisation of QoS conditions that similarly affect a number of media sessions. This is performed through variation of the overall bitrate on the concerned network path by modifying the characteristics of only a selected subset of sessions, and
- for the support of CAC decisions with respect to the QoS conditions that the requested sessions will presumably encounter.

This thesis proposes a method to identify media sessions having in common similar QoS conditions due to being affected by similar network performance. The respective process of identification and, subsequently, logical classification of media sessions into groups (whose member sessions encounter comparable QoS conditions) is referred to as **QoS profiling**, which is described in detail in chapter 5. The principle of QoS profiling, proposals regarding reference implementations and extensions, as well as deployment descriptions and results of several field trials and simulations have been been published in (Weber *et al.*, 2009a), (Weber *et al.*, 2009b), and (Weber *et al.*, 2010).

Determination of QoS conditions

The following example illustrates the described effect. The underlying scenario is depicted in Figure 4.1.

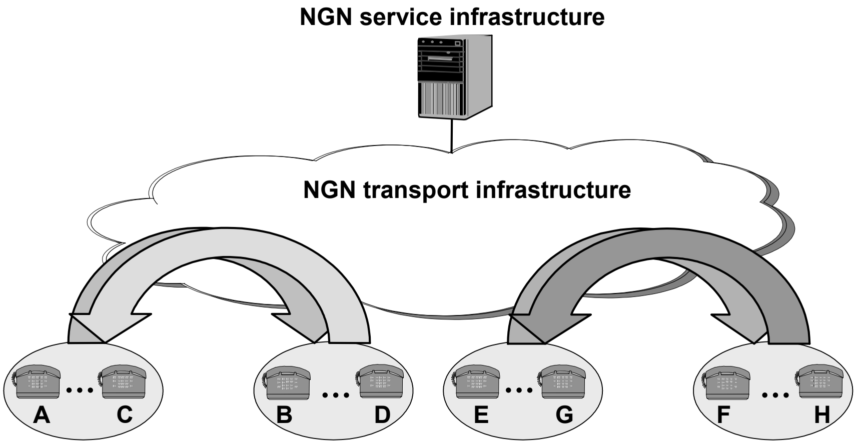


Figure 4.1: Scenario for QoS profiling application

Be it that two NGN subscribers A and C exchange media sessions (such as in VoIP telephony calls) of the same medium and codec with subscribers B and D, respectively. In case that the media sessions from B to A and from D to C share the same network path being the main factor for the network performance equivalently encountered by both sessions, the sessions from B to A and from D to C can be said to experience similar QoS conditions. In this case, in order to collect information about the QoS experienced by subscribers A and C for their respective media sessions, it is sufficient to obtain QoS-relevant information (such as the network performance parameters delay, jitter, and packet loss rate) of one of these two media sessions. This information is likewise meaningful for the QoS experienced by both subscribers A and C for their respective session, and by any potential subscriber whose received media stream is affected by the same QoS conditions. Thus, it can be said that if two or more media sessions can be identified sharing the same network path which decisively determines the network performance experienced by these sessions, one of the sessions can be chosen as a reference for the QoS experienced by both sessions.

However, conclusions on a perceived QoS can not be drawn from information on the QoS experienced by a user which receives media transmitted in the opposite direction over the same network path. Hence, still considering the above-described scenario, the media session A→B will presumably encounter similar QoS conditions as the session from C→D, but not necessarily similar conditions as the media session B→A. Furthermore, another pair of communicating subscribers (E and F) will typically not experience similar QoS conditions as the pairs AB and CD, if the overall network path is not shared for this session. However, media sessions between E and F might experience similar QoS conditions as a session between another pair of communication partnerns, G and H.

Hence, assumed that all media sessions have been classified into groups by the similarity of encountered network performance, it is sufficient to monitor the QoS encountered by only a subset of all sessions to obtain a comprehensive overview on the QoS conditions experienced by any media-receiving subscriber. This effect is utilised within the proposed framework for the resource-saving but comprehensive determination of QoS conditions. To achieve this goal, a concept had to be developed to select the most appropriate reference sessions per group. Furthermore, a concept had to be developed for the centralised continuous estimation of QoS under real-time conditions. These methods are introduced in detail in chapter 6.

Optimisation of QoS conditions, and CAC support

Be it that the overall traffic volume on a network path, which is shared by a number of media sessions, essentially influences the network performance of this path. In this case, reducing the traffic volume injected by a subset of all media sessions would help to restore or optimise potentially degraded QoS conditions for any media session which is forwarded over this path. This effect is utilised within the proposed framework for the optimisation of QoS conditions and CAC support. Therefore, a concept had to be developed for the centralised feedback control-based decision-making regarding the need for and the extend of QoS optimisation action. Furthermore, a concept had to be developed for the selection of communication sessions adversely affected by QoS optimisation action. These methods are introduced in detail in chapter 7.

4.3 Framework components

Within this section, an overview of the components of the solution framework is given, and the sub-entities and functionalities of these components are described. Note that, for simplicity reasons, a single SIP Call Server is used throughout the remainder of the thesis as a simplifying symbol for any SIP-based service provider infrastructure (which may also consist of an IMS as proposed by ITU-T and ETSI for the use in NGN).

4.3.1 Framework components overview

To integrate the proposed framework with an NGN, in order to provide the functionalities generally described in section 4.2, an additional central component has to be connected to the NGN. This component termed **QoS Manager** is required to retrieve and analyse information on the QoS experienced by media stream receivers. Furthermore, based on QoS analyses results, the QoS Manager triggers the optimisation of QoS conditions through the modification of media-related parameters such as codec selections. Therefore, it controls the renegotiation of media codecs between concerned user terminals. Furthermore, the QoS Manager provides CAC support to the central SIP infrastructure. The QoS Manager is supported by a central database function.

In order to provide QoS information to the QoS Manager, but also to reliably execute session modification decisions made by the QoS Manager, specific functionalities are required at the media receiver sides. Therefore, in order not to limit the choice of user terminal equipments, intermediary entities are defined which are logically interconnected between the user terminals and the access networks. These entities are termed **User Access Gates** (**UAG**s). Providing a communication interface for the QoS Manager, the UAGs collect information on the network performance that affects the QoS perceived by the users. Furthermore, by the order of the QoS Manager, UAGs supervise the modification of media-related session parameters such as codec choices between communicating user terminals. Note that a UAG might come as a separate device, or it might come integrated with an IAD (Integrated Access Device) or with a user terminal such as an IP telephone device.

Figure 4.2 illustrates the integration of both QoS Manager and UAGs with an NGN.

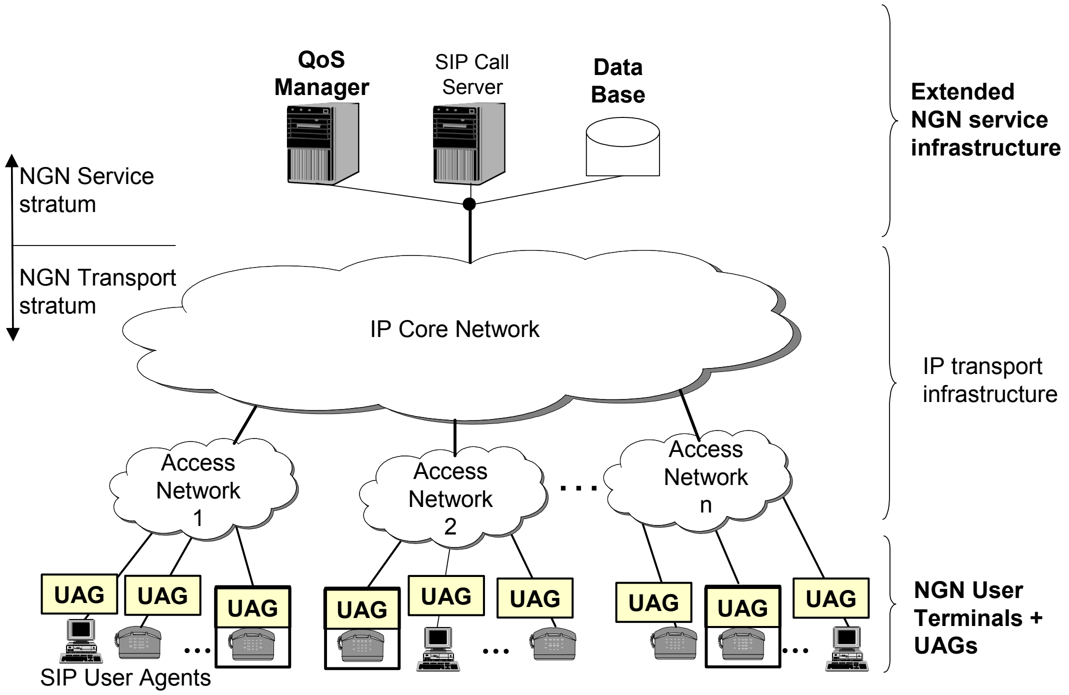


Figure 4.2: Integrated framework for comprehensive QoS optimisation in SIP-based NGN (Weber and Trick, 2008)

Within the transport stratum, the NGN might consist of an IP Core Network (CN), typically connecting a number of Access Networks (ANs). The IP transport infrastructure may be based on any arbitrary technology (such as Ethernet, MPLS, ATM in the CN and Ethernet, DSL or DOCSIS in the ANs) and may or may not come with respective QoS mechanisms (such as VLAN (Virtual Local Area Network), DiffServ, IntServ/RSVP, or any technology-dependent mechanism). The framework introduced within this chapter is designed to provide reasonable QoS without making use of any QoS mechanism which may be available in the transport stratum (however, if desired, the framework can be easily extended to support the utilisation of transport network QoS mechanisms). Therefore it is assumed that the IP transport infrastructure, including ANs and CN, is designed to generally provide sufficient resources according to the expected traffic volume, considering resource requirements coming along with respective use cases (such as flat rate customers), offered services (such as SIP session management and Internet access), and different types of media which may be available for the users (such as VoIP and video telephony).

4.3.2 Framework component layout

Within the following paragraphs illustrate in more details the layout of these components. Details of the interworking of QoS Manager and UAGs and their included sub-entities are further illustrated in the context of the overall framework functionalities in sections 4.4 to 4.6. Therefore the application of the concepts introduced in chapters 5, 6, and 7 is assumed, namely QoS profiling, profile-based QoS estimation, and feedback control-based QoS optimisation.

QoS Manager

This centralised entity covers the functionalities shown in Figure 4.3. It is provided with interfaces to be directly connected to both, the NGN Call Server and a database which keeps QoS-related data processed by the QoS Manager. A third interface is provided to connect the manager to the NGN IP network to allow for information exchange with several UAGs.

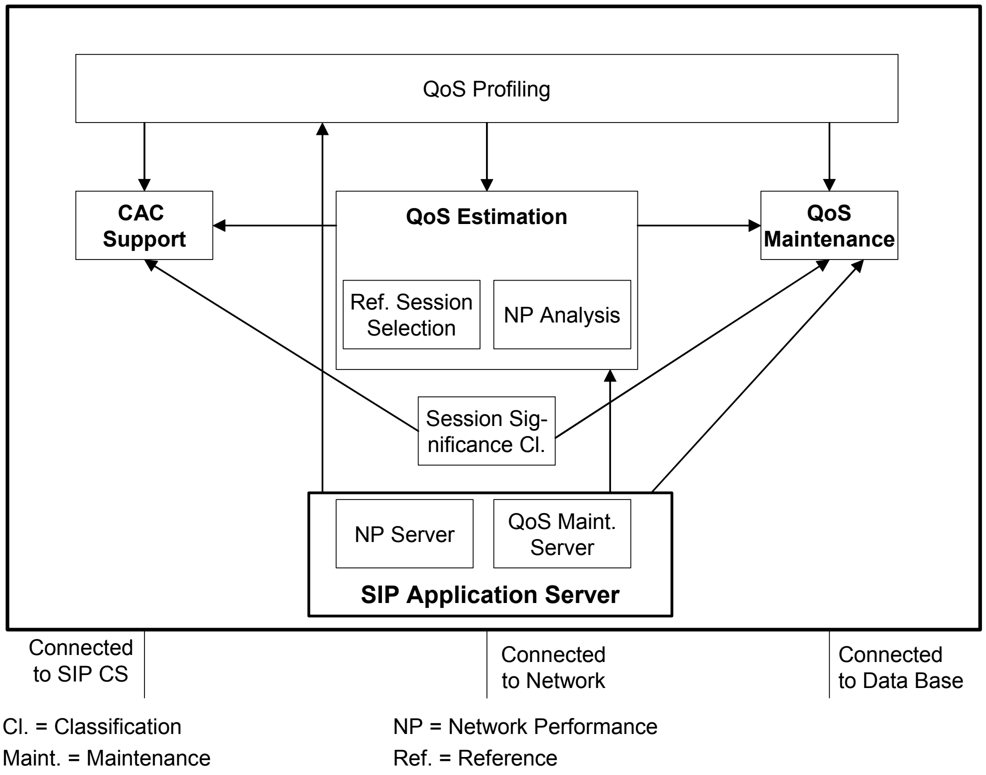


Figure 4.3: QoS Manager block diagram

The QoS Manager consists of the following logical components.

- SIP Application Server as an interface for SIP-based information exchange with the UAGs. The included Network Performance Server provides QoS-relevant data received from UAGs to both, the QoS profiling and the QoS Estimation unit. The included QoS Maintenance Server provides a top-down interface for the enforcement of QoS-optimising decisions made by the QoS Maintenance unit within the UAGs.

- QoS Profiling unit, allowing for efficient and resource-saving QoS estimation and management.

- QoS Estimation unit, to assess QoS conditions of both, recently requested and already existing media sessions. It includes two sub-entities, providing Network Performance analysis and the selection of Reference Sessions for QoS monitoring.

- CAC Support unit, facilitating QoS-related Call Admission Control decision performed by the NGN Call Server.
- QoS Maintenance unit for the active control of QoS conditions provided to ongoing media sessions.
- Session Significance Classification, which facilitates both, CAC Support and the QoS Maintenance.

User Access Gate (UAG)

A UAG covers the functionalities shown in Figure 4.4 and must be trusted by the SIP service provider. In order to deploy the introduced QoS optimisation framework, every point of access to the respective NGN is equipped with a UAG.

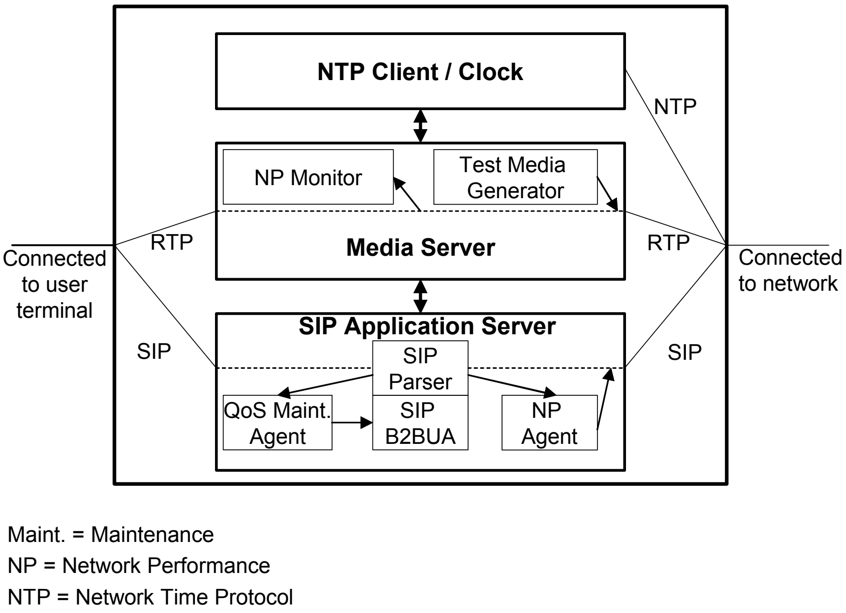


Figure 4.4: User Access Gate (UAG) block diagram

A UAG consists of the following logical components.

- SIP Application Server, to allow for the interpretation and mediation of SIP signalling originating from both the user terminal and the central SIP infrastructure of the NGN, including the QoS Manager. The SIP AS includes a SIP parser for the analysis of SIP messages, a Network Performance (NP) Agent for the SIP-based transmission of QoS-related information upon request of the QoS Manager, and a QoS Maintenance Agent and SIP Back-to-Back User Agent (B2BUA) to enforce and negotiate the modification of media-related parameters with the user terminal on behalf of the QoS Manager.

- Media Server, including a Test Media Generator for the generation of media streams required for pre-session QoS profiling, and a Network Performance Monitor to analyse the network performance by which incoming media streams are affected.
- NTP Client / Clock, required to provide synchronised time stamps for network performance data that are transmitted to the QoS Manager.

4.4 Initial profiling and Call Admission Control integration

The framework for comprehensive QoS optimisation provides a Call Admission Control support mechanism. This mechanism allows for the consideration of the network performance of a concerned path through the transport network within the Call Admission Control (CAC) process. Note that, from a general point of view, beside QoS concerns, the CAC decision taken by a service provider instance may be influenced by additional factors which are not considered within this thesis.

The application of the introduced CAC support mechanism requires that those media sessions that would result from a granted communication have to be initially assigned to virtual groups of media sessions that are affected by similar QoS conditions.

Regarding QoS, the following two factors can be considered within the process of Admission Control.

- The influence of the given network performance (and hence, QoS conditions) on the requested session (will the session experience adequate network performance if granted?)
- The potential impairment of the requested session on the network performance (and hence, QoS conditions) of coexistent sessions (will the requested session impact the network performance experienced by other sessions?)

In order to take into account either of these cases, the conditions (the effective network performance and the joint utilisation by other sessions) on the concerned network path must be determined. This is afforded by the CAC support mechanism described within the following.

Figure 4.5 shows the first part of the process of CAC support as provided by the introduced framework.

User A wants to initiate a session to User B. Therefore, the SIP User Agent of A (UA A) sends a SIP INVITE request (1) to the NGN Call Server (CS). Note that the UAG of A acts transparently regarding the INVITE request. The CS answers the received INVITE with the SIP response 100 Trying (2) and queries the QoS Manager (3) for QoS-related CAC regarding the requested session. Note that any appropriate protocol such as Diameter is proposed to be used for the CAC-related signalling.

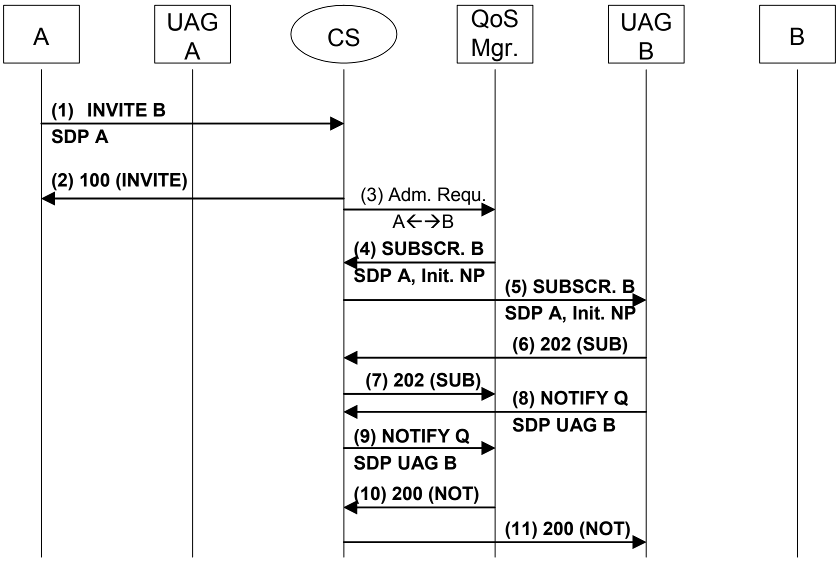


Figure 4.5: CAC support of the solution framework, part 1

Within the QoS Manager, the CAC request is processed by the CAC Support entity. The QoS Manager now has to determine the conditions on the respective network paths which would be involved in the media exchange between user terminals A and B. This is performed by utilising SIP-specific event notification (see section 2.2.2). The Network Performance (NP) Server, being a component of the SIP Application Server entity of the QoS Manager, generates the SIP request SUBSCRIBE (4), including the SDP media description which UA A has originally transmitted in its INVITE request. Note that the CS has to extract the media description, and make it available to the QoS Manager, typically by filing it to a Data Base function.

Sending the SIP SUBSCRIBE request (4), the QoS Manager literally subscribes to the network performance that user B would experience for a media session originating from user A. Note that, according to (IETF RFC 3265, 2002), the type of event for which a subscription is requested has to be defined within the event header field of the respective SUBSCRIBE request. Therefore, regarding the initial subscription to the network performance experienced by a media session, an event named *NPExcerpt* is defined.

The SUBSCRIBE request (4), sent by the QoS Manager and addressed to UA B, is forwarded by the CS (5) and is intercepted by the UAG of B. The interception is performed by the SIP Parser of the SIP Application Server included in the UAG, applying the filter rule that any SUBSCRIBE request for an event named *NPExcerpt* is meant to be processed by the UAG itself and hence, has not to be forwarded to the user terminal.

The request is internally passed to the NP agent of UAG B which accepts the subscription (6), (7) and generates the compulsory NOTIFY request (8). This request is addressed to the QoS Manager (Q) and includes an SDP media description which is assumed by UAG B to be the most adequate answer to the media description of A (which is the offer in the sense of the SDP offer/answer model according to (IETF RFC 3264, 2002), (IETF RFC 6337, 2011)). The NP Server entity of the QoS Manager, upon receipt of the NOTIFY request (9), sends a 200 OK response (10), (11).

In the next step, in order to consider the bidirectional manner of a media exchange, the QoS Manager subscribes to the network performance experienced by A when receiving media from B. Therefore another SUBSCRIBE request is generated (see Figure 4.6) (12), defining the *NPExcerpt* event within the SIP event header field so that the parser of the UAG of subscriber A can identify this request to be intercepted and processed by itself (13). The request also includes the media description received from UAG B in (9).

The NP agent of UAG A accepts the subscription to the *NPExcerpt* event (14), (15) and sends the compulsory NOTIFY request (16), (17) which is answered by the QoS Manager (18), (19).

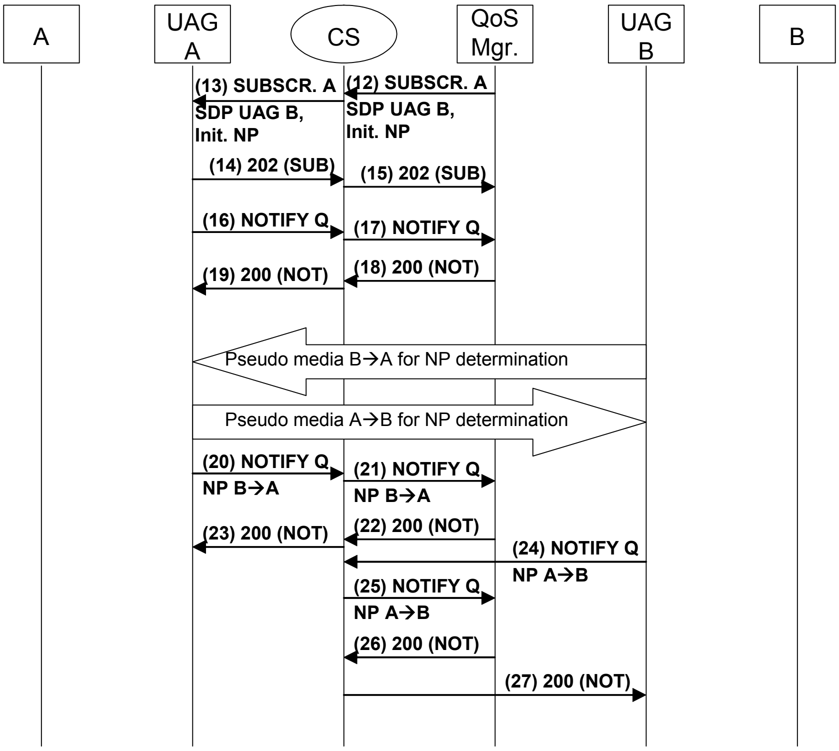


Figure 4.6: CAC support of the solution framework, part 2

Now that both UAGs are aware of the media descriptions of the respective remote party, their included Media Servers generate and exchange pseudo media streams of defined length, consisting of RTP packets with arbitrary payload. Note that, in order to set up conditions similar to the real media stream exchange to be followed, the UAGs have to derive from the given SDP media descriptions the relevant parameters such as desired media and codecs, and apply their respective characteristics such as payload size per packet and packet intervals. This is performed by the NP agents included in the UAGs.

While receiving the media packet sequence from the respective remote UAG, both UAGs monitor the network performance in terms of delay, jitter, and packet loss rate. Note that, as it is assumed that all UAGs connected to an NGN are synchronised by the use of NTP (Network Time Protocol; (IETF RFC 5905, 2010)), the timestamp of the RTP packets (providing absolute NTP time according to (IETF RFC 3550, 2003)) can be used to determine the delay by which each packet has been affected on its way through the network from a sending UAG to the respective remote UAG. Therefore an NTP infrastructure providing sufficient synchronisation accuracy for all UAGs must be available within the respective NGN.

Following the network performance monitoring phase, the NP agents of both UAGs send SIP NOTIFY requests (20), (21) and (24), (25) to provide the respective network performance characteristics to the QoS Manager which confirms their receipt with 200 OK SIP responses (22), (23) and (26), (27).

The QoS Manager now applies QoS profiling to the media sessions of either communication direction (A→B and B→A, respectively). Hence, it assigns the pseudo media sessions to groups of sessions each of which is affected by similar network performance characteristics (see Figure 4.7). If no group exists representing the network performance characteristic of a considered media session, a new group is opened up with the considered session as its first member.

After having assigned both media sessions to a respective group, QoS estimation is performed, and a QoS-related Call Admission Control decision is made. Therefore different factors are taken into account, such as the QoS experienced by members of the considered groups, or the potential effect of a further media session on the QoS characteristic of the group. The CAC decision process is also influenced by the objective significance of the sessions as determined by the Session Significance classification entity of the QoS Manager (see section 7.3.2). However, the CAC result is provided to the Call Server (28).

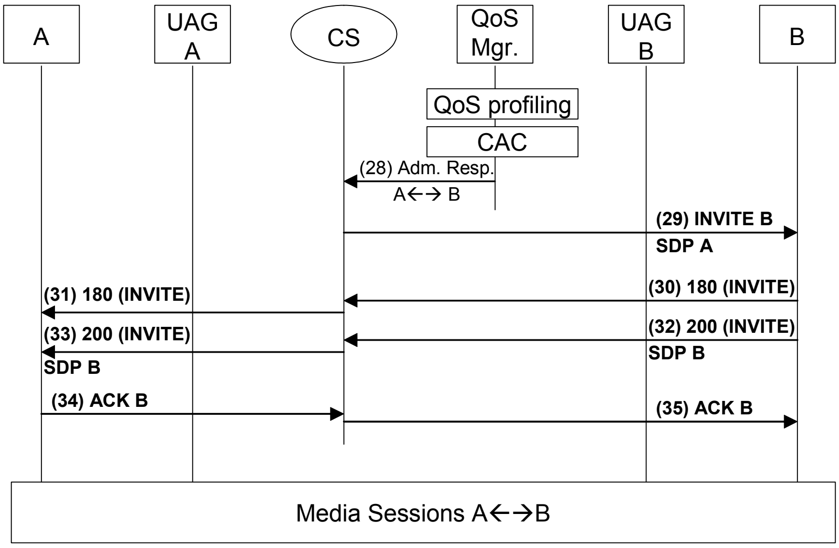


Figure 4.7: CAC support of the solution framework, part 3

Given that the CAC process has resulted in a “grant” decision, the Call Server forwards the SIP INVITE request (29) to user terminal B (UAG B appears transparent). The following exchange of SIP messages (30) to (35) represents the standard procedure of a SIP session initiation with a stateful SIP Proxy Server involved, resulting in a media exchange between user terminals A and B. Both UAGs appear transparent for the media streams. However, they are involved in network performance monitoring to deliver characteristics for the re-grouping of existing media sessions (see section 4.5.1), and for continuously generating network performance characteristics in case the respective media session is chosen as a group reference (see section 4.5.2).

4.5 Comprehensive QoS estimation

This section proposes two mechanisms that, when combined, form the essential function of comprehensive non-invasive QoS estimation provided by the introduced framework. These two mechanisms are

- the algorithm for updating the classification of similar sessions into clusters
- the procedure for continuously receiving QoS information related to a group beyond the lifetime of a reference session.

4.5.1 Recurrent mid-session verification of group affiliations

Before a CAC decision is made, every media session has been classified into a group of sessions which are affected by similar network performance characteristics, and hence QoS conditions (see section 4.4). Since the process of classifying sessions by the similarity of their network performance characteristics is potentially subject to statistical uncertainties, the classification of media sessions to their respective groups should be repeated throughout the sessions' existence. This is already considered by the classification process initiated in line with CAC. By subscribing to the *NPExcerpt* event, the NP Server entity of the QoS Manager instructs UAGs to provide excerpts of characteristics of their encountered network performance not only for the initial classification process but also – within defined intervals – for the verification of the originally group affiliation. Figure 4.8 shows the respective transactions. For the sake of clarity, only the media session B→A is considered within both, the following description and Figure 4.8.

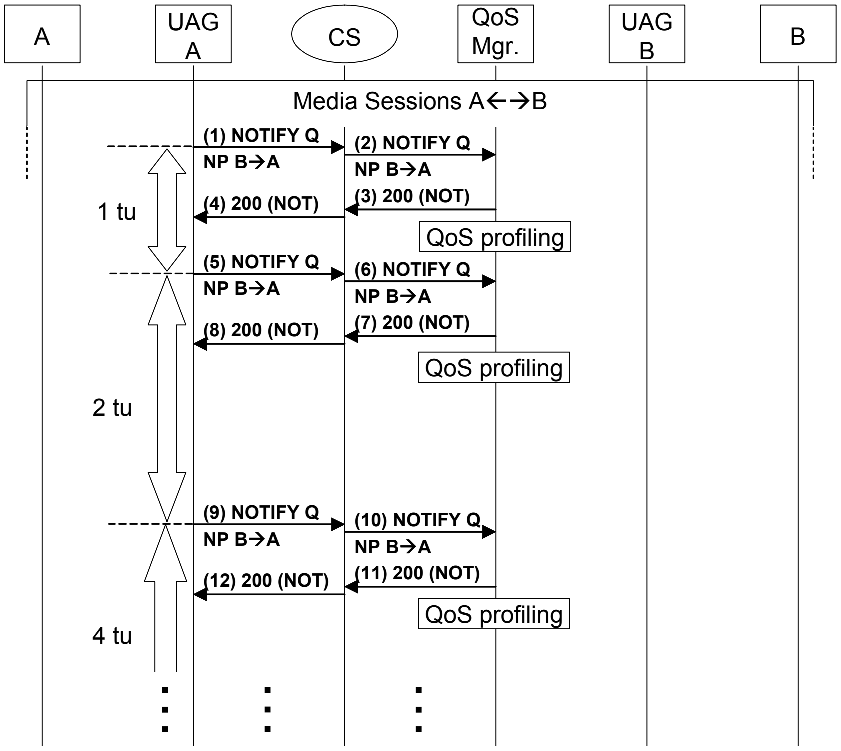


Figure 4.8: Mid-session verification of group affiliation

Be it that a SIP session has already been established between subscribers A and B. In this case, both media sessions have already been classified into a respective group throughout the CAC process. As shown in section 4.4 this process is triggered by the QoS Manager that subscribes to the *NPExcerpt* events of those UAGs receiving media streams. After the session is established, the UAGs again repeatedly record the network performance encountered for the respective session, utilising their internal NP Monitor. An NP excerpt is created and passed to the QoS Manager in a further SIP NOTIFY request which is sent by the UAG's NP Agent component (see Figure 4.8, steps (1), (2)). Upon receipt the NP Server entity of the QoS Manager acknowledges (3), (4) and provides the NP data to the QoS Profiling unit in order to verify the group affiliation of media session B→A. Given that the re-classification result confirmed the initial assignment, the UAG automatically provides an updated sample of network performance data after 1 time unit (tu), resulting in the same procedure as described above (steps (5)…(8)). Note that the absolute length of this interval is due to be set by the NGN service provider. Its value should be chosen depending on the accuracy experienced for the QoS profiling process (which depends on the overall behaviour of the transport network).

Unless otherwise instructed by the QoS Manager, the UAG doubles the interval times after each successful re-verification. Hence, before the next verification procedure (steps (9)…(12)) the UAG waits for a period of 2 tu, and 4 tu for the respective next one after et cetera. If the group affiliation of a session is changed due to the result of a mid-session re-classification the obsolete group affiliation is discarded in favour of the new classification result. As a consequence the QoS Manager instructs the UAG to reset the timer for the re-evaluation interval.

4.5.2 Continuous collection of QoS information

As mentioned before, for the proposed framework, the NGN transport network appears transparent. Hence, information regarding the QoS encountered by media sessions can only be tapped at the respective receiving user terminals. On one hand, this situation simplifies the process of comprehensive QoS estimation, because for every media session a reference point is predetermined from which reliable QoS-related information can be obtained (namely the UAG to which the media-receiving user terminal is connected). However, if every UAG involved in a session transmitted information on the encountered QoS to a centralised collection point, this would drastically increase the per-session bandwidth required in any part of the transport network. In order to diminish the impact of this issue and, at the same time, however, afford the awareness of the QoS experienced by every media session, QoS profiling as introduced in section 4.2 is applied.

As mentioned before, given that a group of media sessions exist, of which it is known that all member sessions are affected by similar network performance conditions and hence, encounter similar QoS. In this case, obtaining information on the network performance encountered by a media session allows for the estimation of the QoS as experienced by any member session of the respective group.

Figure 4.9 and Figure 4.10 show the principle action of request and transfer of network performance characteristics required by the central QoS optimisation entity as a basis for the derivation of the QoS encountered by any member session of a respective group.

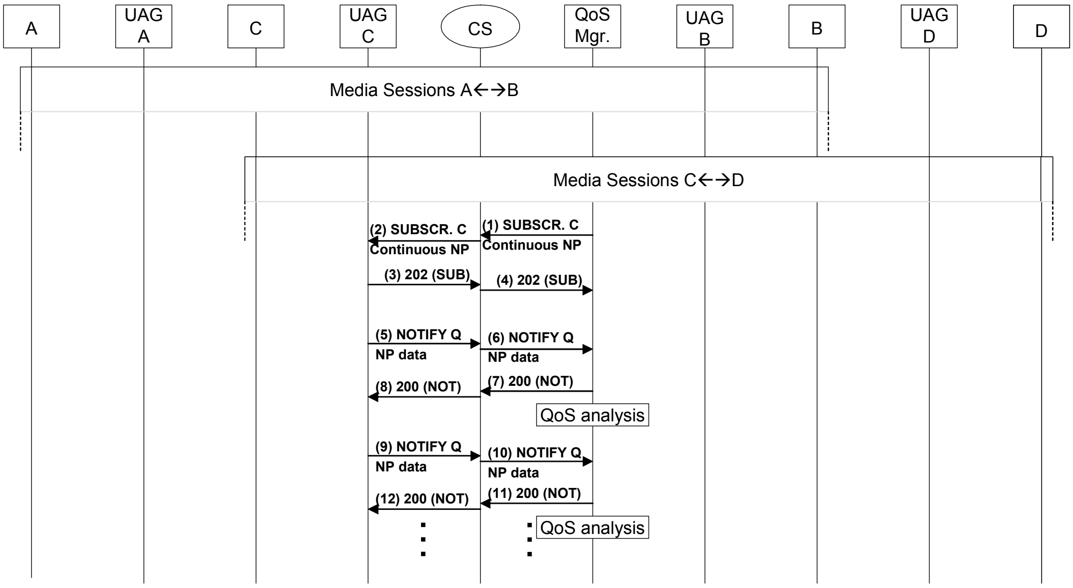


Figure 4.9: Request and transfer of reference network performance characteristics, part 1

It is assumed that SIP sessions have already been set up between subscribers A and B and between subscribers C and D, respectively. In line with the CAC process (see section 4.4) all media sessions (A→B, B→A, C→D, and D→C) were classified regarding their group affiliations by the use of QoS profiling. Within this classification process, media sessions B→A and D→C have been identified as members of the same group (note that, in this case, with a certain probability, media sessions A→B and C→D both are likewise members of another group. However, this is not necessarily the case.).

Be it that, for any reason, the Reference Session Selection entity of the central QoS Manager has chosen media session D→C as a representative session of its respective group. In this case the NP Server of the QoS Manager subscribes to the network performance encountered by the respective media session (see Figure 4.9, messages (1), (2)). Note that, in contrast to the network performance subscription in line with the CAC process (see section 4.4), the subscription introduced herewith refers to the continuous transmission of the whole set of network performance data throughout the remaining life time of the respective session, which is identified by the respective key term *NPReference* within the SIP event header field within the SUBSCRIBE request.

Identifying the specific event, the SUBSCRIBE message is intercepted by the UAG (see section 4.4 for details regarding the interception of SIP messages in UAGs). The UAG NP Agent answers the request (3), (4) and the UAG NP monitoring entity starts to continuously monitor the network performance experienced by the respective media session data stream. The collected information is send to the QoS Manager within the message body of SIP NOTIFY requests (5), (6) and (9), (10), which are acknowledged by the QoS Manager (7), (8) and (11), (12). Upon receipt of a NOTIFY request including network performance characteristics, the NP Server of the QoS Manager passes the data internally to the NP Analysis entity which derives a QoS characteristic for the respective time frame. This QoS characteristic is assumed to be shared among all media sessions being a member of the respective group, given that all sessions are based on the same codec (as all members of the respective group are known to be subject to the same network performance). Hence, by tracing the network performance in one UAG and transmit it to the QoS Manager, the QoS of all member sessions of the respective group is monitored.

If a session is terminated (see Figure 4.10, (13)…(16)) which has served as a reference session, the QoS Manager is informed through a respective session data base entry updated by the Call Server (this step is not included in Figure 4.10). In this case, however, a new reference session (such as session B→A) is chosen by the Reference Session Selection entity of the QoS Manager, whose NP Server in turn subscribes to the network performance of the session replacing the terminated session as group reference (18)…(20).

From this moment the corresponding UAG collects and transmits the network performance characteristics to the QoS Manager (21)…(28), which performs QoS analysis and so forth.

Note that, for the sake of continuity and completeness of the data representing the network performance (and hence, QoS) characteristic, at least two coexistent sessions of the same group should be monitored.

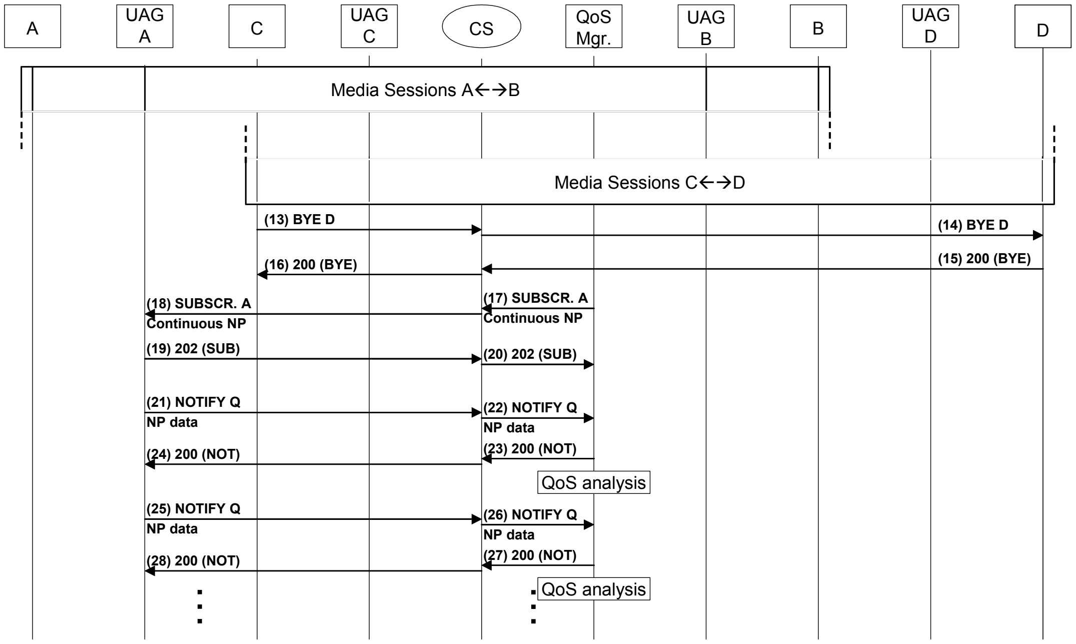


Figure 4.10: Request and transfer of reference network performance characteristics, part 2

4.6 Optimisation of network QoS conditions

As described in section 4.5, the framework for comprehensive QoS optimisation provides methods and mechanisms for the continuous evaluation of QoS conditions experienced by any active media session existing in a SIP-based NGN. The function of QoS profiling (the classification of sessions into groups by the similarity of encountered network performance) plays a major role in this context, as it allows for the minimisation of traffic volume required for QoS monitoring.

Given the case that the process of network performance analysis shows that the reference session of a particular group encounters insufficient network performance, then it is assumed that any member session of this group is concerned by similar conditions, which directly affect the QoS experienced by the respective subscribers.

According to (Zheng *et al.*, 2001), (Calyam *et al.*, 2004), and (Toral *et al.*, 2008), the performance of an IP network path regarding real-time multimedia stream transport capabilities mainly depends on its respective utilisation. Thus, if a particular network path is affected by insufficient network performance, it can be assumed that the actual traffic volume on this path has exceeded the feasible limit, resulting in congestion effects and, hence in forwarding conditions inadequate for real-time media streams. As the overall traffic on a path consists of several media sessions, reducing the traffic volume of at least some of the media sessions is a functional method to restore adequate conditions on the respective path.

The bandwidth required for the transport of a media stream directly depends on the selected media encoding algorithm (codec). For real-time media such as voice and video, several data-compressing codecs exist whose appliance comes with an acceptable trade-off between media quality and traffic volume. This fact is capitalised by the framework introduced within this thesis. In case of deteriorating network performance on a particular network path, selected media sessions on this path are downgraded, which means that their permitted maximum media bitrate becomes restricted, and hence codecs have to be changed to low bitrate codecs. This results in optimised traffic conditions on the path. Also the cancellation of selected media sessions is considered, if inevitable for the effective recovery of network performance. Note that those sessions are selected carefully, as described in detail in section 7.3.2)

The following sections describe the procedures required for downgrading (section 4.6.1) or cancellation (section 4.6.2) of sessions for the purpose of network performance optimisation.

4.6.1 Downgrading of media sessions

Figure 4.11 shows how the downgrading of a media session is induced. It is assumed that the QoS Estimation entity (being a component of the QoS Manager) has identified insufficient performance on a network path represented by a particular virtual group. By the use of the Session Significance Classification unit, session A→B, which is a member session of the concerned group, has been selected by the QoS Maintenance entity (being another component of the QoS Manager) to be downgraded in order to reduce the traffic volume on the respective network path. Note that any arbitrary algorithm or criteria, such as the objective significance of a media session or of the involved subscribers, respectively, might have been applied to perform this selection.

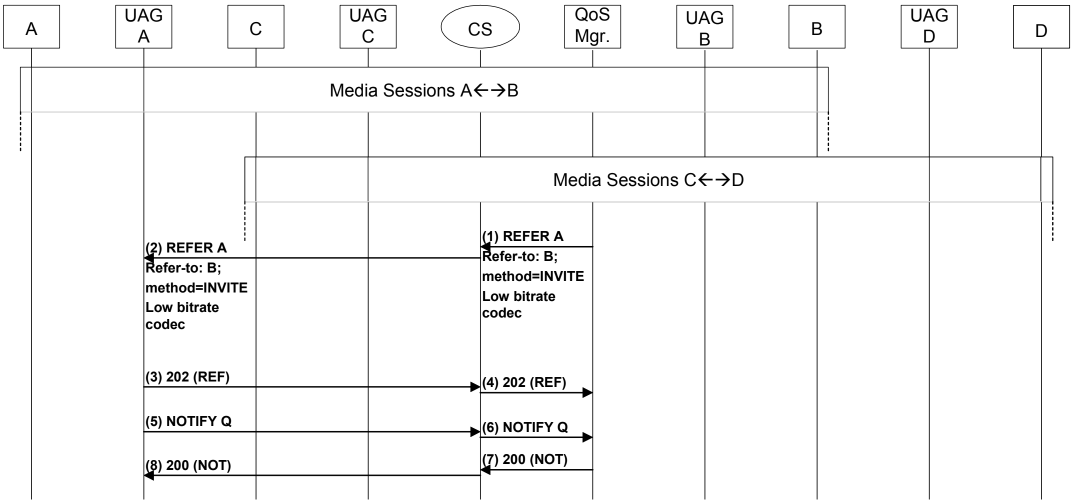


Figure 4.11: Initiation of a session downgrading

In this case, the QoS Maintenance Server (component of the SIP Application Server of the QoS Manager) sends a SIP REFER request (1) to assign the UAG of the media stream sender (in this example: UAG A) to downgrade the session to a low bitrate codec. UAG A intercepts the REFER request (2) and passes it to the internal QoS

Maintenance Agent. The agent accepts the request (3), (4) and NOTIFYs the QoS Manager that the downgrading action is going to be performed (5), (6), which, in turn, is acknowledged by the QoS Manager (7), (8).

In order to perform the downgrade, UAG A now activates its included B2BUA (Back-to-Back User Agent) to interconnect itself logically as a SIP dialog-serving entity. As shown in Figure 4.12, UAG A initiates the downgrading action by sending a SIP INVITE request (9) to be received by user terminal B (10). Note that this request refers to the existing SIP session between user terminals A and B by using the respective SIP dialog identifiers. Hence, this INVITE request is considered as a re-INVITE which does not initiate a new SIP session, but results in another SIP three way handshake exchanged between the parties of an already existing SIP session. In line with this three way handshake, SDP offer and answer are typically exchanged, which can be used to renegotiate parameters related to the media sessions such as media and codec choice. This procedure, known as "SIP session modification", is fully compliant with the SIP standards according to (IETF RFC 3261, 2002).

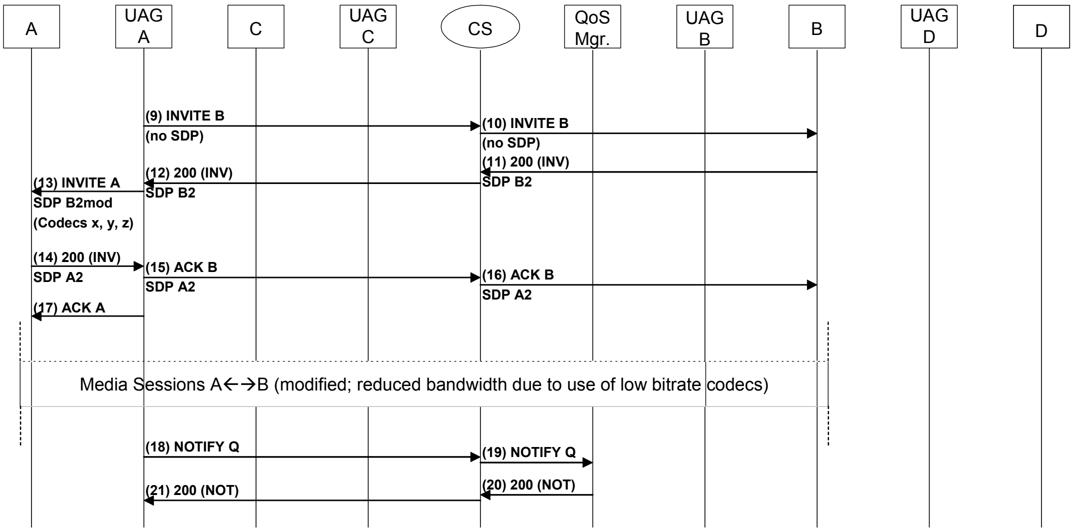


Figure 4.12: Session downgrading by SIP-based codec renegotiation

Also note that the INVITE request (9), (10) does not contain an SDP offer. This is compliant to (IETF RFC 3264, 2002), resulting in the generation of an SDP offer at the receiver of the INVITE request (user terminal B in this example). The SDP offer (SDP B2) is transmitted to UAG A within the SIP 200 response (11), (12) acknowledging the INVITE request.

Before generating the ACK request to complete the SIP three way handshake with user terminal B, UAG A sends a re-INVITE request (13) to user terminal A to keep it involved in the codec renegotiation process. Within this INVITE request, UAG A passes the SDP offer of user terminal B to user terminal A, after manipulating the variety of codecs (SDP B2mod) so that only low bitrate codecs are offered to user terminal A (as ordered by the QoS Manager within the REFER request (1) shown in Figure 4.11). User terminal A accepts the re-INVITE with a SIP 200 response (14), passing the SDP answer of User Agent A (SDP A2) to UAG A. This SDP answer is transmitted to user terminal B within the outstanding ACK request (15), (16), and another ACK request (17) is sent to user terminal A to complete the SIP three way handshake. Note that this interrelation of SIP three way handshakes and SDP offer/answer exchanges is fully compliant with (IETF RFC 3725, 2004).

Subsequently, on behalf of the QoS Manager, the media sessions between A and B have been modified regarding the codecs used. Low bitrate codecs have been chosen in order to optimise the utilisation of the respective network path. The UAG informs the QoS Manager of the successful session modification by sending a SIP NOTIFY request (18), (19), which is answered with a 200 response (20), (21).

4.6.2 Session cancellation

As described in section 4.6.1, the QoS Manager is able to decide and initiate the downgrading of media sessions to reduce the bandwidth utilisation on congested network paths. However, the QoS Manager also can decide to completely cancel single sessions in order to spend less effort and, at the same time, achieving a stronger effect of bandwidth release.

Figure 4.13 shows the proposed framework action for the cancellation of sessions driven by the QoS Manager. Like in the downgrading scenario described in section 4.6.1, a SIP REFER request is sent to UAG A (1), (2). In the case of a cancellation to be performed, the UAG is instructed to terminate the respective session (*method=BYE* argument passed within the Refer-To SIP header field). The REFER is accepted by the UAG (3), (4), and a NOTIFY request is sent to the QoS Manager (5), (6), which acknowledges upon receipt with a 200 response (7), (8). UAG A terminates the existing SIP session with both involved parties (user terminals A and B) by performing third party call control, acting as a SIP Back-to-Back User Agent (B2BUA). Therefore, SIP BYE requests are sent to both user terminals B (9), (10) and A (11), terminating the SIP session. The user terminals immediately stop exchanging media, and the SIP requests are answered with SIP 200 responses (12), (13), (14). UAG A sends a SIP NOTIFY request (15), (16) to inform the QoS Manager that the session termination has been accomplished successfully. Finally the server responds to the notification with a 200 response (17), (18).

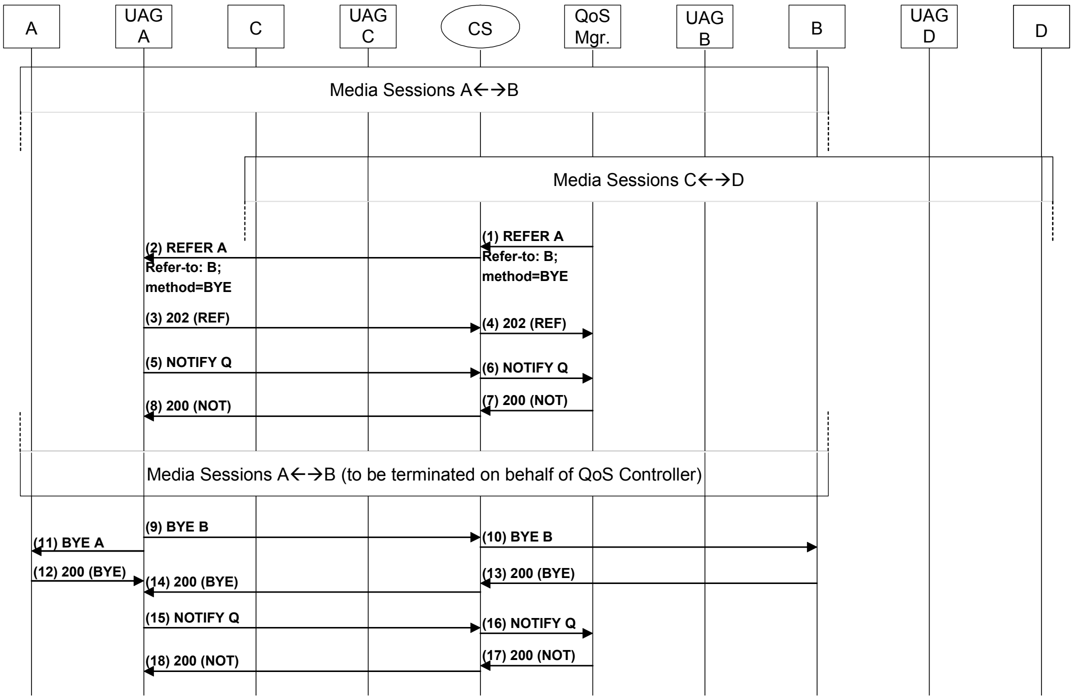


Figure 4.13: Session cancellation on behalf of the central QoS Manager

4.7 Summary

Within this chapter, a framework for comprehensive QoS optimisation in SIP-based NGN has been introduced, which fulfils the requirements for QoS provision in SIP-based NGN as declared in chapter 3. Crucial side conditions have been defined, and from these, three tasks have been derived, describing the main functionality which have to be fulfilled by the solution framework.

- Task 1: QoS estimation
- Task 2: Integration with Call Admission Control
- Task 3: QoS optimisation

The two main types of framework components, namely so-called User Access Gates and a QoS Manager, have been introduced. Both types of components consist of a multitude of integrated sub-elements, whose respective functionalities are demonstrated inline with the description of the overall framework regarding the fulfillment of the three above-mentioned tasks. As indicated throught the description of the framework, the concept that underpins its functionality is the principle of QoS profiling. Chapter 5 outlines this concept, based on AI analysis of the session characteristics.

5 QoS profiling – grouping media sessions by encountered network performance

As mentioned in section 4.2 coexisting media streams experience similar QoS conditions if they share those portions of a transport path that have a significant impact on the overall network performance encountered by those streams. Utilising this effect in a reverse manner means that, from matching network performance characteristics of media streams, conclusions can be drawn on the similarity of transport conditions encountered. Hence media streams can be logically grouped by the similarity of the encountered network performance. This methodology, which has been investigated and utilised within this research, has been termed *QoS profiling*, and has been considered valuable regarding both, efficient QoS estimation and QoS maintenance. As an example, measuring QoS-relevant network performance parameters (delay, jitter, and packet loss rate) of one media stream allows for the estimation of the QoS experienced by any media stream which has been assigned to the same logical group by the use of QoS profiling.

This chapter begins by introducing the principle of QoS profiling (section 5.1). In a further step, the application of an ART 2 Artificial Neural Network for the purpose of QoS profiling is described (section 5.2). In order to utilise ART 2 Neural Networks for unsupervised QoS profiling, both an accuracy self-scaling mechanism and a bootstrap mechanism were developed, which are presented in sections 5.3 and 5.4, respectively.

5.1 The principle of QoS profiling

Data flows that have in common the same packet size and packet intervals are similarly affected when concurrently sent over those parts of the transport network that contribute relevant network performance influences.. This was verified by practical research published as a conference paper (Weber *et al.*, 2009a), and was substantiated by several simulation-based research analyses also published as conference papers (Weber *et al.*, 2009b), (Weber *et al.*, 2010). In line with these analyses it was found that the similarity of the network performance affecting several different multimedia over IP packet flows is reflected in their respective network performance parameter evolution. In particular, matching jitter value sequences of several media streams allows for the identification of media packet streams exposed to similar network performance conditions, which already could be observed throughout measurement series published in (Abu Salah *et al.*, 2008).

In the proposed QoS profiling concept, all media sessions identified as being exposed to similar network conditions are logically assigned to the same group. From each group at least one session is selected as a group reference at any time. The group reference selection process is further described in section 6.1. A reference session represents the network performance that affects any media session affiliated with the respective group. Hence, from the network performance encountered by the reference session, QoS conditions can be estimated for every member session of the respective group (note that the QoS estimation of media sessions from encountered network performance is further described in section 6.2). At the same time, this session is used as group representative regarding the QoS profiling process.

5.1.1 Assigning media sessions to virtual groups

Applying QoS profiling means that, whenever a new media session is established, it has to be assigned to a logical group which represents its respective network performance characteristic. Therefore, the jitter variation and characteristics of the candidate session is matched with the jitter evolution of those media sessions being selected as representatives of already existing groups. This is performed within the QoS Manager, a centralised component of the solution framework residing at the service provider. See section 4.3 for further details.

Note that all jitter characteristics have to be synchronised by the use of time stamps, as they represent an ever-changing value progression. Figure 5.1 shows synchronised jitter characteristics of five different media sessions. In this example, jitter characteristics are available from four reference sessions, each representing a different group (grey-coloured characteristics Ref. 1 … Ref. 4), while the fifth characteristic was extracted from a media session which has to be assigned to one of the groups represented by the reference characteristics. By comparing the jitter characteristics given in Figure 5.1 and considering the full time interval shown, it can be determined that the session to be assigned (bottom line, black) is most likely a member of the group represented by Ref. 3 (third grey line from the top).

The jitter value sequences to be matched are captured by the respective functionality of UAGs (see section 4.3), be it integrated with the respective user terminal, or looped in between the access network and the user terminal. Therefore, if an session initiation request is received by the SIP service infrastructure of the provider, the QoS Manager queries both involved UAGs for the determination and submission of a network performance sample (see section 4.4). These samples include the jitter value sequences used within the initial media session grouping process.

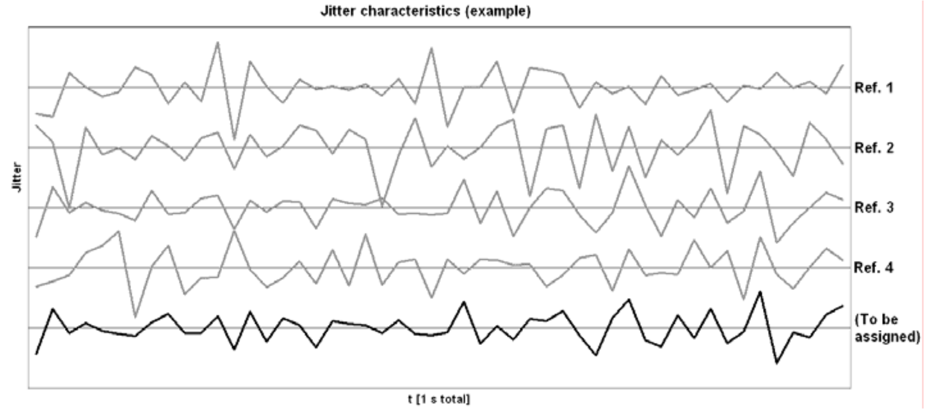


Figure 5.1: Jitter characteristics, monitored in NGN user terminals

UAGs involved in media sessions that have been selected as group reference sessions have to capture all network performance parameters and transmit the respective characteristics block-wise within the message bodies of SIP messages to the centralised QoS Manager (see section 4.5). Hence the QoS Manager is aware of the network performance characteristics of any logical session group including their jitter characteristics, which in turn is also used for the matching process of the new media session.

The variation of jitter values of both, the reference sessions and the candidate session are matched. If the jitter value evolution of the candidate session exhibit significant similarity with one of the group reference sessions, it is assumed that both jitter characteristics are mainly influenced by similar QoS characteristics, and hence the candidate session is assigned to the respective group. If sufficient similarity could not be detected among the jitter characteristics of the candidate session and any reference sessions, the candidate session is designated as the first member session of a new group to be established. In this case, as the first group member, this session is automatically selected as reference session of the respective group. The procedure for the selection of reference session is further described in section 6.1).

5.2 Application of an ART 2 Neural Network for QoS profiling

As stated in section 5.1, in NGN QoS profiling, media sessions are classified into groups by the similarity of jitter characteristics encountered. That is, from a given set of jitter value sequences (each representing an individual media session) to be classified, criteria must be identified by which those sequences are best distinguished. In a next step, considering the identified distinctive features, the given jitter sequences are analysed regarding similarities and differences. Finally, the sequences have to be classified to groups, while the final number of groups is unknown before the classification process is completed. The classification must be performed so that all sequences featuring a defined level of similarity are assigned to the same group, although they may show individual properties and noticeable differences among each other. At the same time, the members of any group must be clearly distinguishable from any member of any other group.

The procedure described above clearly matches the definition of unsupervised pattern recognition as given by (Theodoridis and Koutroumbas, 2009), (Shih, 2010), (Sá, 2001), and (Ripley, 2005).

5.2.1 Pattern recognition

In general, pattern recognition (PR), being a part of most decision making “intelligent” systems, can be defined as the scientific discipline to classify objects into a number of classes (Theodoridis and Koutroumbas, 2009). In this sense, a pattern is “any distinghuishable interrelation of data” (Shih, 2010), a representation of an object suitable for the intended type of processing (Sá, 2001). Note that, in the context of pattern recognition, the terms *pattern* and *object* are often used interchangeably.

Pattern recognition methods and systems are applied in a variety of both, research and practical disciplines. Typical examples are

- Character (letter or number) recognition, OCR systems (Optical Character Recognition),
- Speech recognition,
- Identification of finger prints,
- Seismic analysis,
- Computer-aided (medical) diagnostics,
- Data mining / knowledge discovery.

Every pattern recognition process involves a phase of learning, in which the knowledge on how to distinguish patterns of a given set is determined by the pattern recognition system. Depending on the availability of training samples and / or a priori requirements regarding the total number and the properties of groups, a learning phase is either said to be supervised, unsupervised or semi-supervised. This differentiation of learning modes is also adopted to specify various types of pattern recognition processes according to amongst others (Theodoridis and Koutroumbas, 2009), (Shih, 2010), (Sá, 2001), and (Ripley, 2005).

A system designed for **supervised pattern recognition** is typically trained prior to its in situ deployment. Therefore it is assumed that sample patterns (so-called labelled patterns) are available whose assumed class associations are known to the PR system in order to determine distinctive features by which patterns can be classified. The number of output classes may be predetermined, and also a reference pattern per class, defining the 'perfect class member'. A typical example for the application of supervised pattern recognition is an OCR system which is used to identify hand-written characters from bank order forms. Only a limited number of output classes (one for each allowed character, such as letters from A to Z and numbers from 0 to 9) exist, each of which is identified by a reference sample, defined by the used character set. Also these systems are typically trained with sample patterns before real life application to allow for best possible performance.

In contrast, in **unsupervised pattern recognition**, typically no pre-knowledge exists to support the learning and classification process. While in operation, a set of unweighted (unlabelled) patterns is provided to the system, which has to identify the features forming a class in line with the classification process. This process of unsupervised pattern recognition is also called clustering. It may be compared to a sorting task in which unknown objects (such as random-shaped figures) have to be clustered by similarity without any further given rule (such as 'cluster by number of angles' or 'cluster by colour').

Semi-supervised pattern recognition refers to a learning process including both, labelled and unlabelled training patterns. In this case, the labelled data can be used to express rules or constraints to be considered by a (unsupervised) clustering process.

Within the following, the focus is mainly set on unsupervised pattern recognition, being considered as a fundamental basis of QoS profiling within the framework for comprehensive QoS optimisation.

5.2.2 Adaptive Resonance Theory 2 (ART 2)

As stated in section 5.2, the basic process of QoS profiling can be described as unsupervised pattern recognition, also referred to as clustering.

Beside the three most common clustering algorithms (namely k-means, hierarchical clustering and Self-Organising Maps (SOM) (Boutros and Okey, 2005)), a multiplicity of further mechanisms have emerged to solve the task of unsupervised pattern recognition. The majority of these approaches are based on Artificial Neural Networks (ANN) providing Competitive Learning, the most common of which are SOM, Vector Quantization, and ANN based on Adaptive Resonance Theory (ART). In contrast to concurrent approaches, ART Neural Networks have explicitly been designed to solve a number of common issues of ANN, such as the stability-plasticity-dilemma (Carpenter and Grossberg, 1987); (Henning, 2002); (Xu and Wunsch, 2008). Furthermore, ART combines methods for both, adaptive filtering and competition (Shih, 2010) and builds a bridge among those pattern recognition approaches utilising ANN and those approaches which are based on ‘hard’ clustering algorithmic schemes such as k-means (Ripley, 2005). (Rani and Renganathan, 2003) found that, compared to SOM, ART 2 Neural Networks require less memory storage and computation time. A further potential advantage of ART is that, in contrast to other pattern recognition mechanisms, the clustering is critically re-analysed by the algorithm itself (Freriks *et al.*, 1992). Therefore, ART has been chosen to be utilised for unsupervised pattern recognition embedded within the QoS profiling approach introduced within this work.

Artificial Neural Networks (ANNs) are used in numerous technical applications in order to perform complex tasks such as pattern recognition (or pattern classification), function approximation, prediction/forecasting, optimisation, content-addressable memorising, cybernetics, as well as clustering/categorisation. The latter application is denoted as unsupervised pattern classification in (Jain et al., 1996).

Adaptive Resonance Theory aims at the question how ANN can be designed in a way that allows for preserving the knowledge already learned about a number of samples (stability) while, at the same time, being able to classify and learn information about further presented samples (plasticity). This is realised by a two-step mechanism, which enforces a backward similarity verification before allowing for the updating (and hence, potential adulterating) of already learned information.

While the original ART-1 ANN model (Grossberg, 1976a), (Grossberg, 1976b) aimed at the classification of binary value samples only, the extended ART 2 model (Carpenter and Grossberg, 1987) provides the ability to process patterns consisting of continuous values. Several research work involving ART 2 Neural Networks have been performed, such as regarding the recognition of anomalous IP traffic in intrusion detection (Ding *et al.*, 2008), the analysis of ultra-sonic echoes (Solís *et al.*, 2008), the determination of flow velocities of liquids (Jambunathan *et al.*, 1997), fault diagnostics in chemical engineering (Aradhye *et al.*, 2004), and human signature verification (Mautner *et al.*, 2002).

ART 2 neural networks can be described as unsupervised-learning neural networks with the ability to match analogue continuous value sequences with the objective to classify the sequences by their similarities (Carpenter and Grossberg, 1987). An input sequence, also referred to as a pattern, is interpreted as an m-dimensional vector, where m is the number of single values comprised by the respective input pattern. If an arbitrary number of m-dimensional patterns are presented to an ART 2 network, after a predefined number of learning cycles, the neural network tries to map each pattern to one of n output classes by accomplishing a multi-step comparison process for each pattern. Patterns showing typical similarities are assigned to the same output class.

Figure 5.2 shows the principle architecture of an ART 2 Neural Network. It consists of two subsystems, namely the Attentional Subsystem and the Orienting Subsystem.

Within the Attentional Subsystem, a candidate value pattern I is loaded into a so-called Short Term Memory (STM) which is represented by a collection of functions abstractly called F1 layer. Within this layer, the candidate value pattern is preprocessed, which includes several steps of normalisation, contrast amplification, and noice reduction. Subsequently, individual recognition features of the pattern are identified, extracted and memorised. In a next step the candidate pattern is matched with already existing classes, represented by value samples kept in a Long Term Memories (LTM). The LTM is represented by the individual weightings of bidirectional links between the F1 and F2 layers. The latter layer contains a collection of *n* output classes, to one of which an input candidate value pattern should finally become assigned.

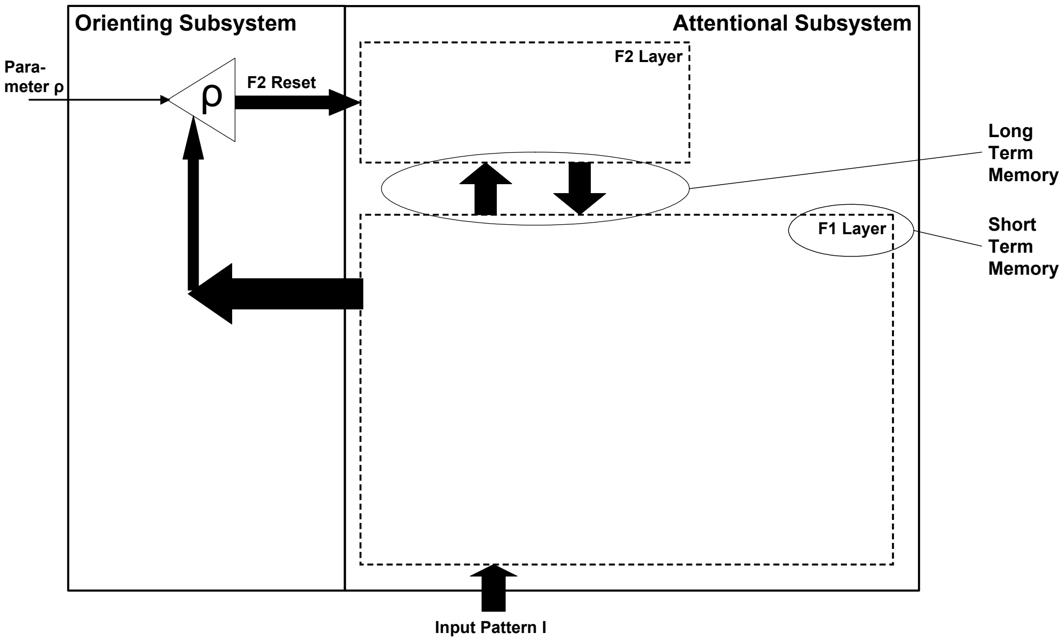


Figure 5.2: Principle architecture of an ART 2 Neural Network

In a first classification attempt, a best fitting ("winning") output class is determined. Now a function of the Orienting Subsystem becomes involved to assess the accuracy of the matching. This is performed by evaluating whether the candidate pattern will be sufficiently represented by the respective class, once it has been designated as its member. If this is the case, the candidate pattern is set as a member of this class, and the features of the pattern are introduced ("learned") to the LTM copy representing this class. If – in contrast – the chosen class is found not to represent the candidate pattern sufficiently, the Orienting Subsystem triggers a reset, resulting in a new classification attempt within the Attentional Subsystem, after excluding ("blocking") the previous winner class from the classification. Note that, regarding the adequacy decision, the Orienting Subsystem directly refers to an external parameter *ρ* ("vigilance parameter"), which defines the deviation tolerance of the classification process.

Note also that the value of ρ has to be chosen manually by the operator of the ART 2 ANN. According to (Hermann, 1992), reasonable limits for ρ are defined by

$$\frac{1}{2}\sqrt{2} \leq \rho \leq 1 \tag{5.1}$$

Both the architecture and internal processes of ART 2 ANN are formally described in (Carpenter and Grossberg, 1987) in detail.

5.2.3 Selection and Verification of ART 2 ANN for QoS profiling

Generally, the utilisation of Artificial Intelligence for the classification of media sessions perceiving similar QoS conditions was first discussed in (Weber *et al.*, 2008). Based on this proposal, (Schnaidt, 2009) implemented and tested the application of an ART 2 ANN for the classification of RTP media sessions by their jitter characteristics. As a conclusion it was found that, although the chosen experimental set-up was considered insufficient in retrospect, ART 2 ANN are generally suitable for the purpose of media session classification.

Following the results achieved by (Schnaidt, 2009), the applicability of ART 2 ANN for the purpose of QoS profiling was further verified. Therefore, an experimental testbed was set up within the Laboratory for Telecommunication Networks at University of Applied Sciences Frankfurt, Germany. Figure 5.3 shows the testbed layout.

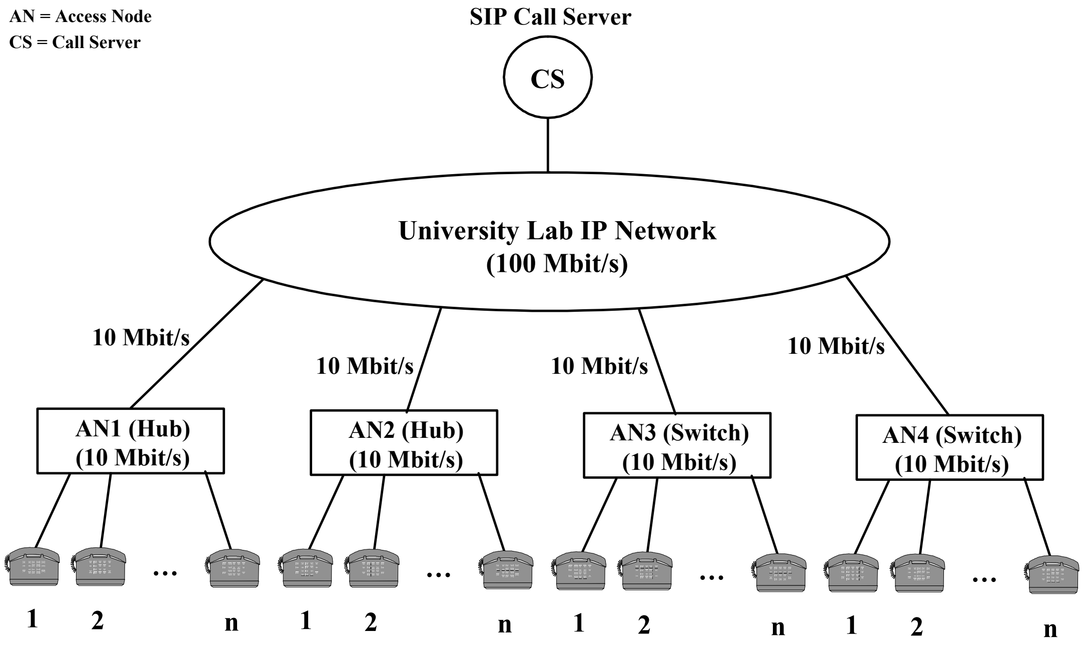


Figure 5.3: Laboratory testbed layout

Within this test bed, different VoIP communication scenarios (VoIP streams sent between defined terminals connected to defined access nodes) were arranged. Different simultaneously existing communication situations were considered in all communication scenarios. Each communication situation comprised different numbers of concurrent VoIP streams sent among the considered ANs. Table 5.1 shows the assignment of simultaneously existing communication situations arranged per communication scenario, and, within the correlation fields, the numbers of VoIP streams considered per communication situation. Within Table 5.1, for communication situations comprising streams exchanged between two different access nodes, the numbers of streams for each communication direction is provided, separated by a slash.

Table 5.1: Communication situations concurrently arranged per communication scenario

	Communication situations considered per communication scenario						
Communi-cation scenario	Internal streams AN1	Streams AN1↔AN2	Streams AN1↔AN3	Streams AN2↔AN3	Streams AN2↔AN4	Internal streams AN3	Internal streams AN4
I			5/5		7/7		
II	6			5/5			8
III	3	4/4		3/3		3	8

To obtain the QoS characteristics of the VoIP data streams, during the tests, all streams were simultaneously captured IP packet-wise at their respective receiving user terminals.

In a further step, the capture files recorded by the VoIP phones were analysed subsequently for each communication scenario. From all captures, the packet-by-packet jitter (delay variance) characteristics of the respective VoIP stream were extracted by calculating the variation of the inter-arrival time for each pair of consecutive IP packets. Hence, for each IP packet of all concurrent VoIP streams, an individual jitter value was achieved, so that jitter value sequences could be composed. Subsequently, to achieve comparability, all jitter sequences obtained from VoIP streams associated with a respective communication scenario were synchronised in time by the use of the time stamps that were included automatically throughout the traces.

In order to generate value sequences that could be processed by an Artificial Neural Network in real-time, the sequences were cut to a length representing one second of the associated VoIP stream (jitter values obtained from 50 subsequent IP packets with a payload sequence of 20 ms each). In a further step, all jitter sequences were smoothed by a running mean algorithm, taking into account five consecutive jitter values.

From the running mean algorithm, for each VoIP stream, one smoothed jitter sequence was obtained, each consisting of 46 discrete analog values. For the comparison and classification of the jitter sequence patterns associated with a respective communication scenario, an ART 2 neural network was implemented by the use of JavaNNS, a Java-based version of SNNS (University of Tübingen, 2008). The neural network was provided with 46 input units, so that every jitter pattern could be presented to and processed by the neural network at once in full length.

To compensate potential inaccuracies within the classification process, the neural network was provided with ten output units, each of which could represent a specific jitter characteristic. The expected numbers of output classes to be assigned per communication scenario can be read from Table 5.1. Within the tests performed, it was assumed that, provided 100 percent classification accuracy, jitter characteristics resulting from the same communication situation could be mapped to one class (for VoIP streams sent among user terminals connected to the same AN) or two classes (one for each communication direction for communication involving two different ANs), respectively.

For each communication scenario, a multitude of classification runs were performed. Therefore, the jitter patterns obtained from the analysis of the VoIP streams associated with the respective communication scenario were presented to the ART 2 neural network as a set of pattern sequences. The ART 2 vigilance parameter ρ was manually varied among the runs in order to evaluate the most exact classification result per pattern set. Furthermore, different sequential orders of the jitter patterns within a respective pattern set were tested. The number of learning cycles to be performed by the ART 2 ANN before the actual classification was set to 100 for all runs.

Further details on the performed tests and the achieved results were published as a research paper (Weber *et al.*, 2009a). As a summarising result, Table 5.2 shows the achieved classification accurateness (in percent) of ART 2-based classification of media sessions by the similarity of perceived QoS. The term accurateness refers to the correctness of the classification of those media streams to the same group that are exchanged between the same pair of access networks.

Table 5.2: Classification accurateness achieved from ART 2 verification scenarios

Communication scenario	I	II	III	Total Average (I, II, III)
Classification accurateness [%]	87.5	62.5	57.1	69.0

Considering the details of the respective communication scenarios given in Table 5.1, it is obvious that the precision decreases with the increase of the complexity of the communication scenarios. However, it is also obvious that even without any refinement of the classification methodology or applied algorithms, a classification accuracy of 87.5 percent could be achieved for communication scenario I, which was considered to be sufficient as a basis for the application of ART 2 ANN for QoS profiling.

Note that throughout this research, further enhancement and refinement of the methodology of ART 2-based QoS profiling could be achieved, resulting in a general improvement of the classification reliability and hence, an increase of the overall classification accurateness as demonstrated in section 8.3.2. Sections 5.3 and 5.4 introduce the respective enhancements.

5.2.4 Principle of ART 2-based QoS profiling

For the automated matching and grouping of jitter characteristics within this research, an ART 2 ANN (see section 5.2.2) has been selected, because of its suitability in classifying analogue value sequences. The ART 2 ANN is provided with *m* input cells and *n* output cells, with *m* being the number of discrete values to be considered within the ART 2 classification run. Hence, the ART 2 has to be provided with 50 input cells if a number of 50 consecutive jitter values is chosen to be considered. This corresponds to a VoIP media stream segment of a length of one second, assuming G.711 coding and 160 Byte payload per packet. Using jitter value characteristics which represent one second of a media stream segment provides a reasonable compromise between classification accuracy on one hand and the delay resulting from capturing, processing and classifying of the jitter characteristics on the other hand.

The number of output cells *n* refers to the number of distinguishable virtual groups to be considered within the ART 2 classification run. Thus, if four reference jitter characteristics are available, each representing the QoS conditions of one existing virtual group, the ART 2 ANN has to be provided with four output cells.

As described in section 5.2.2, a sequence of consecutive values to be classified within an ART2 classification run is referred to as a pattern. Hence, a set of value sequences, each representing one group, is referred to as a pattern set. Within the following, a jitter value sequence is named a pattern.

The collection of jitter patterns to be considered (one reference pattern per existing virtual group plus the pattern to be classified, referred to as 'Pattern Under Test' (PUT)) are provided to the ART 2 ANN as one pattern set. Within one classification run, the ANN performs a number of internal matching and weighting tasks, involving an unsupervised learning process, resulting in a stable state of classification. When this state is reached, each pattern has been either assigned to exactly one class, or it could not be assigned to any class. Note that several patterns within one pattern set could also be assigned to the same class if they showed significant similarities.

Note that the result of an ART 2 classification run is strongly influenced by the choice of the ART 2 vigilance parameter ρ (see section 5.2.2), providing a value range of ρ = $[(0.5 \cdot \sqrt{2})...1.0]$. If ρ is chosen relatively low, the ANN shows a more tolerant classification behaviour. Hence, the classification is potentially imprecise. If ρ is chosen relatively high, the ANN shows a strict classification behaviour, coming along with the risk of being unable to identify similarities among patterns within the respective pattern set (see (Carpenter and Grossberg, 1987) for details on the function of ρ). Thus, the result of an ART 2 classification can only be considered reliable if a most adequate value for ρ is determined. In section 5.3 a method for the automated discovery of a suitable ρ value is discussed.

5.3 Accuracy self-scaling for ART 2-based QoS profiling

The most suitable value for the ART 2 vigilance parameter ρ mainly depends on the texture and characteristics of the patterns to be classified, and hence, can generally not be pre-defined. In (Rayón Villela and Sossa Azuela, 2000) a method is introduced for the determination of ρ in ART 2 ANN. However, as this approach is optimised for the best possible discrimination of several patterns of a pattern set, this approach is not suitable for the process of distinguishing reference patterns and, within the same classification run, identifying the reference pattern which a respective candidate pattern matches best.

Because for our purpose it is understood that the reference patterns must be distinguishable from each other (as they are known to represent different QoS characteristics), we can use this fact for approximating the most suitable value of ρ. Therefore a number of classification runs are performed with the same pattern set, but with varied ρ (note that within the following, performing multiple classification runs with the same pattern set is referred to as one classification process). Table 5.3 shows an example for the influence of ρ on the result of ART 2 classification runs, and for how the most suitable value ρ can be evaluated.

Table 5.3: Example for the influence of ρ on ART 2 classification results

Classification Process No.	**Vigilance (ρ)**	**Pattern 1**	**Pattern 2**	**Pattern 3**	**Pattern 4**	**Pattern 5**
		(Ref. 1)	(Ref. 2)	(Ref. 3)	(Ref. 4)	(PUT)
1	0.85	1	1	1	2	1
2	**0.934375**	**1**	**2**	**3**	**4**	**3**
3	0.94375	1	2	3	4	0

For the classification runs considered, in run

- no. 1 (ρ = 0.85), all patterns except pattern 4 were assigned to the same class (class 1). Pattern 4 was assigned to a separate class (class 2). Hence, as the classification result is too imprecise to distinguish among the reference patterns (patterns 1…4), the considered value for ρ has obviously been chosen too low.
- no. 2 (ρ = 0.934375), all patterns representing references (patterns 1…4) were assigned to separate classes (classes 1…4), while the PUT (pattern 5) was assigned to class 3. As pattern 5 was assigned to the same class as pattern 3, it is evident that the user terminal represented by pattern 5 belongs to the same group as the user terminal serving as reference 3. However, ρ might still be too low, as the classification behaviour of the ART 2 would perhaps be more exact with a higher value of ρ.
- no. 3 (ρ = 0.94375), all patterns representing references (patterns 1…4) were assigned to separate classes (classes 1…4), while the PUT (pattern 5) could not be assigned to any of these classes (showing "0"). Hence, either the user terminal represented by pattern 5 shows a different jitter characteristic than the user terminals serving as references 1…4 (in this case, a new virtual group is established), or ρ was chosen too high to identify similarities between pattern 5 and another pattern.

Thus, for the exemplary classification process shown in Table 1, the best suitable value for ρ must be in the range of (0.934375…0.94375). In order to further pinpoint the best suitable value for ρ, further classifications have to be run with ρ values taken out of the given range. However, if these further classification runs do not disprove the provisional result evident from Table 1, the PUT (pattern 5) can be said to be classified into the group represented by reference 3. This is due to the result of classification run no. 2, in which the reference patterns are distinguished from each other and, at the same time, the PUT could be assigned to one of them.

In order to define a consistent procedure for the validation of ART 2 classification run results, the following cases of possible results are distinguished. Each case provides a statement regarding the valuation of the applied ρ value, aiming towards a most suitable *ρ*. This statement provides the basis for the choice of *ρ* to be set for the subsequent classification run. Following these statements, a most reliable assignment of the PUT can be provided, and a most suitable value for *ρ* can be determined.

- Case a) If at least one reference pattern could not be assigned to any class number, the classification process was performed too strict. Hence, ρ was chosen too high.
- Case b) If two or more reference patterns share the same class number, the classification process was performed too loose. Hence, ρ was chosen too low.
- Case c) If all reference groups could be distinguished correctly from each other and if a class number was assigned to the PUT, this class number is a potential group candidate. Hence, the classification process might have been performed well, but it also might have been performed too loose. Hence, ρ was potentially chosen too low.
- Case d) If all reference groups could be distinguished correctly from each other and if no class number was assigned to the PUT, this could have been caused by two different situations. In any case, ρ was potentially chosen too high.
    - The classification process might have been performed well. In this case, the PUT simply does not match any of the reference groups and, hence, is the first recognised representative of a new group. It is assumed that this is the case if this PUT could not be assigned to a reference pattern within another classification run performed.
    - The classification process might have been performed too strict to identify similarities between the PUT and any of the reference patterns.

In order to determine a most suitable value for ρ, multiple classification runs have to be performed based on the same pattern set. In each run, ρ has to be adapted subject to the outcome of the previous run (cases a) … d)). Regarding the adaption, the below-mentioned rules are suggested to be followed, resulting in a statistically optimised number of required classification reruns (i = number of classification runs performed with the same pattern set. See case explanations above for the meaning of “too low/high”).

$\rho_{(0)min} = 0.7; \rho_{(0)max} = 1.0$ $\rho_i = (\rho_{(i)max} + \rho_{(i)min}) / 2$

If $\rho_{(i-1)}$ = “*too low*”: $\rho_{(i)min} = \rho_{(i-1)}\,; \rho_{(i)max} = \rho_{(i-1)max}$

If $\rho_{(i-1)}$ = “*too high*”: $\rho_{(i)max} = \rho_{(i-1)}\,; \rho_{(i)min} = \rho_{(i-1)min}$

This classification process is continued in an automated manner until a most suitable value for ρ is determined (until the assignment of the PUT to either one specific class or to no class has been verified within a preselected number of classification runs, or until a most suitable value for ρ could be narrowed down to a predefined range of accuracy).

The results of an ART 2-based jitter pattern classification process can be further verified by repeating the process for jitter patterns originating from the same user terminals, but which represent a later time interval of the same respective communication situations. After having performed several classification processes, the provisional results can be compared and the most likely final result can be determined.

5.3.1 Test and conclusion

The introduced ART 2 accuracy self-scaling approach has been verified by the use of a simulation environment. Based on the research prototype introduced in chapter 8, a SIP-based NGN architecture has been set up, allowing for the simulation of different communication scenarios. The architecture consisted of one core network and four access networks, to each of which different numbers of user terminals were connected. Upon session initiation, media flow packets (simulating VoIP calls with G.711 codec) were bidirectional exchanged in a peer-to-peer manner between the user terminals. By varying the distribution of parallel calls among the user terminals connected to different access networks, different numbers of virtual groups of media sessions were set up, each being exposed to different QoS conditions. All media packets were recorded and time-stamped at their respective receiving user terminals. The collected data were post-processed to extract the per-packet inter-arrival jitter of each media flow. For each virtual user group, one user terminal was randomly chosen whose jitter characteristic served as the respective group's reference characteristic. Finally, the synchronised jitter sequences were bundled as pattern sets and provided to the ART2 ANN, and the classification procedure was applied as introduced within this section.

Test results

Table 5.4 shows the results of the initial proof-of-concept tests performed for the evaluation of AI-based QoS profiling including the accuracy self-scaling approach introduced within this section. These results were published in (Weber *et al.*, 2009b).

Table 5.4: Results of initial tests of the jitter classification

Scenario	No. of virtual groups considered	Accuracy of classification
a)	4	93%
b)	8	71%
c)	10	69%

As shown in Table 5.4, within the tests, three different communication scenarios (a) … c)) were considered, which differ in the number of considered virtual groups. It is observed that the classification accuracy decreases from scenario to scenario with the increase of the number of virtual groups. Although it is evident that the principle of AI-based QoS profiling, applied in conjunction with the self-scaling procedure introduced within this section, results in a suitable degree of accuracy.

The self-scaling approach could be further improved, especially regarding the consideration of a larger amount of virtual groups to be considered, as proposed in (Weber *et al.*, 2009b). It was decided to generally use pattern sets consisting of three patterns only, composed of the respective PUT and two reference patterns. Depending on the number of references to be considered within a respective scenario, multiple pattern sets were assembled, and presented consecutively to an ART 2 Neural Network, which was equipped with three output cells only. The self-scaling approach introduced within this section was applied on all pattern sets separately, and the reference pattern returning the highest value for the ART 2 vigilance parameter ρ was considered to be the best matching reference.

5.4 A bootstrap mechanism for ART 2-based QoS profiling

As stated in section 5.3, the mechanism for the virtual grouping of media sessions is based on the distinguishability of reference value patterns, each representing an autonomous virtual group. Within the following subsections, a bootstrap issue existing with the virtual grouping mechanism is described, and a solution is outlined.

5.4.1 Bootstrap issue

The distinguishability of references is considered mandatory in order to determine the most suitable ρ value for a specific pattern set. In turn, the most suitable ρ value is required to reliably identify the group affiliation of the respective PUT (Pattern Under Test).

Reference patterns originate from User Access Gates (UAGs) as described in section 4.3. These UAGs are involved in media sessions and selected as representatives of their logical groups by the procedures introduced in section 6.1. Note that the assignment of a media session to its logical group is performed before this session might be selected as a group representative.

Also note that, like any other media session, a media session nominated as a group representative originally was assigned to its group, which typically requires the application of the grouping procedure described in section 5.3. However, the application of this procedure requires the reliable knowledge of the group affiliations of the reference patterns. Hence, the introduced QoS profiling approach lacks of a start-up mechanism providing the required information. Within the following subsection, a solution concept for this issue is introduced.

5.4.2 Initiating the QoS profiling process

As previously mentioned, the NGN QoS profiling approach introduced in (Weber *et al.*, 2009b) shows a bootstrap issue. This issue results from the unawareness of the group memberships of the first media streams which are initially exchanged after the start-up of the QoS optimisation framework described in chapter 4.

Figure 5.4 shows a mechanism that solves this bootstrap issue. Note that at least three jitter patterns (monitored synchronously at different UAGs) must be available in order to apply this mechanism. These jitter patterns might be derived from the first three media data streams exchanged after the framework start-up. The idea of this bootstrap mechanism is to detect mutual similarity and discrimination features among those three patterns and, hence, identify their group affiliations. No previous knowledge is required regarding any relationship of the pattern sources.

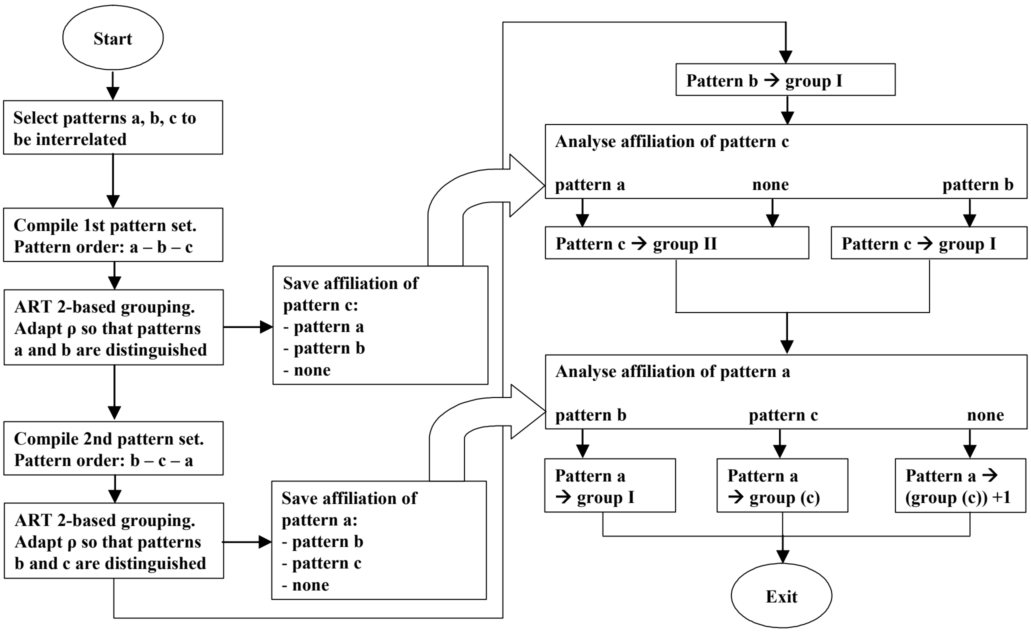

Start
Select patterns a, b, c to be interrelated
Compile 1st pattern set. Pattern order: a – b – c
ART 2-based grouping. Adapt ρ so that patterns a and b are distinguished
Save affiliation of pattern c:
- pattern a
- pattern b
- none
Compile 2nd pattern set. Pattern order: b – c – a
ART 2-based grouping. Adapt ρ so that patterns b and c are distinguished
Save affiliation of pattern a:
- pattern b
- pattern c
- none
Pattern b → group I
Analyse affiliation of pattern c
pattern a
none
pattern b
Pattern c → group II
Pattern c → group I
Analyse affiliation of pattern a
pattern b
pattern c
none
Pattern a → group I
Pattern a → group (c)
Pattern a → (group (c)) +1
Exit

Figure 5.4: Bootstrap mechanism for QoS profiling

Three jitter value patterns (named *a*, *b*, and *c*) are arranged as a pattern set. According to the mechanism described in section 5.3, a first ART 2-based grouping process is performed. During this grouping process, the pattern set is presented to an ART 2 ANN equipped with two output units (hence, two different classes can be distinguished). The ART 2 vigilance parameter ρ is adapted so that patterns *a* and *b* are distinguished, and each of them is assigned to a different output class. Pattern *c*, as a member of the same pattern set, is also considered within the grouping procedure. Depending on similarity characteristics, it might be assigned to the same output class as pattern *a* or to the same output class as pattern *b*. However, if no sufficient similarity is given, pattern *c* will not be assigned to any existing output class. In any case, the grouping result for pattern *c* is stored for further processing.

Subsequently, a second grouping process is accomplished, with the same patterns considered. This time, ρ is adapted so that patterns *b* and *c* are distinguished and assigned to different output classes. In any case, the affiliation of pattern *a* is stored.

In the next steps, the patterns are assigned to virtual groups according to their interrelations. Note that pattern *b* (which represented an ART 2 output class in both classification processes) is considered as the default representative of the virtual group I.

First, the affiliation of pattern *c* to a specific virtual group is determined. If pattern *c* was assigned to the output class that had been represented by pattern *b* in the first classification process, it is obvious that patterns *b* and *c* must be considered as members of the same virtual group. In this case, pattern *c* is associated with virtual group I. If pattern *c* was not assigned to the class represented by pattern *b*, pattern *c* is considered as the default representative of a further virtual group (group II).

Finally, the affiliation of pattern *a* is analysed, resulting from the second ART 2 classification process. If pattern *a* had been assigned to a class either represented by pattern *b* or pattern *c*, pattern *a* is assigned to the respective virtual group. If pattern *a* was not associated with either pattern *b* or pattern *c*, pattern *a* is considered as the default representative of a further virtual group (group II or III, depending on whether class II has already been established).

Table 5.5 shows all possible affiliations and group associations.

Table 5.5: Possible affiliations and group associations of the introduced bootstrap mechanism

#	Affiliation of pattern c	Affiliation of pattern a	Group I	Group II	Group III
1	a	b	b, a	c	
2	a	c	b	c, a	
3	a	none	b	c	a
4	b	b	b, c, a		
5	b	c	b, c, a		
6	b	none	b, c	a	
7	none	b	b, a	c	
8	none	c	b	c, a	
9	none	none	b	c	a

With the bootstrap mechanism introduced within this section, the group affiliations of three synchronously monitored jitter patterns can be autonomously identified. Note that no precognition is required regarding group references or most suitable ρ values. Hence, after two virtual groups have successfully been distinguished by the bootstrap mechanism introduced within this section, NGN QoS profiling as described in section 5.3 can be successfully applied.

5.4.3 Test and conclusion

The introduced bootstrap mechanism for ART 2-based QoS profiling has been evaluated by the use of the research prototype introduced in chapter 8. A SIP-based NGN was set up, and different communication scenarios were simulated. In all scenarios G.711 VoIP calls were set up and several packets of all media streams were recorded and time-stamped at their respective receivers. The resulting data were synchronised and post-processed. In a further step inter-arrival jitter values were calculated for each packet. Several pattern sets were arranged, each comprising three patterns to emulate a bootstrap scenario.

Table 5.6 shows several scenarios considered. The scenarios differ in the number of comprised virtual groups and in the order of the patterns representing the groups. The accuracy stated in Table 5.6 is related to the correct assignment of patterns to virtual groups by the bootstrap mechanism introduced within this section, with ten different pattern sets (each comprising three patterns) tested per scenario.

Table 5.6: Test scenarios for QoS profiling bootstrap mechanism

Scenario No.	No. of virtual groups included	Order of patterns within sets (by group numbers)	Accuracy of group assignment achieved
1	3	I - II - III	100%
2	1	I - I - I	67%
3	2	I - II - II	100%
4	2	I - II - I	100%
5	2	I - I - II	100%

Table 5.6 shows that the bootstrap mechanism introduced within this section provides excellent assignment accuracy for those scenarios in which the considered jitter patterns comprise more than one virtual group. However, in scenario 2, in which all three patterns included within a pattern set belong to the same virtual

group, a limited accuracy is experienced. This is due to the fact that the effectiveness of the introduced bootstrap mechanism depends on the distinguishability of patterns. However, the distinguishability of patterns belonging to the same virtual group is naturally limited. Those patterns must provide significant similarities in order to be defined as members of the same virtual groups.

The bootstrap mechanism for ART 2-based QoS profiling as introduced within this section has been published as a conference paper (Weber *et al.*, 2010).

5.5 Summary

This chapter has discussed the methodology of QoS profiling in detail. Both, the motivation for the creation of this methodology, and the background principle have been presented. QoS profiling is based on the assignment of several media streams to virtual groups by the similarity of the network performance encountered. As the network performance of a media stream comes as a fluctuating set of discrete delay, jitter, and packet loss rate values, a method was proposed for matching value sets regarding shared similarities. This problem has been identified as being solveable by unsupervised pattern recognition. Therefore, the basic characteristics of the most common approaches for unsupervised pattern recognition have been discussed. As a consequence, the applicability of ART 2 ANN for QoS profiling has been further investigated. Therefore both the fundamentals of ART 2-based pattern recognition as well as the verification of its applicability for QoS profiling have been described.

Subsequently, the principle of ART 2-based QoS profiling has been outlined, which is based on the recognition of similarities among several jitter value sequence patterns. Because the characteristics of the jitter value sequences are subject to various influences, the dynamic range among several patterns must be considered. Therefore, in order to allow for the dynamic adaptation of the pattern recognition throughout a QoS profiling process, a method for accuracy self-scaling was developed, which has been introduced within this chapter. Furthermore, in order to allow for the definition of the most appropriate initial classification accuracy when starting the QoS profiling process, a bootstrap mechanism was developed throughout this project. This mechanism has also been described within this chapter.

6 Profile-based QoS estimation

Applying the NGN QoS optimisation framework introduced in chapter 4 involves the estimation of QoS conditions experienced by the receivers of media data streams. Therefore, every media endpoint is equipped with a QoS monitoring entity within its related User Access Gate, be it integrated with the user terminal, or looped-in into the transmission path between user terminal and access network, respectively. These QoS monitoring entities measure the impact of the network performance on IP multimedia data streams that pass the device. The entities are SIP-aware regarding QoS information requests from the central service infrastructure. Upon such requests, SIP is also used to immediately transmit value sequences of resulting QoS characteristics to the centralised unit.

The concept of NGN QoS profiling as introduced in chapter 5 is applied, amongst others, with the benefit of reducing the number of QoS monitoring points required. Within the following subsections, the research done regarding the profile-based estimation of QoS is presented, beginning with a methodology for the selection of suitable QoS reference points in section 6.1. Subsequently, the application of the ITU-T E model for online QoS estimation is discussed, involving several novel extensions to this model (section 6.2).

6.1 Selection of QoS monitoring points

While it might be feasible to let every user terminal transmit a single QoS summary report upon completion of a media session as suggested in (IETF RFC 6035, 2010), the continuous transmission of QoS characteristics from every user terminal during each ongoing multimedia session results in a substantial amount of additional traffic within the transport network. This traffic occupies valuable bandwidth and potentially impairs the network performance. On the other hand, as mentioned in section 3.2.1, it is a mandatory requirement that the central service infrastructure is at any moment aware of the QoS conditions affecting any ongoing media session. This discrepancy is considered within the proposed framework by utilising a QoS profiling-based approach.

The following subsections describe how a comprehensive overview of the QoS provided to every active session can be achieved. Therefore only a subset of all media-receiving UAGs has to be monitored.

6.1.1 Concept for the selection of QoS monitoring points

For the sake of network traffic efficiency, the proposed framework assumes that QoS monitoring is required only in a subset of media-receiving entities at the same time in order to obtain a sufficient overview on the QoS provided to any session. Within the following sections, rules and algorithms are described to achieve this goal.

Based on the following rules, the determination of a UAG is performed for being considered as a monitoring point for the QoS of the media session it currently receives.

- If a virtual group that represents a defined QoS characteristic consists of only one media stream, the UAG receiving this media stream is the only available QoS monitoring point for that virtual group, and hence is selected as the group's QoS monitoring point.

- A UAG, once selected as a QoS monitoring point, communicates to the QoS Manager information on the experienced QoS while the related media session is active.
- If a media session of a UAG acting as monitoring point is terminated and further media sessions exist associated with the related virtual group, new QoS monitoring points for the related communication situation have to be assigned.
- UAGs are typically chosen as QoS monitoring points based on their respective relevance (see section 6.1.2).

As a variant for the selection of monitoring points a set of exceedingly relevant UAGs within a certain group could be preselected as QoS monitoring points. These UAGs are queried in advance to share QoS information of all media sessions in which they will be involved in future with the QoS Manager. If the QoS information obtained from the collection of preselected monitoring points covers communication situations among all active UAG groups no additional UAGs have to be queried.

Preselecting QoS monitoring points can help reducing the signalling traffic for the explicit querying of UAGs for QoS information. If required, preselection can be used in combination with the formerly-mentioned method of per-session selection and in-session querying of monitoring points by their relevance. Depending on the characteristics given in a respective NGN regarding the distribution of UAGs among different virtual groups and the subscribers' communication behaviour, this combination can possibly result in a comprehensive overview on the QoS conditions effective among all identified groups and, at the same time, minimise the amount of signalling traffic required for QoS information querying.

6.1.2 Ranking of reference sessions for QoS monitoring

To activate the monitoring function of a UAG during an ongoing session, additional SIP signalling is required. To increase the efficiency of the monitoring process, media sessions to be monitored should be carefully selected in a way that contributes to the minimisation of the signalling overhead. Hence media sessions that can be expected to be available as references for a relative longer period should be preferred over sessions whose remaining session period can be expected to be short lived.

Prior research identified that call durations can be statistically described as a lognormal distribution (Jedrzycki and Leung, 1996), (Barceló and Jordán, 2000), and (Boggia *et al.*, 2005). Equation 6.1 shows how the elapsed time t of a session relates to the probability density function of a collection of call duration values, following a lognormal distribution, with $\hat{\vartheta}$ being the maximum likelihood estimator (m.l.e.) of the logarithmic mean session duration of the total number of n sessions (see equation 6.2) with T_i being the session duration of the i-th session. According to equation 6.3, $\hat{\sigma}$ corresponds to the maximum likelihood estimator of the standard deviation of $\hat{\vartheta}$.

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \vartheta)^2}{2\sigma^2}} \quad ; \ \sigma, \vartheta > 0; t \geq 0 \qquad (6.1)$$

$$\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} \ln(T_i) \qquad (6.2)$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\ln T_i - \hat{\vartheta})^2} \qquad (6.3)$$

For the lognormal distribution, the point of global maximum (mode) is defined as given in equation 6.4.

$$Mode[X] = e^{\vartheta-\sigma^2} \tag{6.4}$$

The mode corresponds to the maximum value of the probability density.

In order to determine the applicability of a media session as a QoS monitoring reference, a behaviour profile is generated for every subscriber and continuously updated by the respective function of the QoS Manager (see section 4.3). Therefore the length of each incoming and outgoing call of the respective subscriber is considered. According to (Smoreda and Licoppe, 2000), for most individuals, the durations of self-initiated and answered phone calls typically differ. Hence, regarding the selection of reference media sessions for groupwise QoS estimation, it should also be taken into account whether the media receiver was the initiator (i) or the answerer (a) of the communication session, respectively. Thus, for either case (i) and (a) values for $\hat{\vartheta}$, $\hat{\sigma}$, and *Mode*[*X*] according to equations 6.2 to 6.4 are separately determined, and are updated after each call of the respective subscriber.

Table 6.1 shows an example for the heuristic determination of the relevance of media sessions as group references. All media sessions considered in this example are assumed to have been assigned to the same virtual group by the use of QoS profiling (see chapter 5). The receivers of the media streams are identified as subscribers A, B, C, …, J. Subscribers A to D have initiated (i) the communications to which their received media streams belong, while subscribers E to J have answered (a) invitations to communication sessions. For any subscriber a behaviour profile has been generated, separately considering the length of any incoming and outgoing call. Thus, value pairs for $\hat{\vartheta}$(i), $\hat{\sigma}$(i), and $\hat{\vartheta}$(a), $\hat{\sigma}$(a) are available, and so are values for the corresponding modes. For the sake of clarity, only values relevant for the considered case are shown in Table 6.1. Thus, for subscribers A to D, only values for $\hat{\vartheta}$(i), $\hat{\sigma}$(i), and the corresponding *Mode*[*X*](i) time (in seconds) are given, while for subscribers E to J, only $\hat{\vartheta}$(a), $\hat{\sigma}$(a), and *Mode*[*X*](a) are displayed.

Table 6.1: Examplary group reference ranking list

Subscr. ID	Role in current session (i) / (a)	$\hat{\vartheta}$ (i)	$\hat{\vartheta}$ (a)	$\hat{\sigma}$ (i)	$\hat{\sigma}$ (a)	Relevant *Mode*[*X*] time [s]	Elapsed session time [s]	(Mode time) – (Elapsed time) [s]	Reference relevance ranking pos.
A	(i)	5.603		1.112		(i) 79	477	-398 (*)	(*)
B	(i)	5.25		1.373		(i) 29	235	-206 (*)	(*)
C	(i)	6.394		0.167		(i) 582	332	250	**2**
D	(i)	5.464		1.359		(i) 37	22	15	**5**
E	(i)	6.472		1.115		(i) 187	46	141	**3**
F	(a)		5.534		1.434	(a) 32	42	-10 (*)	(*)
G	(a)		6.286		0.585	(a) 381	56	325	**1**
H	(a)		6.12		0.419	(a) 382	255	127	**4**
I	(a)		4.034		0.832	(a) 28	26	2	**6**
J	(a)		6.006		1.083	(a) 126	463	-337 (*)	(*)

(*): Not considered as potential reference, since elapsed session time has exceeded mode time

In order to evaluate an ongoing media session for being potentially relevant as a reference session, the already elapsed session time (given in seconds) has to be considered, being subtracted from the time given by *Mode*[*X*]. In case of a negative result, the considered media session can not be considered as a group reference. In Table 6.1, this is indicated by '(*)' (see subscribers A, B, F, and J).

As stated above, in order to minimise the signalling traffic required for the activation of network performance monitoring functions of UAGs, those sessions should be selected as group reference which can be expected to last relatively long. Hence a relevance ranking is continuously performed for every existing virtual group. In the example given in Table 6.1, the media stream received by subscriber G is most relevant as a group reference, since its current session shows a difference between the mode and the already elapsed session time of 325 seconds, which is the maximum value compared to any other listed subscriber. Following this calculation scheme, the media session received by subscriber C is the second most relevant session, and so on.

Note that the ranking is continuously updated. Hence, the media session received by subscriber I (considered as no. 6 within the snapshot ranking given in Table 6.1) will soon change to be not considerable, since the difference between its mode and the elapsed session time is only 2 seconds.

Note also that at least two reference sessions of the same virtual group should be monitored in parallel, in order to consider unexpected termination of QoS monitored reference sessions (see section 4.5.2).

6.2 Application of the E model for profile-based QoS estimation

Once a media session has been selected as a reference for a virtual group of sessions, the UAG of the respective media stream receiver is queried to transmit to the central QoS Manager information describing the encountered network performance in terms of delay, jitter, and packet loss rates (see section 4.5.2). This information has to be interpreted by the QoS Manager in order to derive information regarding the QoS experienced by the member sessions of the respective virtual group.

As introduced in section 3.1.2, the E model published by ITU-T in (ITU-T G.107, 2009), although originally designed as a so-called network planning tool, comes as a suitable approach to estimate the effective QoS in VoIP communications from measured network performance. Within the proposed framework for comprehensive QoS optimisation, the E model is utilised to construe network performance parameters in dependence of used media codecs. Therefore, extensions are provided to the E model that allow for online QoS assessment.

6.2.1 E model-based QoS assessment

Various research has been performed within the field of E model application for QoS estimation in Multimedia over IP environments. From the analysis of recent publications it was found that most of this research combines both, the E model based QoS optimisation, and the optimisation of the E model itself. Representing this two-sidedness, both (Atzori *et al.*, 2006) and (Li *et al.*, 2008) propose jitter buffer algorithms based on extended E model variations to optimise the relation between delay and packet loss with respect to the perceived QoS. Also aiming on both, the optimisation and application of the E model, (Carvalho *et al.*, 2005) present a software tool for the E model based QoS estimation in VoIP environments. This tool explicitly considers the effect first published by (Clark, 2001) that temporary degradation of the network performance has non-instantaneous effects on the QoS perceived by the user. Regarding the impact of temporary QoS impairments on the overall quality experienced for the complete call, (Carvalho *et al.*, 2005) describe a recency effect. They propose that impairments occurring closer to the end of a call have a stronger impact on the result of subjective quality assessment for the whole call than impairments which have occurred within the beginning phase of the call.

The work proposed by (Bandung *et al.*, 2008) aims on the optimisation of QoS provision in low bitrate VoIP environments, determining a trade-off relation among several network performance parameters, different codecs and their individual requirements, numbers of parallel sessions, and perceived QoS. Therefore, the authors utilise the extended E model published by (Ding and Goubran, 2003). Explicitly aiming on NGN environments, (Lewcio *et al.*, 2009) describe an extension to the E model that considers the effect of switching between narrowband and wideband codecs in VoIP QoS estimation.

Summarising this review it can be pointed out that various examples exist for the successful utilisation of the E model as a tool for the estimation of the QoS experienced by VoIP users. However, it is obvious that, in order to adapt the E model to the demands of real-time QoS assessment, optimisation work is required. Within the following sections, further issues regarding the application of the E model for QoS estimation are discussed, resulting in the proposal of an E model adapted for the application as QoS estimation tool within the proposed framework for comprehensive QoS optimisation.

6.2.2 Consideration of jitter in the E model

Assuming default values for those E model parameters which are not related to network performance or codec characteristics, the following QoS-affecting influences are considered by the E model.

- Codec-specific impacts
- Delay
- Packet loss

As mentioned in section 3.1, beside delay and packet loss effects, also the variance of the delay (jitter) contributes to the influence of the network performance on the QoS of conversational real-time media exchange. However, the jitter is not explicitly considered as an E model parameter according to (ITU-T G.107, 2009), which has served as a motivation for several research work within this field. Amongst others, (Ding and Goubran, 2003) proposed an extension to the E model, introducing an additive credential to the R factor for the consideration of jitter impairments. The respective credential has been modelled as a sum of a parabolic function to consider the effect of self-similarity of Internet traffic, and an exponential function to introduce the influence of the jitter buffer size. Several required coefficients and constants of the jitter-representing credential were determined by simulation for the audio codecs ITU-T G.723.1 ACELP (Algebraic Code-Excited Linear Prediction), ITU-T G.723.1 MP-MLQ (Multi-Pulse-Maximum Likelihood Quantization) (both (ITU-T G.723.1, 2006)), and ITU-T G.729 (ITU-T G.729, 2007). In order to apply this extension to the E model considering further codecs such as ITU-T G.711 (ITU-T G.711, 1988) or iLBC (Internet Low Bitrate Codec) (IETF RFC 3951, 2004), their respective coefficients have to be determined and validated beforehand.

Note that further approaches, also introducing E model extensions for jitter consideration have amongst others been published by (Ren *et al.*, 2010) and (Zhang *et al.*, 2011).

Assuming that every VoIP terminal is equipped with a playout buffer for jitter compensation, the jitter effect can also be considered within the E model by means of a de-jitter buffer emulation. According to (Mei *et al.*, 2005) and (Sengupta *et al.*, 2006), the jitter can be assumed to add up to the end-to-end delay in the case that the de-jitter buffer is able to compensate the jitter. If the delay affecting single media packets differs drastically from the average delay, those packets might be dropped by the de-jitter buffer in order to avoid the increase of the playout delay. In this case, regarding the effect recognisable by the subscriber, a packet loss occurs as a result of the jitter effect.

Due to the fact that it adequately reflects the situation given in VoIP user terminals regarding the QoS-related effects caused by jitter, this thesis explicitly proposes the application of a de-jitter buffer emulation to consider the jitter within the E model. For the sake of clarification, a relatively simple reactive and non-adaptive, “native” jitter compensation algorithm according to (Do Valle *et al.*, 2010) has been selected to generally demonstrate its integration with the E model. Considering a non-adaptive de-jitter buffer algorithm effectively reflects a worst case state which might be given in older or low cost user terminals, hence covering a broad range of different VoIP end systems. For optimised adaptation to specific conditions in future applications, the selected jitter compensation algorithm might be replaced by more sophisticated algorithms.

The selected de-jitter buffer algorithm assumes a fixed reference value (*target*) for the de-jitter buffer depth (time value given in ms) chosen within the borders of the value range [(50…250) ms] according to (Do Valle *et al.*, 2010). The following variables have to be considered.

- n_i = Network delay affecting the *i*-th packet of a media stream while being carried through the network
- d_i = Overall end-to-end playout delay scheduled for the *i*-th packet
- b_i = Time for which the *i*-th packet is effectively retarded within the de-jitter buffer; $b_i = [0\ldots target]$

The context among $n_{i,}$ d_i and b_i is given according to equation 6.5.

$$d_i = n_i + b_i \tag{6.5}$$

The determination of b_i is performed per-packet according to equation 6.6.

$$b_i = \begin{cases} (target - n_i) & \text{if } (n_i < target) \\ 0 & \text{else} \end{cases} \tag{6.6}$$

Note that, according to equation 6.5, the total end-to-end playout delay d_i of the *i*-th packet already includes both, the network delay and the delay potentially added by the de-jitter buffer.

In case that a packet is massively delayed during its transport through the network, the time n_i for which the packet remains in the network might exceed the scheduled overall playout delay d_i. In this case, the packet is dropped by the de-jitter buffer in order not to retain additional playout delay. The respective packet has to be considered being lost according to equation 6.7, with L_{Ji} being a Boolean flag, marking the *i*-th packet with 1 in case it has to be counted as a lost packet due to de-jitter buffer timeout.

$$L_{Ji} = \begin{cases} 1 & \text{if } (n_i > d_i) \\ 0 & \text{else} \end{cases} \tag{6.7}$$

6.2.3 Considering packet loss in continuous online QoS estimation

Beside the issues that have already been addressed by various research work as introduced in sections 6.2.1 and 6.2.2, another crucial point regarding the application of the E model exists, concerning the determination and comprehension of packet loss in continuous real-time QoS estimation. Originally the E model had been designed to determine a statistical output value representing the average QoS perceived by users in dependence of statistical input values (such as codec-specific characteristics, delay, and parameters related to packet loss occurence). In this sense, in order to consider the loss of packets, the following parameters have to be provided to the E model, entering into the calculation of Ie_{eff} according to (ITU-T G.107, 2009) (see equation 6.8; Ie represents a codec-specific parameter related to zero packet loss ($Ppl = 0$)).

- Bbl = Packet-loss robustness factor (codec-specific) [-]
- Ppl = Random packet-loss probability [%]
- $BurstR$ = Packet-loss burst ratio [-]

$$Ie_{eff} = Ie + (95 - Ie) \cdot \frac{Ppl}{\frac{Ppl}{BurstR} + Bpl} \tag{6.8}$$

Note that Ppl represents the percentage rate of random packet loss occurring throughout a call. As such, it is determined by the ratio of (number of packets lost) over (number of packets sent). Hence, a reasonable Ppl value can not be calculated on a per-packet basis, because Ppl would be 100 percent in case the packet is lost. Beside the fact that considering solely this snapshot value would result in an unrealistic break within a respectively derived QoS characteristic, the permitted range for *Ppl* is defined as $(0 \leq Ppl \leq 20)$ % according to (ITU-T G.107, 2009).

In order to consider this limit and to allow for realistic influence of single packet loss occurance in E model based real-time QoS estimation, a novel extension to the E model is proposed within this thesis.

Explicitly this extension focuses on the consideration of random packet loss as a running mean value of *x* consecutive packets, re-calculated for every packet (see equation 6.9). This allows for continuous real-time QoS estimation based on the E model.

$$Ppl_k = \frac{100}{x} \sum_{i=j}^{k} L_i \quad ; j=1+k-x \tag{6.9}$$

According to equation 6.9, Ppl_k reflects the uncorrected percentual packet loss rate as a running mean value, computed for the *k*-th packet of the media stream.

x refers to the number of packets to be considered for the running mean calculation. According to this research it has been found that a maximum of 100 considered packets results in acceptable feedback to occurring packet loss (see equation 6.10).

$$x = \begin{cases} j & \text{if } j<100 \\ 100 & \text{else} \end{cases} \tag{6.10}$$

According to equation 6.11, L_i represents a Boolean indicator for the loss of the *i*-th packet.

$$L_i = \begin{cases} 1 & \text{if } i\text{-th packet is considered to be lost} \\ 0 & \text{if } i\text{-th packet is considered to be played out} \end{cases} \qquad (6.11)$$

Note that, for integration of the E model extensions proposed in section 6.2.2, L_i has also to be set for packets that are discarded by the de-jitter buffer. Hence, L_i according to equation 6.11 integrates L_{Ji} according to equation 6.7 (see section 6.2.2).

In order to consider the permitted value range of *Ppl* as defined in (ITU-T G.107, 2009), if required, Ppl_k is corrected according to equation 6.12, referred to as Ppl_{kc}.

$$Ppl_{kc} = \begin{cases} Ppl_k & \text{for } 0 \le Ppl_k \le 20\ \% \\ 20\ \% & \text{for } Ppl_k > 20\ \% \end{cases} \qquad (6.12)$$

6.2.4 Proposed E model refinements

In order to introduce the extensions and contributions to the E model discussed in sections 6.2.2 and 6.2.3 adaptions have only to be considered for those components of the E model which involve delay and packet loss variables. Due to the fact the definitions made in sections 6.2.2 and 6.2.3 affect any occurrence of the respective delay and packet loss parameters within the E model, its general definition and computation algorithms remain unchanged. Therefore only the concerned E model parameters are redefined according to Table 6.2. This thesis proposes these modifications regarding the utilisation of the E model for continuous online QoS estimation on a per-packet basis. Within Table 6.2, the indices *i* and *k* refer to the *i*-th or *k*-th packet of a media data stream, respectively.

Table 6.2: Redefinition of E model parameters for continuous online QoS estimation

E model Parameter abbreviation	Parameter meaning	Description of modification	Formal equation of modi-fication	Related thesis section and equation
T	Mean one-way delay	Inclusion of de-jitter buffer emulation playout delay for jitter consideration	$T_i = d_i$	Section 6.2.2, equation 6.5
Tr	Round-trip delay; typically *Tr* = 2*T*	Inclusion of de-jitter buffer emulation playout delay for jitter consideration	$Tr_i = 2\ d_i$	Section 6.2.2, equation 6.5
Ta	Absolute (one-way) delay; typically *Ta* = *T*	Inclusion of de-jitter buffer emulation playout delay for jitter consideration	$Ta_i = d_i$	Section 6.2.2, equation 6.5
Ppl	Random packet-loss probability	Inclusion of de-jitter buffer emulation packet discard for jitter consideration, and running mean valuation of packet loss percentage for real-time E model application	$Ppl_k = Ppl_{kc}$	Section 6.2.2, equation 6.7 and Section 6.2.3, equations 6.9…6.12.

6.3 Summary

Within this chapter, the framework functions for comprehensive estimation of QoS conditions as encountered by any media session have been described. Due to the QoS profiling-based virtual grouping of media streams which are affected by similar QoS conditions, it is sufficient to monitor the encountered QoS only in a subset of user terminals and assume similar QoS conditions for any other member session of the same virtual group. In order to increase the efficiency of this monitoring process, the selection of QoS monitoring points is a crucial topic which has been outlined within this section. In order to provide efficient QoS monitoring, it has been found that especially those sessions should be considered as reference sessions which are assumed to show the longest remaining session duration. A heuristic ranking algorithm based on lognormal distribution has been described to determine those sessions best considered as reference sessions, which are sorted and provided as a ranking list of sessions to be considered as reference sessions for QoS monitoring.

Furthermore, within this chapter, the utilisation of the ITU-T E model has been discussed regarding its applicability for continuous online QoS estimation of VoIP media streams. It was found that the E model is generally applicable for the purpose of QoS estimation inline with the proposed QoS optimisation framework. However, some modifications to the E model are required. An approach for jitter consideration (which is not included in the original E model) has been proposed, based on the introduction of a de-jitter buffer emulation.

Furthermore, a contribution has been described regarding the consideration of packet loss occurence in continuous online QoS estimation. A refinement of the E model has been proposed, redefining several delay- and packet loss-related E model parameters, preparing the E model for continuous online QoS estimation within the proposed framework for comprehensive QoS optimisation.

7 Feedback control-based optimisation of QoS conditions

Beside QoS estimation, the optimisation of QoS conditions for real-time media streams is the second main aspect of QoS management in NGN. Since the top-down allocation and release of transport network resources comes with potential disadvantages such as cross-layer signalling (see section 3.3.3), a more passive approach is pursued within this project, avoiding the active control of transport network elements.

IP network paths showing a lower level of utilisation provide improved network performance (thus, QoS conditions) compared to considerably busy network paths (Kim and Jeong, 2011). Hence it can be said that for multimedia over IP, the performance of a transport network path can be considered as an indicator for the level of tolerable utilisation of the respective path. Following this comprehension, it has been decided that – within this project – the optimisation of QoS conditions for real-time communication media streams shall be based on the control of the volume of media traffic carried on respective IP transport paths. This control has to be performed under avoidance of the active management of IP transport resources (such as resource reservation) and, hence is termed *passive control*. This chapter introduces methods and algorithms to optimise QoS conditions for real-time media communication in NGN, utilising the passive control of network utilisation. Section 7.1 introduces the principle of passive network utilisation control, and summarises the research in this area. Within the proposed framework, the passive control of the network utilisation is performed by a feedback control system that is introduced in section 7.2. Subsequently algorithms are presented for the selection of sessions that are considered for passive network utilisation control action (section 7.3). Finally, the evaluation of the proposed QoS optimisation approach is demonstrated based on the simulation of a communication scenario (section 7.4).

7.1 Efficient passive control of network utilisation in NGN

According to (Kim and Jeong, 2011), the utilisation of bandwidth required for the transmission of VoIP media streams can be significantly reduced by adopting audio codecs which are optimised for providing acceptable speech quality and, at the same time, require a considerably lower amount of bandwidth. Figure 7.1 gives an overview of the speech quality levels achievable (in MOS LQO) with a variety of prevalent low bitrate voice codecs, and their respective gross bitrate demands. The E model has been used for the determination of the MOS values.

As shown in Figure 7.1, as an example, applying the GSM-EFR audio codec (Global System for Mobile communications – Enhanced Full Rate) saves more than 57 percent of gross bitrate compared to the G.711 audio codec (38.6 kbit/s (GSM-EFR) versus 90.4 kbit/s (G.711)), providing an only marginally reduced level of achievable speech quality (MOS 4.29 (GSM-EFR) versus MOS 4.41 (G.711)). Hence, by dynamically forcing the use of appropriate compressing codecs, if required, further transport resource capacity is made available, resulting in the optimisation and recovery of acceptable QoS conditions without appreciable loss of media quality.

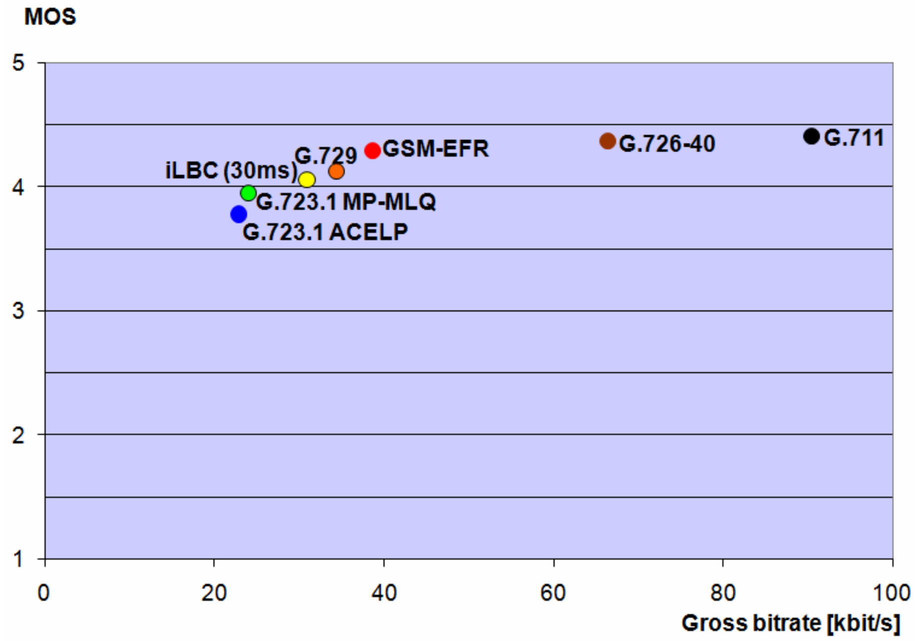


Figure 7.1: MOS LQO over gross bitrate demands of different audio codecs

Within the process of SIP session initiation, the selection of the media codecs to be used is typically based on an end-to-end negotiation procedure between the involved user terminals according to the SDP offer-answer model. Managing the overall media traffic volume by manipulating codec selections requires a centralised entity, able to restrict the codec choice if required. Note that this task should be fulfilled in conjunction with Call Admission Control (CAC) being executed within the service provider infrastructure upon request of a recent communication session. Hence, if a communication session is requested which, if granted, would result in the emergence of a further media session being a member of a virtual group (see section 5) concerned by potential shortage of network performance, this session might be granted only under "downgrade" conditions, which means that the user terminals are forced to use a bandwidth-efficient media codec. The general principle to employ codec downgradings to manage the network utilisation for passive QoS control has been published as a research paper in (Weber *et al.*, 2008).

If network performance conditions within a virtual group are rather critical, a recently requested communication session might also be rejected by CAC, if granting would result in potential further deterioration of the network performance of a respective virtual group of media sessions.

In order to allow for immediate reaction on impaired network performance, beside taking influence on the codec choice of recently upcoming sessions, also the dynamic change of codecs of already existing media sessions is destined for the management of network utilisation by the framework which has been developed within this project. That is, in order to recover QoS conditions in case of the congestion of a network path, from the concerned virtual group existing media sessions might be "downgraded" by being forced to switch to more bandwidth-efficient media codecs. In case of critical network performance, selected existing sessions might also have to be terminated by remote in order to immediately release occupied network capacity and hence, restore acceptable QoS conditions. The respective procedures and signalling are described in section 4.6.

7.1.1 Existing approaches for passive network utilisation control

Some research has been carried out by different teams, all describing the principle of downgrading (or degrading, respectively) and cancellation/rejection of multimedia sessions (or data transfer, respectively) for the benefit of optimising either the utilisation or QoS conditions of network paths. (Abdelzaher and Shin, 1998) describe an algorithm for the dynamic assignment of bandwidth available on a transport path among several users, resulting in a balanced overall network performance. In consequence, every user is provided with a certain amount of bandwidth within minimum and maximum limits which are guaranteed to the user through a SLA (Service Level Agreement). However, although in this approach network resources are not actively managed by a central control entity, this approach assumes the application of end-to-end IP QoS mechanisms such as RSVP to control effective QoS conditions in the transport network.

Recently passive network utilisation control has been especially proposed as a method for QoS adaptation in multimedia transmission over wireless networks by several different research teams. For example, (Yang *et al.*, 2005) describe a resource management approach for heterogeneous wireless networks. Applying this approach, a service provider can control the throughput for different media (such as voice, video, and data) by the stepwise degradation of the respective sessions among four so-called quality levels (*Excellent*, *Good*, *Basic, Rejected*). Per quality level, bandwidth limits are defined per medium. The quality level *Rejected* refers to the cancellation or refusal of the respective communication. However, although taking into account different user priorities, the history of the QoS perceived by users is not considered within this approach.

(Sfairopoulou *et al.*, 2008) have investigated the effects of different CAC and downgrading policies on the resulting call blocking and dropping probabilities and experienced average quality (in terms of MOS values) provided for voice transmission over WLAN. As a result, an approach is presented to find the best trade-off among the three measures call blocking, call dropping, and perceived QoS, combining both, the downgrading and rejection/cancellation of recently requested and already existing sessions. To achieve this, a new Grade of Service (GoS) factor is introduced, combining both call quality and quantity measures. In (Sfairopoulou *et al.*, 2011) an updated approach is proposed, including a redefined GoS factor called VGoS factor. However, both approaches do not consider the selection of sessions to be downgraded or cancelled, respectively. Furthermore, the quantification of sessions to be downgraded or cancelled/rejected is not considered in (Sfairopoulou *et al.*, 2008). The updated approach (Sfairopoulou *et al.*, 2011) adopts an algorithm which originally has been proposed by (Hole and Tobagi, 2004). This quantification algorithm explicitly considers characteristics specific to 802.11 MAC and physical layers. Hence, it is only applicable for the control of the network utilisation of wireless access networks based on the 802.11 specification.

A variant of this algorithm has also been applied within a different approach by (Tüysüz and Mantar, 2010), who propose a cross-layer approach to improve voice quality and data throughput over multirate WLAN. This approach also utilises the principle of media codec degradation, additionally considering the manipulation of the codec-dependent frame size of the media packets. The feedback on the actually perceived QoS is determined from both, RTCP (RTP Control Protocol, provided by the applications) and information obtained from the MAC layer, provided by the WLAN access points. Hence, this approach is also limited to WLAN access networks.

The session-based QoS management architecture introduced by (Tebbani *et al.*, 2008) also has explicitly focussed on multimedia transmission over WLAN, allowing for the assessment and control of provided QoS conditions. This cross-layer architecture allows for both, DiffServ-based QoS manipulation within the network, and network utilisation control based on the ability to cancel existing sessions. The approach is refined in (Tebbani *et al.*, 2009), introducing an extended architecture and an algorithm for the network utilisation control. Therefore, specific network elements are introduced, providing cross-layer information transmission and processing. In (Tebbani and Haddadou, 2008), both media codec downgrading (from G.711 to G.729 audio codecs) and codec frame size adjustment have been additionally introduced. However, since all considered variants of this approach are based on cross-layer signalling and require a modified transport network infrastructure, potential issues arise with this approach, such as described in section 3.3.5.

Summarising the current research in the field of QoS management for multimedia over wireless access networks, (Sfairopoulou *et al.*, 2011) provide an overview about the most relevant approaches which utilise mechanisms of codec downgrading and session cancellation/rejections for wireless access network utilisation control.

In order to optimise the utilisation of network paths, already existing sessions can be dynamically forced to switch to more efficient media codecs (“downgrading”) in order to reduce the overall bandwidth demand for the concerned virtual group (note that the definition and formation of virtual groups is described in chapter 5). The protocol-based execution of a downgrading process for existing media sessions is described in section 4.6.1. However, if a virtual group of media sessions is concerned by severe lack of network performance, the downgrading of sessions might not be sufficient to immediately recover QoS conditions. In this case, selected existing communication sessions might also be cancelled. The process of a central-driven session cancellation for the sake of QoS recovery is further described in section

4.6.2. Summarising, if critical network utilisation is detected which affects members of a considered virtual group, the following four cases of action can be distinguished.

A recently requested session …

- might be granted under downgrade conditions or
- might be denied.

Existing sessions …

- might be downgraded or
- might be intentionally cancelled.

Dedicated algorithms are required in order to decide whether and how many sessions are going to be downgraded or cancelled/rejected, respectively. Within the following sections, those algorithms are introduced which have been developed within this research project, to be applied within the framework for comprehensive QoS optimisation.

7.2 Feedback control system for the optimisation of QoS conditions

As stated in section 7.1, the performance of a network path (and hence, the QoS conditions it provides) can be passively controlled by adapting the overall bandwidth demands for media sessions on the respective path. Therefore, two methods have been identified: release of bandwidth through codec downgradings or session cancellations. Regarding voice communication, a codec downgrade can be performed in a way that it only marginally impairs the QoS experienced by the concerned subscribers depending on the achievable MOS provided by both, the original and the subsequently used codec (see Figure 7.1 in section 7.1). In contrast, as derived from (Yang *et al.*, 2005) and (AlQahtani and Mahmoud, 2006), both the rejection of a session request and especially the unmeant termination of an ongoing session are experienced as considerable discomfort by the concerned subscribers. On the other hand, releasing network resources through the enforced termination/rejection of whole communication sessions comes as a more efficient method compared to the optimisation of bandwidth demands by the use of codec downgrading procedures. Therefore, within this project, both methods have been considered and applied in combination. However, their respective applications are well-distinguished and limited. Within the following subsections, the management of QoS conditions through the control of the network utilisation is described.

7.2.1 Proposed feedback control circuit

Within the defined framework, no knowledge of the architecture of a transport network involved in media stream transport is assumed. However, the framework classifies media streams to virtual groups by the similarity of the network performance to which they are exposed. Since the monitoring of the network performance and the subsequent assignment to virtual groups is performed in a fine-grained manner, it can be assumed that those media streams which are assigned to the same virtual group are commonly carried over those transport network paths whose network performance characteristics have the strongest influence on the QoS conditions perceived by the respective media sessions. Hence, in order to control QoS conditions encountered by those media sessions which form a virtual group, the utilisation of the respective transport network path has to become controlled.

Since the framework developed throughout this research project is not able to actively take an influence on transport network characteristics, the control of the utilisation of network segments can only be performed by varying the controllable portion of traffic induced. Regarding the scope of this research, this is provided by the manipulation of the characteristics of media sessions or their life times, respectively. Figure 7.2 shows the control circuit which has therefore been defined as a part of the developed framework for QoS optimisation.

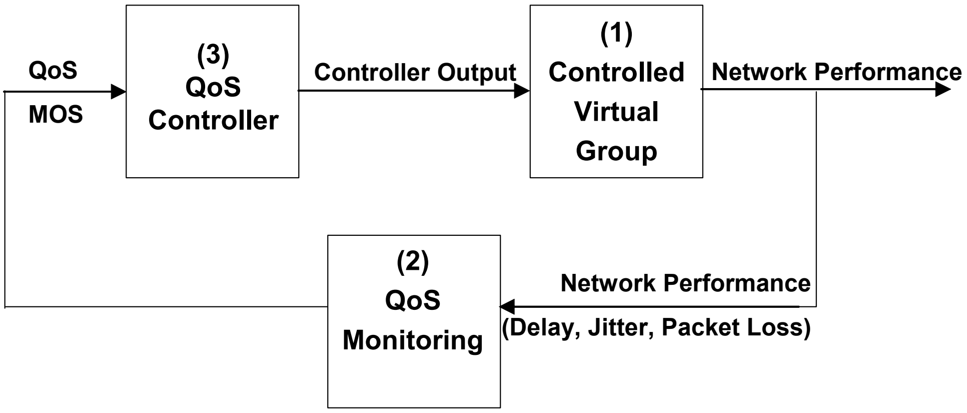


Figure 7.2: Control circuit for QoS optimisation (per virtual group)

The control circuit consists of three logical units, numbered (1)…(3). The functions comprised by these units are described in the following.

(1) As system to be controlled, the control circuit comprises the considered virtual group, including any media session assigned to it. Every media session contributes a certain amount of traffic to the virtual group. The bitrate demands of all member sessions add up to the overall bitrate occupied by the considered virtual group (being one of its main characteristics). This characteristic significantly influences the network performance encountered by the media streams assigned to the considered virtual group. Hence, by manipulating the overall bitrate demand of a virtual group, the network performance provided to its member sessions can be modified. For more

information on the formation and characteristics of virtual groups, see chapter 5.

(2) In order to be aware of the network performance perceived by the member sessions of a virtual group, a QoS monitoring system is applied. This system measures the network performance in terms of packet delay, jitter, and loss rates as encountered by member sessions of the considered virtual group, and derives MOS values as a measure for the QoS perceived by the subscribers. See chapter 6 for further information on the QoS monitoring and estimation functions included in the developed framework.

(3) In order to allow for the optimisation of QoS conditions provided by the considered virtual group, the output of the QoS monitoring is used as input for a QoS controller, which in turn provides an output signal processed by the controlled system. Since the framework is designed solely for the optimisation of QoS conditions rather than for the absolute control of the transport network performance, the QoS controller is implemented as an extremum seeking controller. In contrast to standard controllers applied to most control circuits, this type of controller does not require an external reference input, but includes functions to optimise the performance of the controlled system in itself. This is achieved by continuously readjusting the behaviour of the controlled system in a way so that the feedback measure provided to the controller indicates the best performance of the system achievable at the respective point in time. The detailed function provided by this controller is described within the following section.

7.2.2 QoS controller functionality

Within section 4.5, the general method of comprehensive QoS estimation as proposed within this project is outlined. Section 6.2 describes in detail the derivation of QoS-representing MOS values from the network performance measured in reference QoS monitoring points provided by so-called User Access Gates. The selection of these references is described in section 6.1.

Since the active control of bandwidth amount or network performance provided by transport network resources typically requires cross-layer signalling, it is not considered as an adequate QoS control technique to be used within this research project. The developed framework rather aims on the optimisation of QoS conditions on dedicated network paths through the control of the demanded traffic volume. This comes by the cost of only having relative influence on effective QoS conditions provided on a considered network path. Therefore, it has been decided to choose a rather soft control mechanism not based on absolute QoS measures, but on distinct conditions such as the gradient of the MOS evolution. This control mechanism is implemented as an extremum seeking controller, provided with a simple but effective control algorithm which is introduced within the following.

Observing the evolution of QoS characteristics of a virtual group of media sessions within a defined time period can provide several different outcome value sequences, representing different conditions of the QoS evolution. Within this project, the QoS controller has been designed to distinguish and consider the following five conditions identified by the analysis of the QoS evolution (assuming the consideration of the mean value evolution after smoothing).

1. Stable MOS above a defined threshold value that represents the limit for required intervention (in form of optimisation action)
2. Continuously increasing MOS, such as observed in case of a QoS recovery effect after a previously occured QoS deterioration

3. Continuously decreasing MOS, such as observed in case of deterioration of QoS conditions. If required, this condition can be further subclassified regarding the strength of the decrease.
4. Stable MOS below a defined threshold value that represents the limit for required intervention
5. Oscillating MOS among a defined threshold value that represents the limit for required intervention

In order to manage the QoS provided on a network path, depending on the respective condition, different action has been identified to be required in order to provide stabilised QoS conditions on an acceptable level. Table 7.1 provides an overview of the distinguished conditions, the respective action which has to be triggered by the controller, the expected follow-up condition, and the resulting consequence if the condition does not respond to the proposed action (B/W = bandwidth; Min = Minute). Regarding the detailed settings, the following assumptions are made.

- Condition identification: The identification of the QoS condition is based on the monitoring and analysis of the MOS evolution of a selected QoS reference session. The MOS characteristic is preprocessed by the use of a running mean algorithm, taking into account the variation of the actual MOS of the previous two seconds, which is computed for every received media packet. The tendency of the MOS variation is derived from the continuous comparison of the current running mean MOS value with the corresponding value effective before a previous interval of two seconds.
- Conditions 3 a) and 3 b), 4), and 5): Condition 3 a), 4), and 5) refer to light continuous QoS decrease between 0.3…0.89 MOS/Min., while condition 3 b) refers to strong continuous QoS decrease of 0.9 MOS/Min. or more. The percentages for both, the initial soft bitrate reduction (performed as codec downgradings) and hard bitrate reduction have been set to 10 percent. However, it is assumed that hard bitrate reduction provides a stronger effect

regarding the immediate release of occupied bandwidth, compared to the gradual reduction of bandwidth occupation in case of soft bitrate reduction. If the continuation of soft bitrate reduction after an initial action is indicated, the effective reduction of bandwidth occupation can exceed 10 percent.

Regarding the controller setup as proposed within this project, the effectiveness of the assumed rates has been empirically confirmed by prototypical simulation (see section 7.4). If required, these rates might be adapted to respective meaningful levels for use in modified framework scenarios.

In case that the exhaustive application of a proposed action does not result in successful QoS recovery, the service operator must be notified, be it for informal reasons, or in order to take any arbitrary action out of scope of this thesis to manage the active control of network resources. Therefore, if a condition does not respond to the destined action initiated by the controller, the effective amount of bandwidth is kept at the given minimum level and an alarm is triggered, informing the service provider of the identified issue.

Figure 7.3 shows the QoS optimisation procedure performed by the QoS controller for every identified virtual group, considering the QoS conditions and actions as introduced in Table 7.1.

As described above, the QoS controller adjusts QoS conditions for media sessions by controlling the network utilisation through codec downgradings and/or cancellations and rejections of media sessions. Within the following section, the procedures for the selections of the concerned sessions are described.

Table 7.1: Distinguished QoS conditions and required action

Condi-tion no.	Condition description	Action to be taken	No. of follow-up condition expected in case of success	Consequence if condition does not respond to action as expected
1	Stable QoS above threshold MOS	none	n/a	n/a
2	Continuous QoS increase	Allow for gradual B/W increase by granting recently requested sessions without limitations (no active upgrades!!)	1	n/a
3	a) Light continuous QoS decrease	Initial action: Soft reduction of B/W occupation by 10 percent through codec downgradings. If condition still persists: continue downgrading until every session is downgraded to its respective minimum. Follow-up action (if condition persists): Hard reduction of B/W occupation by session denial/cancellations until left occupation (in terms of B/W) is reduced by 10 percent.	2 or 4	Maintain achieved minimum B/W level. **TRIGGER ALARM**
	b) Strong continuous QoS decrease	Initial action: Hard reduction of B/W occupation by session denial/cancellations until left occupation (in terms of B/W) is reduced by 10 percent. Follow-up action (if condition persists): Soft reduction of B/W occupation until every session is downgraded to its respective minimum.		
4	Stable QoS below threshold MOS	As for condition 3 a)	2	Maintain achieved minimum B/W level. **TRIGGER ALARM**
5	QoS oscillates among threshold MOS	As for condition 3 a)	1, 2, or 4	Maintain achieved minimum B/W level. **TRIGGER ALARM**

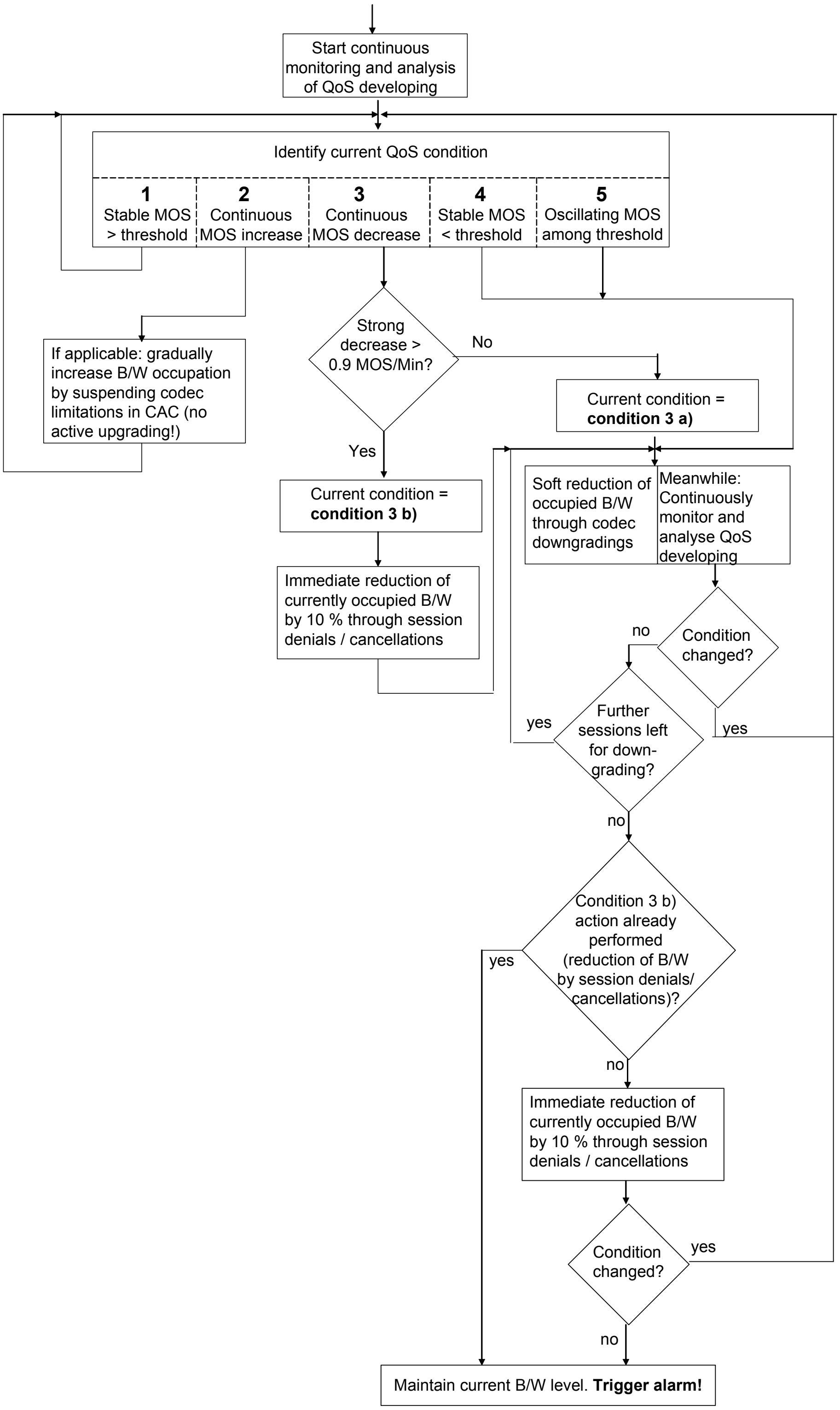

Start continuous monitoring and analysis of QoS developing
Identify current QoS condition
1 Stable MOS > threshold
2 Continuous MOS increase
3 Continuous MOS decrease
4 Stable MOS < threshold
5 Oscillating MOS among threshold
If applicable: gradually increase B/W occupation by suspending codec limitations in CAC (no active upgrading!)
Strong decrease > 0.9 MOS/Min?
No
Yes
Current condition = condition 3 a)
Current condition = condition 3 b)
Immediate reduction of currently occupied B/W by 10 % through session denials / cancellations
Soft reduction of occupied B/W through codec downgradings
Meanwhile: Continuously monitor and analyse QoS developing
Condition changed?
no
yes
Further sessions left for down-grading?
yes
no
Condition 3 b) action already performed (reduction of B/W by session denials/ cancellations)?
yes
no
Immediate reduction of currently occupied B/W by 10 % through session denials / cancellations
Condition changed?
yes
no
Maintain current B/W level. Trigger alarm!

Figure 7.3: QoS optimisation procedure performed by the QoS controller

7.3 Selection of sessions considered for passive network utilisation control action

As stated amongst others in (Sfairopoulou *et al.*, 2008), both the downgrading of media sessions to low bitrate codecs and the rejection or cancellation of communication sessions are potential opportunities to recover and maintain acceptable network performance in case of QoS shortage due to increases in network utilisation. Among several proposals regarding the passive control of the network utilisation summarised in section 7.1.1, different approaches are described for the selection of those media sessions to be considered for being downgraded or cancelled (or rejected), respectively. (Yang *et al.*, 2005) propose a bonus malus point system, assigning reward points to every media session, depending on its respectively granted quality level, the medium and the priority class of the respective user. For sessions that are rejected, intentionally cancelled or accidently dropped, malus points are assigned, indicated as negative values. In case of an identified bandwidth shortage, in order to determine which particular session should be considered for being downgraded or rejected, (Yang *et al.*, 2005) propose a degradation utility value calculated according to equation 7.1.

$$\textit{Degradation utility} = \textit{(Released bandwidth) / (Lost reward points)} \qquad (7.1)$$

If required, those sessions are supposed to be concerned by any degrading action that provide the highest degradation utilities. However, although generally considering individual user priority levels, (Yang *et al.*, 2005) do not explicitly take into account the individual experience of users which were previously already concerned by degradation, be it codec downgrading or session cancellation/rejection.

For the sake of simplification, the passive network utilisation approaches proposed in (Sfairopoulou *et al.*, 2008) and (Sfairopoulou *et al.*, 2011) assume that sessions considered for downgradings or cancellations/rejections are chosen randomly. That is, sessions are not provided with any preference regarding downgrading or cancellation/rejection decisions.

Regarding the selection of sessions to be considered for downgrading or cancellation, (Tebbani *et al.*, 2008) propose that those sessions should be considered which provide a lower priority compared to other sessions. The priority of a session is derived from the priority of the involved users, which is assumed to be determined according their respective SLAs. However, (Tebbani *et al.*, 2008) have not considered the effect described by (Yang *et al.*, 2005), that codec downgrading, cancellation of existing sessions, and rejection of requested sessions each come with different levels of impairment regarding the user experience of a session.

Therefore, within this thesis, different metrics are identified to be meaningful regarding the selection of sessions either considered for downgrading action or cancellation/rejection, respectively. While switching a media session to a low bitrate codec has a relatively low impact on the QoS of a session, the rejection or cancellation of a requested or already active session is typically experienced as a strong impairment of the service quality. Due to this fact, a clear distinction is made regarding the selection of sessions which are either downgraded or rejected/cancelled, respectively. While downgrading decisions are suggested to be based on efficiency factors such as the possible amount of traffic volume saved over the time, regarding the selection of sessions to be considered for rejection/cancellation the objective session significance has been identified as an essential factor. That is, a session which is objectively more important should be preferred over a less important session regarding its continuance. Beside the user priority reflecting the customer state (such as being a VIP, premium, or standard customer) according to a SLA contract closed between a user and a service provider, further credentials (such as the overall quality level experienced by a user in the past) should be considered for the determination of the objective session significance.

Within the following sections algorithms are introduced for the selection of media sessions to be downgraded, denied, or cancelled in order to control the utilisation of transport network paths associated with particular virtual groups, and hence to manage their experienced QoS conditions. Therefore voice has been selected as reference medium, although the algorithm can easily be adapted to any other medium.

7.3.1 Selection of sessions considered for downgrading action

If the downgrading of media sessions is required in order to recover QoS conditions for member sessions of a particular virtual group, sessions have to be selected for being considered as downgrade candidates.

The permanent downgrading of a media session to a low bitrate codec comes along with the following characteristics.

- Reduction of the bitrate demands of the session for its remaining life time
- Reduction of the maximum QoS (in terms of MOS values) provided to the concerned user for the remaining session life time
- Additional signalling for codec renegotiation

Considering these factors, for every existing virtual group, every media session has to be classified into a ranking list of sessions to be considered for being downgrade candidates. Therefore, for each session a quantitative index D_i is calculated which represents the reasonableness that this session should be considered for downgrading action, compared to any other concurrent session. For the calculation of D_i the efficiency of the codec to switch to is considered, as well as the achievable bitrate saving and the estimated remaining session duration.

Equation 7.2 shows the proposed way of indexing media sessions into the ranking list of downgrade candidates.

$$D_i = E_{nC} \cdot \Delta BW_{oC/nC} \cdot t_s \tag{7.2}$$

This equation comprises the following variables.

- D_i = Downgrade index of a particular media session, to be used as a quantitative index for the classification of a media session into the ranking list of sessions to be considered for downgrading. A session providing a high D_i value is preferred to be downgraded over another session providing a relatively lower D_i value. Formally, for audio sessions, D_i comes with the abstract unit [MOS · s].
- E_{nC} = Efficiency Quotient of the most efficient codec available for a particular session. Within this research it is proposed that the efficiency of narrowband audio codecs is defined according to equation 7.3. For other media types, the codec efficiency might be defined in a different way.

$$E_{nC} = \frac{MOS_{nC}}{GrossBitrate_{nC}} \tag{7.3}$$

From all codecs available for a particular media session, the codec providing the highest efficiency among all available codecs is considered as a downgrade choice. Hence, its E_{nC} is considered for the calculation of D_i. Formally, E_{nC} comes with the abstract unit [MOS / (kbit/s)]. For a selection of common narrowband audio codecs, Table 7.2 provides Efficiency Quotient values in descending order, assuming ideal transport network performance.

Table 7.2: Efficiency Quotients and related characteristics of selected narrowband speech codecs

Codec Name	**Typical payload per packet [Byte]**	**Packet Interval [ms]**	**Netto Bitrate [kbit/s]**	**Gross Bitrate [kbit/s]**	**R Factor**	**MOS CQE**	**Efficiency Quotient [MOS / (kbit/s)]**
G.723.1 ACELP	20	30	5.3	22.9	74	3.78	**0.1651**
G.723.1 MP-MLQ	24	30	6.4	24	78	3.95	**0.1646**
iLBC 13.33	50	30	13.33	30.9	81	4.06	**0.1314**
G.729	20	20	8	34.4	83	4.13	**0.1201**
GSM (HR)	14	20	5.6	32	70	3.6	**0.1125**
GSM-EFR	30.5	20	12.2	38.6	88	4.29	**0.1111**
G.728 (16)	40	20	16	42.4	86	4.23	**0.0998**
iLBC 15.2	38	20	15.2	41.6	83	4.13	**0.0993**
GSM (FR)	32.5	20	13	39.4	73	3.73	**0.0947**
G.726-32	80	20	32	58.4	86.2	4.24	**0.0726**
G.726-24	60	20	24	50.4	68.2	3.51	**0.0696**
G.726-40	100	20	40	66.4	91.2	4.37	**0.0658**
G.711 µ (PCMU)	160	20	64	90.4	93.2	4.41	**0.0488**
G.711 a (PCMA)	160	20	64	90.4	93.2	4.41	**0.0488**

- $\Delta BW_{oC/nC}$ = Gross bitrate [kbit/s] which is saved if the codec downgrading from the original codec (*oC*) to the new codec (*nC*) will be performed. See equation 7.4 for the calculation of $\Delta BW_{oC/nC}$.

$$\Delta BW_{oC/nC} = (GrossBitrate_{oC} - GrossBitrate_{nC}) \qquad (7.4)$$

- t_s = Estimated remaining life time of the considered session, calculated as the difference of the most likely session duration and the already elapsed time. See section 6.1.2 for details on the determination of the most likely session duration.

Table 7.3 provides an example for the determination of the relevance of media sessions for being considered to be downgraded. All media sessions considered in this example are assumed to be member sessions of the same virtual group. For the purpose of demonstration, the estimated remaining session life time has been equalised to 120 s. The ranking list positions directly result from the value orders of the Downgrade Indexes D_i, which were calculated according to equation 7.2.

Table 7.3: Exemplary ranking list of sessions considered for downgrading

Session No.	Original codec	Available alternative codec providing improved E_{nC}	E_{nC} [MOS / (kbit/s)]	Gross Bitrate oC [kbit/s]	Gross Bitrate nC [kbit/s]	Delta BW [kbit/s]	t_s [s]	D_i [MOS·s]	Ranking List Position
1	G.711	iLBC 15.2	0.0993	90.4	41.6	48.8	120	581.38	**3**
2	G.711	GSM-EFR	0.1111	90.4	38.6	51.8	120	690.85	**2**
3	G.711	G.726-32	0.0726	90.4	58.4	32	120	278.79	**6**
4	G.711	G.723.1 ACELP	0.1651	90.4	22.9	67.5	120	1337.03	**1**
5	G.726-40	iLBC 15.2	0.0993	66.4	41.6	24.8	120	295.45	**5**
6	G.726-40	iLBC 13.33	0.1314	66.4	30.9	35.5	120	559.73	**4**
7	GSM (FR)	GSM-EFR	0.1111	39.4	38.6	0.8	120	10.67	**10**
8	GSM (FR)	GSM (HR)	0.1125	39.4	32	7.4	120	99.90	**9**
9	G.729	G.723.1 ACELP	0.1651	34.4	22.9	11.5	120	227.79	**7**
10	G.729	G.723.1 MP-MLQ	0.1646	34.4	24	10.4	120	205.40	**8**

7.3.2 Selection of sessions considered for cancellation/rejection

In general the availability of a telecommunication service can be constrained by two events, namely call blocking (that is, the rejection of a requested call by the service provider) and call dropping (that is, the intentional provider-driven termination of an existing session). Both are typically executed within the service infrastructure, triggered by a CAC functionality in case of a shortage of network resources.

Applying the framework for comprehensive QoS optimisation as defined within this research assumes that upon being requested and granted, every media session is assigned to a virtual group of media sessions which perceive similar QoS conditions (see section 5.1.1). In order to recover and maintain the network performance which affect all media sessions within a defined group, the cancellation of existing sessions or the rejection of requested sessions can become necessary (see section 7.2.2).

Recent research work imply that telecommunication subscribers experience stronger impairment about an existing call becoming dropped compared to the rejection of a requested call. Within the bonus/malus point system included in their resource management approach, (Yang *et al.*, 2005) propose that a call which is forcefully cancelled or dropped during handoff comes with a double loss of reward points for the provider compared to a rejected session request. Within the equation for the calculation of the Grade of Service experienced by telecommunication service subscribers, (AlQahtani and Mahmoud, 2006) assume a weighting factor of 10 for the impact of a dropped call compared to a rejection of a call request. This assumption has been adopted by (Sfairopoulou *et al.*, 2011).

Regarding the framework defined within this research project, in order to select communication sessions to be considered for cancellation/rejection, every media session is weighted according to its objective significance. Therefore the objective significance of a media session is defined through its individual role regarding the overall adherence to service availability limits according to the SLA between the concerned subscriber and the telecommunication service provider. Hence, this project proposes that a media session is considered most important if its termination resulted in the violation of service availability limits according to the applicable SLA. Therefore, in equation 7.5, a method is proposed to indicate the objective significance of a requested communication session. This calculation is mainly based on the relation of unsuccessful sessions over all sessions and comprises the following variables.

- S_r = Objective significance of a requested communication session; $S_r = [0\ldots1]$ ($S_r = 0$: Session has no relevance; $S_r = 1$: Session is highly important)
- C_{blA} = Total number of previous session initiation attempts of subscriber A which were denied (blocked) by the service provider (typically due to CAC decisions). From the perspective of subscriber A, only outgoing call attempts are considered
- C_{drA} = Total number of previous communication sessions in which subscriber A was involved (both, incoming and outgoing) and which were intentionally or accidently terminated (dropped) by the service provider
- C_{sA} = Total number of previous communication sessions in which subscriber A was involved (both, incoming and outgoing) and which have been completed successfully
- b_A = Service unavailability probability for subscriber A, according to the applicable SLA; $b_A = [0\ldots1]$. If a subscriber A is guaranteed a service availability of 99 percent, its unavailability probability is (100-99) percent = 1 percent → $b_A = 0.01$
- a = Auxiliary summand, for intentional increase of the objective significance of dedicated sessions (such as to force the successful initiation of emergency calls). $a = [0\ldots1]$

$$S_r = \frac{C_{bl\,A} + C_{drA}}{b_A(C_{blA} + C_{drA} + C_{sA})} + a \text{ ; } S_r = [0\ldots1] \qquad (7.5)$$

Note that within this project the objective significance of a recently requested communication session is assumed to depend on caller-related properties only. That is, neither call statistics nor service availability of the called party are considered.

In contrast, regarding the significance of an already established communication session, both calling (*A*) and called party (*B*) have to be considered. In this case, in a first step, the objective significance of the session is calculated independently for both parties *A* and *B*, considering their respective SLA and recorded call statistics. In the second step, the overall objective significance of the session is set to the significance value calculated for either *A* or *B*, whichever is higher. Equations 7.6 to 7.8 show the respective relations. For subscriber *B*, the variable names as defined above are used respectively with adapted indices. Additionally, the following variables are defined.

- S_A = Objective significance of an existing communication session, considering subscriber *A* perspective only. $S_A = [0\ldots1]$ ($S_A = 0$: Session has no relevance; $S_A = 1$: Session is highly important)
- S_B = Objective significance of an existing communication session, considering subscriber *B* perspective only. $S_B = [0\ldots1]$ ($S_B = 0$: Session has no relevance; $S_B = 1$: Session is highly important)
- S_e = Overall objective significance of an existing communication session, after considering the perspectives of both subscribers, *A* and *B*
- a_A = Auxiliary summand, for intentional increase of the objective significance of a dedicated session from the perspective of subscriber *A* (such as to force the successful initiation of emergency calls). $a_A = [0\ldots1]$

- a_B = Auxiliary summand, for intentional increase of the objective significance of a dedicated session from the perspective of subscriber B (such as to force the successful initiation of emergency calls). a_B = [0…1]

$$S_A = \frac{C_{bl\,A} + C_{drA}}{b_A(C_{blA} + C_{drA} + C_{sA})} + a_A \text{ ; } S_A = [0\ldots1] \tag{7.6}$$

$$S_B = \frac{C_{bl\,B} + C_{drB}}{b_B(C_{blB} + C_{drB} + C_{sB})} + a_B \text{ ; } S_B = [0\ldots1] \tag{7.7}$$

$$S_e = \begin{cases} S_A & \text{if } S_A > S_B \\ S_B & \text{else} \end{cases} \tag{7.8}$$

Objective session significance values are determined for both, requested and established communication sessions. For every virtual group of media sessions, respective communication sessions are arranged in a ranking list by their respective significance. The ranking list position of the respective communication session is considered in decisions regarding the blocking of a requested or the cancellation of an active session, respectively. If a number of sessions have to be cancelled or rejected to sustain optimised QoS conditions, these sessions are chosen which provide the lowest objective significance.

7.4 Evaluation of the introduced QoS optimisation approach

In order to evaluate the introduced QoS optimisation approach, a relevant scenario has been created and simulated by the use of the research prototype developed within this research project. Within the following subsections, both the scenario and the simulation are described, and the simulation results are analysed. All relevant files related to this evaluation are included on the CD-ROM enclosed with this thesis.

7.4.1 Evaluation scenario description

Within the introduced scenario it is assumed that the number of sessions in a particular virtual group increases gradually. Due to the resulting rise of traffic volume, the performance of the involved transport network paths deteriorates smoothly but continuously. The QoS optimisation framework is expected to detect this impending degradation of QoS conditions, and to take adequate action to compensate the effect of network performance impairment on the QoS perceived by the service users.

After the stabilisation of QoS conditititions, the considered virtual group is affected by a sudden and sharp decline of the network performance, resulting in a strong and continuing impairment of the QoS perceived by the concerned users. The cause for this disruption might be a failure of transport network elements, or due to any other event, such as the abrupt infusion of additional traffic not associated with the considered virtual group. Again, it is expected that the optimisation framework is able to detect the resulting impairment, and to choose and apply appropriate methods to recover QoS conditions experienced by the member sessions of the considered virtual group.

In either case it is expected that any action taken by the QoS controller results in the fast and efficient recovery of QoS conditions, providing a maximum amount of sessions with equitable QoS measure.

7.4.2 Evaluation scenario simulation

The evaluation scenario described in section 7.4.1 has been arranged for simulation by the use of the research prototype developed within this research project (see chapter 8). Figure 7.4 shows the deployed architecture.

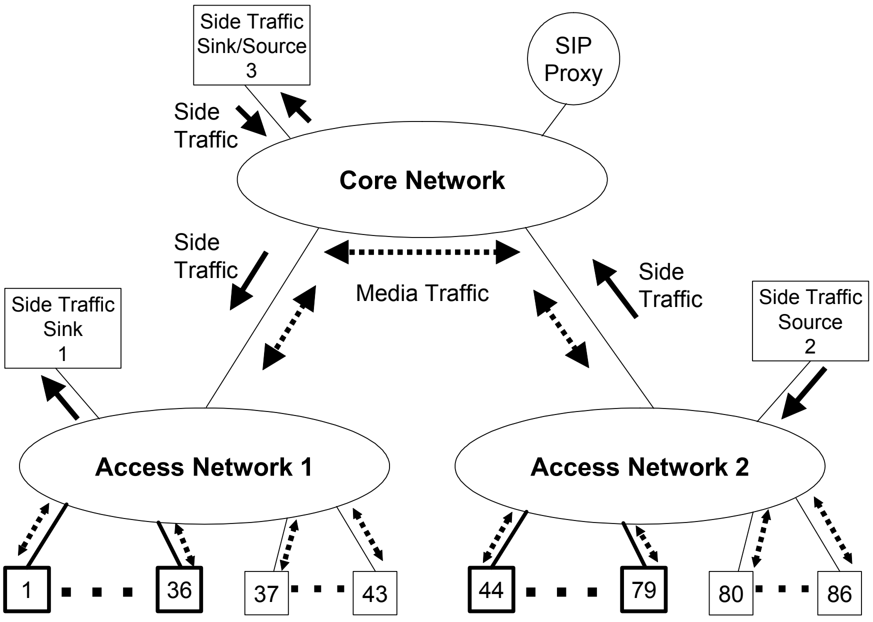


Figure 7.4: Evaluation scenario architecture

The following parameter settings were used.

- Two Access Networks (AN) are connected to the same Core Network (CN) via 3.5 Mbit/s duplex links.
- 43 user terminals are connected to each AN via 0.5 Mbit/s duplex links.
- The CN is provided with the combined Side Traffic Sink/Source 3 which receives and terminates side traffic from Side Traffic Source 2 connected to AN 2, and which independently generates and sends side traffic to Side Traffic Sink 1 connected to AN 1.

- Throughout the whole course of the simulation, both Side Traffic Sources independently generate side traffic according to the following distribution.
    - Exponential on/off
    - Average on / off times: each 10 ms
    - Traffic characteristic during on-times: packet size: 1500 Byte (constant), sending bitrate: 0.5 Mbit/s
- A SIP Proxy Server is connected to the CN.
- At the beginning of the scenario simulation, user terminals 1…36 successively initiate VoIP sessions based on the G.711 voice codec with user terminals 44…79 (note: only media streams received by user terminals 1…36 are assumed to be comprised by the virtual group whose QoS conditions are the subject matter of this evaluation). The first initiated session is considered as QoS reference session throughout the whole course of the simulation. For the simplification of the simulation scenario, this session is not considered for any QoS optimisation action.
- If required, sessions to be concerned by QoS optimisation action are chosen randomly from the available sessions.
- Beside the G.711 voice codec, every user terminal is assumed to support one of the following low-bitrate codecs. The availability of these codecs is assumed to be uniformly distributed among the user terminals.
    - G.726-40
    - GSM-EFR
    - G.729
    - iLBC 13.33
    - G.723.1 MP-MLQ
    - G.723.1 ACELP
- At a certain point in time of the course of the simulation, seven additional G.711 VoIP sessions are immediately initiated between user terminals 37…43 connected to AN 1 and user terminals 80…86 connected to AN 2.

These sessions are considered as additional side traffic and are not concerned for QoS optimisation action such as downgrading or cancellation.

7.4.3 Course of simulation and discussion

Table 7.4 chronologically lists all relevant events, observations, actions, and implications related to the evaluation simulation.

In addition to the simulation run involving the introduced QoS optimisation method, the same scenario was simulated without QoS optimisation. Figure 7.5 provides a graphical comparison of the resulting MOS characteristics for both simulation runs. It is clearly shown that the application of the proposed QoS optimisation effectively compensates the impact of unsteady network performance caused by increased network utilisation. This is not only depicted by means of the concrete MOS characteristic of the reference session (blue and pink curve) but also based on the average MOS over all sessions plotted as flat lines (dashed and chain line). The yellow and amber vertical marks indicate the trigger points for the first and the second QoS optimisation actions, respectively.

Table 7.4: Chronological course of the exemplary evaluation simulation

Time [s]	Event / Observation / Action	Implication
0.00… 0.40	Start of simulation, user terminals register with SIP Proxy/Registrar Server, activation of side traffic sources 2 and 3	
0.50… 10.00	Gradual initiation of 36 VoIP sessions from user terminals connected to AN 1 to user terminals connected to AN 2. All sessions assigned to the same virtual group. Total traffic volume from media sessions comprised by this group: 36*90.4kbit/s = 3254.4 kbit/s	
12.00	Running Mean MOS deviation limit of -0.3 MOS/Min exceeded (current value: 0.303 MOS/Min). Current Running Mean MOS: 4.358	Soft B/W reduction through codec downgradings required. Downgrade rate = 10% of current traffic = 325.44 kbit/s
12.00… 14.50	Execution of codec downgradings of 6 VoIP sessions. Total bandwidth saving: 325.6 kbit/s --> New total traffic volume = 2928.8 kbit/s In parallel to downgrade action: Control that deviation of Mean MOS does not exceed the limit for "hard" bandwidth reduction of -0.9 MOS/Min. If this is not the case, allow for 3 s of QoS recovery before deciding about further action.	
15.00	Re-evaluation of Running Mean MOS deviation. Running Mean MOS recovers with +0.245 MOS/Min. Current Running Mean MOS: 4.358	QoS recovery action successful. No further action required.
17.75	Running Mean MOS has increased to a value of MOS 4.368	
24.00… 24.60	Seven additional G.711 VoIP sessions are initiated immediately. These sessions are considered as additional side traffic, not as members of the concerned virtual group	
25.15	Running Mean MOS deviation limit of -0.3 MOS/Min exceeded (current value: 0.3712 MOS/Min). Current Running Mean MOS: 4.355	Soft B/W reduction through codec downgradings required. Downgrade rate = 10% of current traffic = 292.88 kbit/s
25.20	Initiating codec downgrading of 5 VoIP sessions. Planned total bandwidth saving: 301.6 kbit/s. In parallel to downgrade action: Control that deviation of Running Mean MOS does not exceed the limit for "hard" bandwidth reduction of -0.9 MOS/Min.	
25.24	Running Mean MOS deviation limit of -0.9 MOS/Min exceeded (current value: -1.043 MOS/Min). Current Running Mean MOS: 4.333. One session already downgraded to G.723.1 ACELP --> effective saving: 67.5 kbit/s. Current total traffic volume: (2928.8 - 67.5) kbit/s = 2861.3 kbit/s	Stop session downgradings and immediately release 10% of occupied B/W through cancellations (release of 286.13 kbit/s required).
25.30… 25.60	Cancellation of four G.711 VoIP sessions. Traffic saving: 361.6 kbit/s. Allow for 3 s of QoS recovery before deciding about further action.	
28.30	Re-evaluation of Running Mean MOS deviation. Running Mean MOS recovers with +2.085 MOS/Min. Current Running Mean MOS: 3.946	QoS recovery action successful. No further action required.
28.30… 40.16	Running Mean MOS continuously increases to a final value of 4.360. No further impact detected. End of simulation	
40.16	End of simulation	

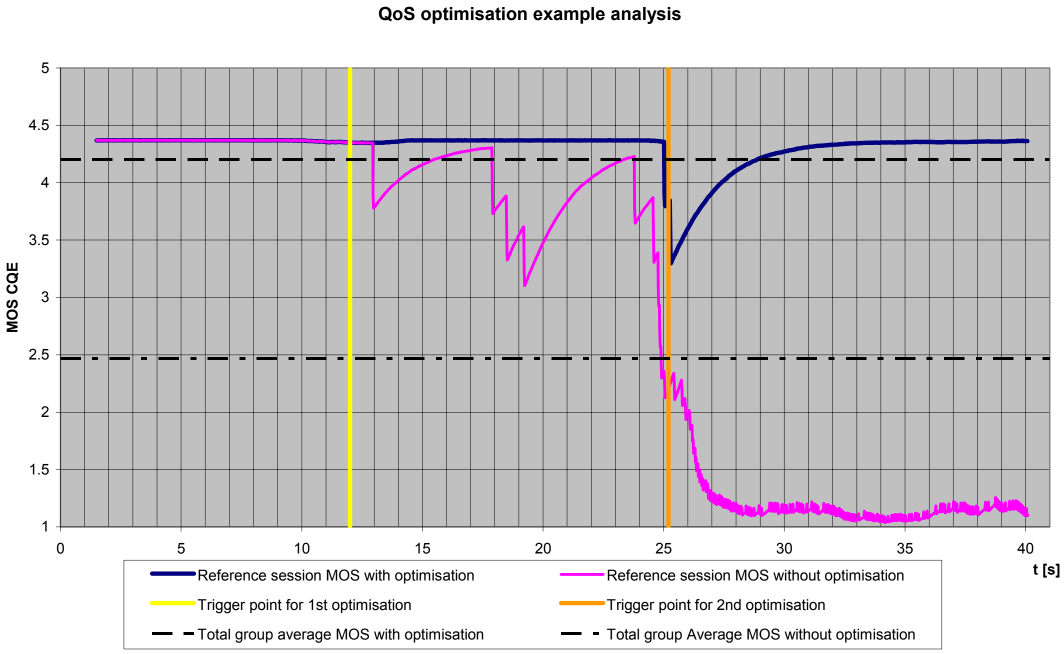


Figure 7.5: Graphical comparison of MOS evolution with and without optimisation

7.5 Summary

Within this chapter, the feedback control-based optimisation of QoS conditions has been described, as proposed within this project. The optimisation is performed through the passive control of the utilisation of concerned transport network paths through limiting the bitrate demands of media streams. Therefore, media sessions can be either modified to switch to low bitrate codecs (that is, being “downgraded”), or media sessions can be cancelled or rejected, respectively. Existing approaches for passive network utilisation control have been analysed, and fundamental issues were identified, which do not allow for utilisation within the proposed QoS optimisation framework.

For the determination and dimensioning of required QoS optimisation action, a feedback control system has been developed, whose control circuit has been described. The system is based on an extremum seeking controller, whose detailed functionality has been introduced. Several initial conditions of the system to be controlled have been distinguished, and threshold values have been defined which determine the range of action of the control circuit. Several action required for the optimisation of QoS conditions has been defined, which are all based on the downgrading or cancellation / rejection of media sessions. The effectiveness of the controller functionality has been successfully evaluated by the use of the research prototype. The respective simulation results have been introduced within the section preceding this summary. It is evident that the introduced QoS system allows for the early detection of emerging QoS degradation and, subsequently, is able to take appropriate action to anticipate further QoS deterioration and to recover QoS conditions for the considered sessions.

Since both the downgrading and the cancellation / rejection of sessions come with different characteristics each introducing potential benefits and disadvantages, sessions concerned by either action must be carefully chosen. In order to determine those sessions considered for downgrading, for every session, a downgrading index is computed, derived from the efficiency of the potential follow-up codec, the possible bitrate savings, and the remaining duration of the session. In contrast, regarding the cancellation or rejection of sessions, the objective significance of every session is determined, computed from the service availability guaranteed to the user according to the applicable SLA in conjuction with the actually experienced service availability of the user.

8 Research Prototype and framework evaluation

This chapter introduces the research prototype which has been used for proof-of-concept evaluation of several framework functionalities. The prototype consists of a collection of software components, which, as a whole, allow for the simulation of user-defined NGN communication scenarios. The scope and general layout of the prototype is discussed (section 8.1), followed by the description of its functional architecture and operation (section 8.2). Section 8.3 demonstrates the application of the research prototype for the evaluation of the proposed framework regarding its applicability for QoS optimisation. Section 8.4 focuses on a quantitative evaluation of the introduced framework itself, including comparisons with the standardised NGN QoS architecture regarding signalling effort, call setup delay and complexity coming along with QoS provision.

8.1 Scope and general layout of the research prototype

The research prototype has been designed with the objective to demonstrate the overall framework functionality based on the sequential execution of several steps which have to be performed to provide QoS optimisation. The sequential processing is the main difference to a potential real-world deployment of the developed framework in an NGN where most framework functions are required to run simultaneously and continuously, providing the overall framework with real-time functionality. However, the sequential prototype architecture has been favoured over a real-time capable software design for the following reasons.

- To allow for the separate adaptation, evaluation, and demonstration of the functionalities of several single framework components without any runtime interdependencies among each other.
- To facilitate the integrability of the framework with both, a recognised packet network simulator (ns-2) (University of Southern California, 2011) and a simulation environment for Artificial Neural Networks (SNNS) (University of Tübingen, 2008).
- To decrease the level of software complexity and hence, expected programming effort.

In spite of this fundamental difference to a real-world layout, the created research prototype can be used for the offline-demonstration of the overall framework functionality, including the interaction among the framework components. The collection of software that is required to deploy and run the research prototype is included on the CD-ROM enclosed with this thesis.

8.2 Functional architecture and operation of the research prototype

Since the prototype is designed for step-by-step processing, several tasks have to be considered sequentially. Figure 8.1 shows the schematic architecture of the research prototype. Within this figure, every framework step to be performed comes as a separate object. Objects including processes which have to be performed manually by the prototype operator are framed by thin contours. Objects which, once initiated, run autonomously are framed by thick contours. Object representing logical parts of the framework, which, however, are not required for the prototype functionality are provided with dashed line contours. Every object as depicted in Figure 8.1 consists of an upper and a lower part. While the upper parts display step numbers and a brief object description, the lower parts provide information about the outputs of the objects.

Within the following, the schematic prototype functionality is described, considering all of its tasks as depicted as objects in Figure 8.1.

1) Before the prototype can be applied, the considered NGN communication scenario has to be created. This is performed through the compilation of a network scenario description in a format which is destined for the use with *ns-2* (namely *tcl* format). In principle a scenario description includes specifications of the network architecture, nodes, links, traffic, timed events, and trace options. Note that the *ns-2* version integrated with this prototype has been upgraded with a third party SIP stack (Prior, 2007), which had to be further extended to allow for the full SIP functionality required for the integration with the proposed framework.

   In addition to the *tcl* file, a text file named *classreadout.fw* has to be created, including a list of all media sessions to be considered for QoS optimisation. This separate list is required to avoid interdependencies between the network scenario description provided to *ns-2* and individual decisions regarding the consideration of sessions for QoS optimisation procedures.

2) Once the NGN communication scenario has been described as a *tcl* file, a first simulation run has to be performed, utilising the adapted version of *ns-2*. As simulation output, a trace file named *out1.tr* is generated, including all data required for the subsequent analysis of the network performance affecting media sessions within the given communication scenario.

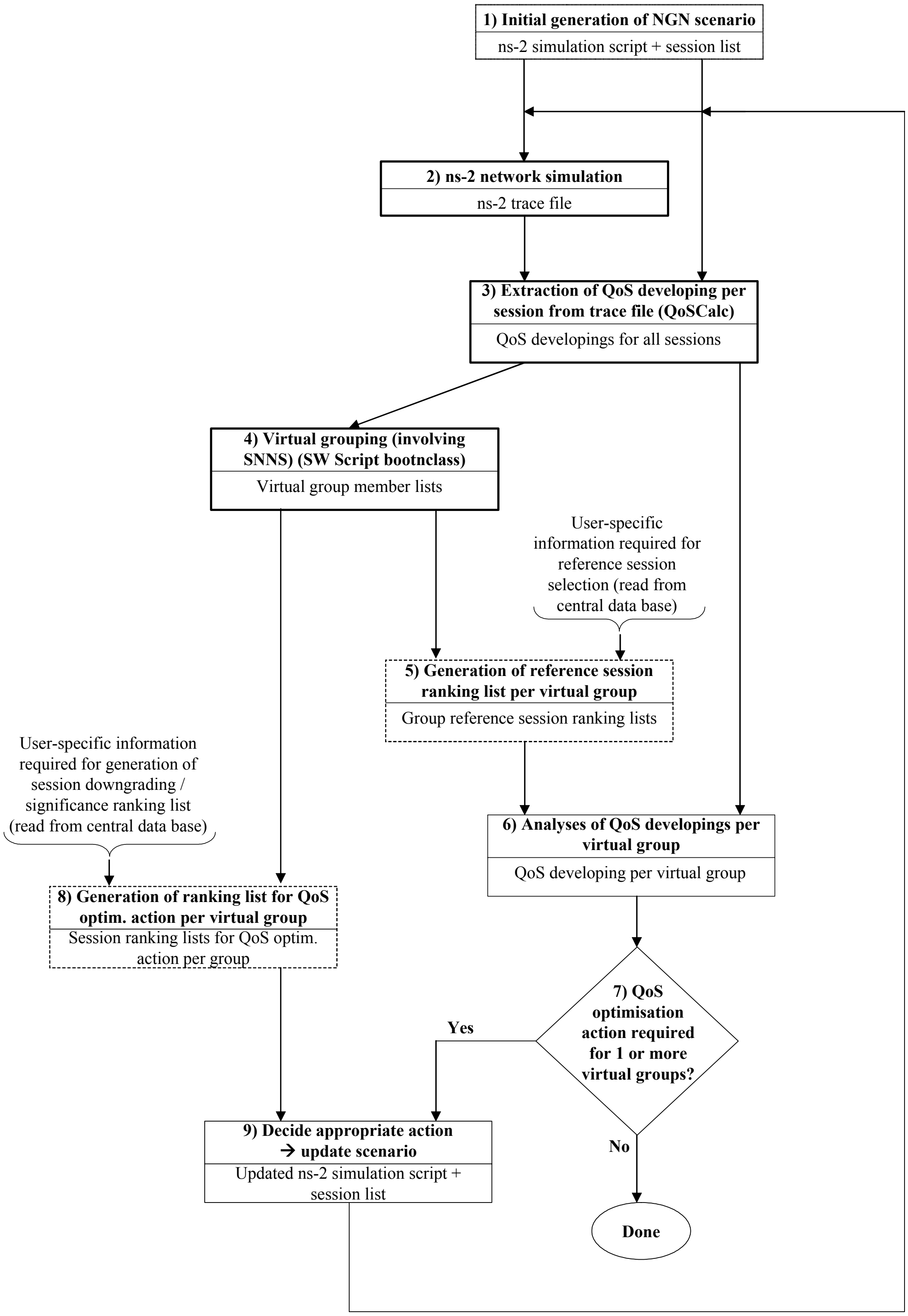

1) Initial generation of NGN scenario
ns-2 simulation script + session list
2) ns-2 network simulation
ns-2 trace file
3) Extraction of QoS developing per session from trace file (QoSCalc)
QoS developings for all sessions
4) Virtual grouping (involving SNNS) (SW Script bootnclass)
Virtual group member lists
User-specific information required for reference session selection (read from central data base)
5) Generation of reference session ranking list per virtual group
Group reference session ranking lists
User-specific information required for generation of session downgrading / significance ranking list (read from central data base)
6) Analyses of QoS developings per virtual group
QoS developing per virtual group
8) Generation of ranking list for QoS optim. action per virtual group
Session ranking lists for QoS optim. action per group
7) QoS optimisation action required for 1 or more virtual groups?
Yes
No
9) Decide appropriate action → update scenario
Updated ns-2 simulation script + session list
Done

Figure 8.1: Schematic functionality architecture of the research prototype

3) For all media sessions listed in *classreadout.fw*, the per-packet network performance characteristics (delay, jitter, packet loss) are extracted from the *ns-2* trace file *out1.tr*. In a subsequent step, the resulting QoS characteristics are derived, coming as MOS value sequences. In addition, average values are computed for every session. These steps are performed automatically by a software script called *qoscalc* which has been coded within this research project.

4) Reusing the jitter characteristics as obtained from step 3), all sessions are now virtually grouped by the similarity of their perceived network performance. Therefore, jitter value patterns are generated by a software script called *trace2patnoref* which has been written inline with this research project. These patterns are subsequently used as an input to an Artificial Neural Network (ANN) of the type ART 2, which is provided by a Neural Network simulator called *SNNS* (Stuttgart Neural Network Simulator). Since multiple classification runs have to be performed by the ANN throughout one virtual grouping process, a middleware called *bootnclass* has been developed within this research project, providing the interworking with *SNNS*. *bootnclass* sequentially presents collections of jitter patterns to the ANN, reads and analyses the results of the separate classification runs and derives the resulting assignments of media sessions to virtual groups. Both, the bootstrap process and the accuracy self-scaling process described in sections 5.3 and 5.4 are performed within this step, managed by the *bootnclass* software. After the completion of the virtual grouping process, a result file called *prot.txt* is generated, including lists of the member sessions of all recognised virtual groups.

5) In order to determine those sessions which are best applicable as QoS reference sessions for their respective groups, continuously updating ranking list have to be generated. These lists are based on the list showing the assignments of sessions to virtual groups obtained in step 4). For the determination of the qualification of a session as a group QoS reference, the algorithm introduced in section 6.1 is used. However, since the outcome of

this process has no relevant effect on the optimisation results obtained from the prototype application, this step has not been implemented. For the further processing it is assumed that per virtual group, at least one QoS reference session is chosen randomly from all member sessions by the prototype operator.

6) The QoS perceived by the receiver of the reference media session represents the QoS as experienced by any receiver of any member session of the respective virtual group. In order to evaluate the QoS provided to a respective group, the QoS evolution of the reference sessions are analysed. Therefore the algorithms described in section 7.2 are applied.

7) Based on the results of the analyses performed in step 6), it is decided whether the QoS conditions affecting one or more virtual groups have to be optimised. If this is not the case, the given scenario did not require any further QoS optimisation. In case that a need for optimisation is detected, the execution of step 8) is required logically before decisions can be made regarding the applicable optimisation action.

8) Within this step, continuously updated ranking lists are generated, providing information which sessions are best applicable to be concerned by QoS optimisation action. Concretely, two ranking lists are provided, one providing priorities of sessions to be considered for codec downgradings, and one which provides the ranking for sessions potentially considered for session cancellations or rejections, based on their objective significance. Both ranking lists are generated by the use of the algorithms introduced in section 7.3. However, since the outcome of this process has no relevant effect on the optimisation results obtained from the prototype application, this step has not been implemented. For the further processing it is assumed that sessions considered for downgrading or cancellation/rejection are chosen randomly from all group member sessions by the prototype operator.

9) Within the final step, decisions are made regarding the action to be taken in order to improve or recover the QoS provided to the member sessions of the concerned virtual groups. If required, session downgradings or cancellations

are concluded in the required dimension as determined by the algorithm introduced in section 7.2. The respective modifications are added to the *ns-2* simulation script and to the session list, which together describe the scenario. Subsequently, the updated scenario is again simulated with *ns-2*, and the follow-up steps are processed again as described above.

By performing several optimisation cycles, the NGN communication scenario is successively adjusted regarding the QoS provided to the included media sessions. This simulates the functionality of the framework for QoS optimisation developed within this research work.

8.3 Prototype-based framework evaluation

Within previous chapters, the functions provided by the research prototype were utilised separately to evaluate the novel concepts coming with the introduced framework (see sections 5.4.3 and 7.4). In order to demonstrate both the main framework and prototype functionalities as a whole a complex test scenario has been defined, and simulated both with and without the application of the QoS optimisation framework. The following subsections introduce the test scenario and its simulation with and without QoS optimisation, and discuss the simulation results. All relevant files related to this evaluation are included on the CD-ROM enclosed with this thesis.

8.3.1 Test scenario

Figure 8.2 provides an overview of the considered test scenario.

Within the scenario an NGN has been set up, consisting of a core network, to which three access networks (AN 1…3) are connected. Every AN is provided with a number of user terminal, and a SIP service control infrastructure is connected to the core network for signalling control.

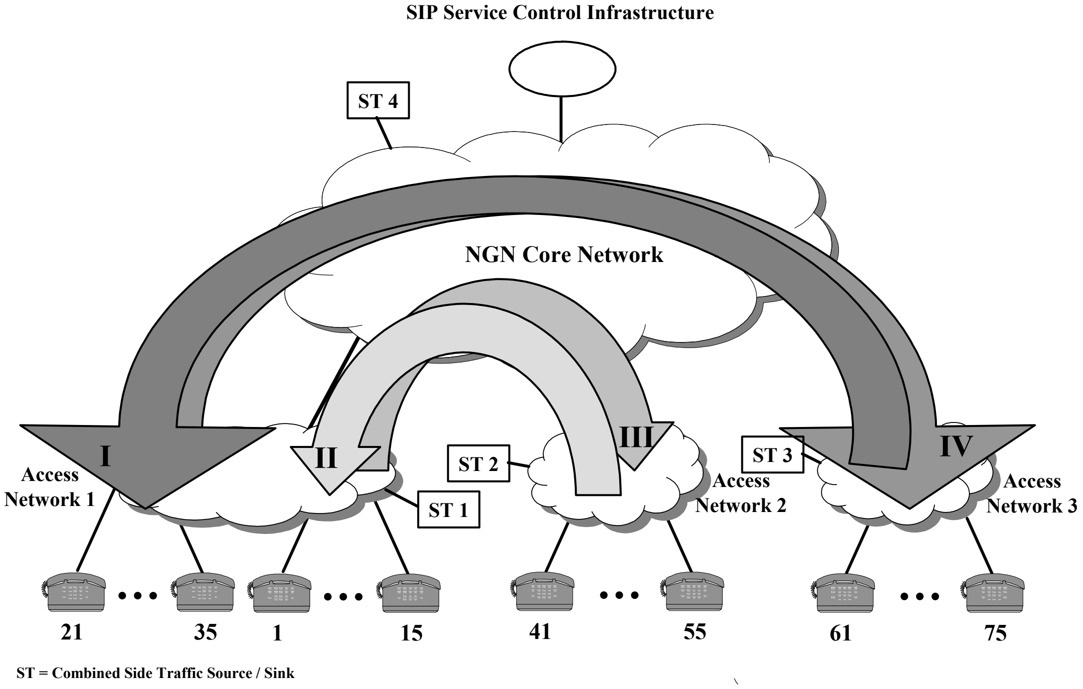


Figure 8.2: Test scenario for framework evaluation

Furthermore, combined side traffic sources and sinks (ST) are connected to every AN as well as to the NGN Core Network. All STs connected to ANs exchange bidirectional packet traffic with the central ST connected to the core network. Hence sending out IP packets with different statistical frequency distributions, the ST sources generate basic traffic loads between the core network and several ANs. This is required to induce statistical network performance variations within the network simulation environment, resulting in a more realistic network behaviour. Therefore, both network link and ST parameters were empirically chosen to adopt realistic network performance characteristics as observed in typical fixed line IP communication scenarios as published in (Abu Salah *et al.*, 2008). Note that some of the given parameters are varied among different network links and STs in order to establish a variety of general network performance conditions for several media streams.

The following setup was provided.

- Access Network 1 (AN1)
    - Connected to CN via a 3.5 Mbit/s duplex link
    - ST 1 generates and sends side traffic to ST4 according to the following distribution
        - Exponential on/off
        - Average on-time: 10 ms
        - Average off-time: 20 ms
        - Packet size: 1500 Byte
        - Sending bitrate (on-time only): 1.5 Mbit/s
    - ST 1 receives side traffic from ST4 according to the following distribution
        - Exponential on/off
        - Average on-time: 20 ms
        - Average off-time: 10 ms
        - Packet size: 1500 Byte
        - Sending bitrate (on-time only): 1.5 Mbit/s
    - 30 user terminals connected (numbers 1…15 and 21…35) via 0.5 Mbit/s duplex links
    - At different points in time user terminals 1…15 initiate SIP sessions for bidirectional G.711 audio media exchange with user terminals connected to AN 2
    - At different points in time user terminals 21…35 initiate SIP sessions for bidirectional G.711 audio media exchange with user terminals connected to AN 3
- Access Network 2 (AN 2)
    - Connected to CN via a 2.0 Mbit/s duplex link

  - ST 2 generates and sends side traffic to ST4 according to the following distribution
    - Exponential on/off
    - Average on-time: 10 ms
    - Average off-time: 10 ms
    - Packet size: 1500 Byte
    - Sending bitrate (on-time only): 1.5 Mbit/s
  - ST 2 receives side traffic from ST4 according to the following distribution
    - Exponential on/off
    - Average on-time: 10 ms
    - Average off-time: 30 ms
    - Packet size: 1500 Byte
    - Sending bitrate (on-time only): 1.5 Mbit/s
  - 15 user terminals connected (numbers 41…55) via 0.5 Mbit/s duplex links, accepting sessions initiated by user terminals connected to AN 1

- Access Network 3 (AN 3)
  - Connected to CN via a 2.0 Mbit/s duplex link
  - ST 3 generates and sends side traffic to ST4 according to the following distribution
    - Exponential on/off
    - Average on-time: 10 ms
    - Average off-time: 10 ms
    - Packet size: 1500 Byte
    - Sending bitrate (on-time only): 1.5 Mbit/s
  - ST 3 receives side traffic from ST4 according to the following distribution
    - Exponential on/off
    - Average on-time: 30 ms
    - Average off-time: 10 ms
    - Packet size: 1500 Byte
    - Sending bitrate (on-time only): 1.5 Mbit/s

  - 15 user terminals connected (numbers 61…75) via 0.5 Mbit/s duplex links, accepting sessions initiated by user terminals connected to AN 1

According to this setup, four media stream directions can be distinguished, which are shown in Figure X.Y as raised arrows, numbered I…IV.

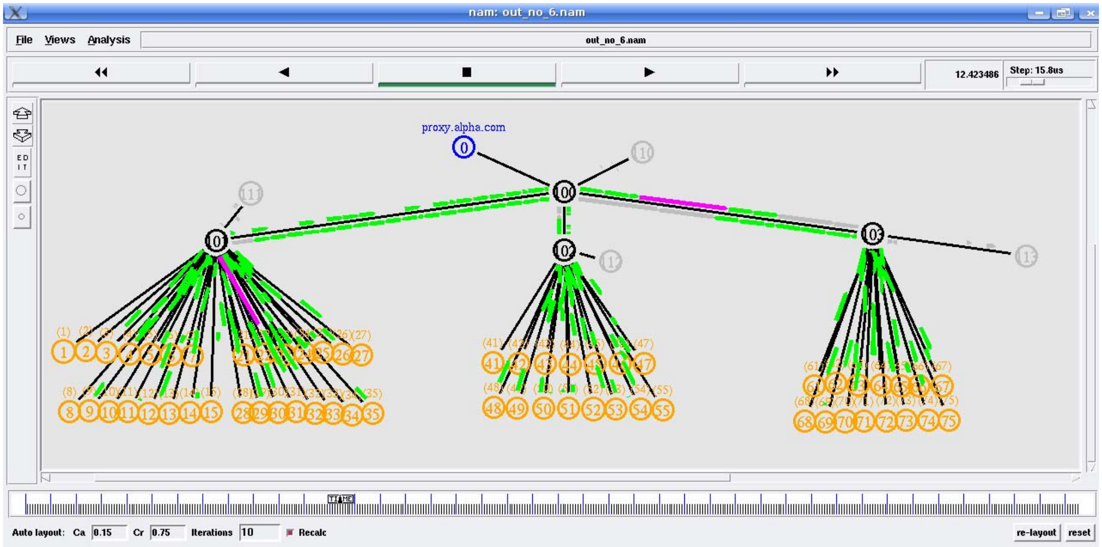
- Number I: Media streams sent from user terminals connected to AN 3 to user terminals connected to AN 1
- Number II: Media streams sent from user terminals connected to AN 2 to user terminals connected to AN 1
- Number III: Media streams sent from user terminals connected to AN 1 to user terminals connected to AN 2
- Number IV: Media streams sent from user terminals connected to AN 1 to user terminals connected to AN 3

Beside G.711, all user terminals are assumed to support one alternative audio codec (G.726-40, GSM-EFR, or G.729).

8.3.2 Course of simulation and discussion

The introduced test scenario was modelled as a ns-2 tcl description file to be simulated by the use of the ns-2 network simulator. Subsequently the concept of the QoS optimisation framework was applied, performing the steps described in section 8.2. User terminals were virtually grouped by applying QoS profiling (see chapter 5), and QoS conditions were determined for all groups. Based on the optimisation rules defined in section 7.2 QoS conditions were optimised for each group. The optimisation cycle was repeated several times, resulting in the step-by-step adaption of the tcl file, until optimal QoS conditions were determined.

Figure 8.3 shows a screenshot of the graphical representation of the ns-2 simulation course. Access and core networks are shown as black-framed circles, while SIP user terminals, which are connected to the access networks, are represented by amber-framed circles. Side traffic sinks/sources are displayed as grey-framed circles. The SIP proxy server comes as a blue-framed circle, connected to the node that represents the core network.



Auto layout: Ca 0.15 Cr 0.75 Iterations 10 Recalc

Figure 8.3: Screenshot of graphical representation of test scenario simulation

On top and underneath the black lines which represent physical connections between nodes, traffic is indicated through arrow bars of different sizes and colours, each representing one packet. Green bars represent RTP media traffic, SIP signalling messages are displayed as magenta-coloured bars, and side traffic is shown as grey bars.

QoS profiling

The QoS profiling of all media sessions was performed periodically at six points in time throughout the course of the simulation.

As already stated in section 8.3.1 four directions of media streams (I, II, III, and IV) could be distinguished. Throughout the simulation it appeared that considerable similarities existed in the QoS characteristics of media streams sent in directions I (AN3→AN1) and IV (AN1→AN3). This can be explained by the combined influence of the following factors.

- Although being transmitted in opposite direction, those media directions share exactly the same paths through the network (path between AN1 and AN3).
- The side traffic that affects the transmission on the considered network path show comparable characteristics in both transmission directions, especially regarding on and off time intervals.

For this reason if applicable, media stream directions I and IV were jointly considered within the classification process.

At a later point in time of the simulation course considerable similarities were also observed in the QoS characteristics of media stream directions III (AN1→AN2) and IV (AN1→AN3). This can be explained by a temporary statistically stronger impact of the portion of the network path which is shared between media streams in these two directions (namely, the processing in AN1, and the transmission path to the core network). For this reason, if applicable, media stream directions III and IV were jointly considered in the fifth and sixth profiling run.

Table 8.1 to Table 8.4 show the achieved grouping results. The average grouping accuracies ranged from 86.16 to 98.33 percent with an overall average of 91.44 percent.

Table 8.1: Grouping results for media streams in direction I

Time of profiling run [s]	3.5	10.5	17.5	24.5	31.5	38.5	Grouping accuracy total average
Direction I (*)							
No. of active media streams	10	13	13	13	13	15	
No. of media streams assigned to a virtual group associated with this stream direction	10(*)	13 (*)	13(*)	13 (*)	8(*)	12(*)	
No. of media streams assigned to a virtual group associated with unrelated stream direction	0	0	0	0	5	3	
Grouping accuracy [%]	100	100	100	100	61.54	80	**90.26**

(*) If indicated, stream directions I and IV were jointly considered

Table 8.2: Grouping results for media streams in direction II

Time of profiling run [s]	3.5	10.5	17.5	24.5	31.5	38.5	Grouping accuracy total average
Direction II							
No. of active media streams	10	13	13	13	13	15	
No. of media streams assigned to one associated virtual group	9	13	13	13	13	15	
No. of media streams assigned to a virtual group associated with unrelated stream direction	1	0	0	0	0	0	
Grouping accuracy [%]	90	100	100	100	100	100	**98.33**

Table 8.3: Grouping results for media streams in direction III

Time of profiling run [s]	3.5	10.5	17.5	24.5	31.5	38.5	Grouping accuracy total average
Direction III (**)							
No. of active media streams	10	13	13	13	13	15	
No. of media streams assigned to a virtual group associated with this stream direction	9	12	11	12	13(**)	13(**)	
No. of media streams assigned to a virtual group associated with unrelated stream direction	1	1	2	1	0	2	
Grouping accuracy [%]	90	92.31	84.62	92.31	100	86.67	**90.99**

(**) If indicated, stream directions III and IV were jointly considered

Table 8.4: Grouping results for media streams in direction IV

Time of profiling run [s]	3.5	10.5	17.5	24.5	31.5	38.5	**Grouping accuracy total average**
Direction IV (*)							
No. of active media streams	10	13	13	13	13	15	
No. of media streams assigned to a virtual group associated with this stream direction	8(*)	12(*)	11(*)	13(*)	13(*) (**)	9(*) (**)	
No. of media streams assigned to a virtual group associated with unrelated stream direction	2	1	2	0	0	6	
Grouping accuracy [%]	80	92.31	84.62	100	100	60	**86.16**

(*) If indicated, stream directions I and IV were jointly considered

(**) If indicated, stream directions III and IV were jointly considered

QoS optimisation

QoS conditions of all sessions involved in the communication scenario were monitored throughout the course of the simulation, and reference sessions were determined for all media stream directions. Based on the QoS characteristics observed from the reference sessions, QoS optimisation action was performed. Table 8.5 to Table 8.7 provide the chronological course of QoS evolution and resulting optimisation action. The left column of the tables provides either consecutive time points or time periods (in seconds), starting at 0.00 at the beginning of the simulation. In the middle column, any occuring event, observation, or applied action is listed. Note that any new row of the tables refers to a further event, observation, or action. In the right column, if applicable, the implication derived from the corresponding event or observation is given.

Table 8.5: Chronological course of the QoS optimisation test scenario part a)

Time [s]	Event / Observation / Action	Implication
0.00…0.40	Start of simulation, user terminals register with SIP Proxy/Registrar Server, activation of side traffic sources ST 1, ST 2, ST 3, and ST 4	
0.50…7.50	Initiation of 14 G.711 VoIP sessions between user terminals connected to AN 1 and AN 2, and initiation of 14 VoIP sessions between user terminals connected to AN 1 and AN 3. Media stream direction AN 3 --> AN 1: direction I. AN 2 --> AN 1: direction II. AN 1 --> AN 2: direction III. AN 1 --> AN 3: direction IV. Total traffic volume per direction: 14*90.4kbit/s = 1265.6 kbit/s	
7.60	Direction IV: Running Mean MOS deviation limit of -0.3 MOS/Min exceeded.	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 126.6 kbit/s
7.70	Direction IV: Running Mean MOS deviation limit of -0.9 MOS/Min exceeded.	Immediately release 10% of occupied bandwidth through cancellations (release of 126.6 kbit/s required --> 2 G.711 sessions).
8.00	Cancellation of 2 G.711 VoIP sessions between AN 1 and AN 3. Traffic saving: 180.8 kbit/s (required: 126.6 kbit/s --> 54.2 kbit/s surplus bitrate release).	
8.00	1 further session requested between AN 1 and AN 3. Granted under downgrade conditions (Codec G.729 --> additional traffic: 34.4 kbit/s)	
8.40	Direction II: Running Mean MOS deviation limit of -0.3 MOS/Min exceeded.	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 126.6 kbit/s
8.50	1 further session requested between AN 1 and AN 2. Granted under downgrade conditions (Codec G.729 --> additional traffic: 34.4 kbit/s)	
8.63	Direction II: Running Mean MOS deviation limit of -0.9 MOS/Min exceeded.	Immediately release 10% of occupied bandwidth through cancellations (release of 161.0 kbit/s required --> 2 G.711 sessions).
8.93	Cancellation of 2 G.711 VoIP sessions between AN 1 and AN 2. Traffic saving: 180.8 kbit/s (required: 161.0 kbit/s --> 19.8 kbit/s surplus bitrate release).	
9.00	1 further session requested between AN 1 and AN 3. Granted under downgrade conditions (Codec GSM-EFR --> additional traffic: 38.6 kbit/s)	

Table 8.6: Chronological course of the QoS optimisation test scenario part b)

Time [s]	Event / Observation / Action	Implication
12.00	Direction IV: Revision of MOS evolution after session cancellation. Still QoS deterioration exceeding -0.9 MOS/Min. Current total bitrate: 1067.2 kbit/s	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 106.7 kbit/s
12.30…12.50	Downgrade 3 G.711 session between AN 1 and AN 3 to codecs G.726-40, GSM-EFR, and G.729 --> total saving: 131.8 kbit/s	
12.94	Direction II: Revision of MOS evolution after session cancellation. Running Mean MOS deviation recovered successfully. No further action required.	
15.00	Direction IV: Revision of MOS evolution after downgrading. Running Mean MOS deviation recovered successfully. No further action required.	
16.81	Direction IV: Running Mean MOS deviation limit of -0.3 MOS/Min exceeded. Total occupied bitrate: 935.6 kbit/s	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 93.6 kbit/s
16.85	Direction IV: Running Mean MOS deviation limit of -0.9 MOS/Min exceeded.	No additional consequence, since session cancellations have already been performed for this media direction.
17.10…17.30	Downgrade 3 G.711 session between AN 1 and AN 3 to codecs G.726-40, GSM-EFR, and G.729 --> total saving: 131.8 kbit/s	
21.00	Direction IV: Revision of MOS evolution after downgrading. Running Mean MOS deviation recovered successfully. No further action required.	
26.74	Direction II: Running Mean MOS deviation limit of -0.3 MOS/Min exceeded. Total occupied bitrate: 1119.2 kbit/s	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 111.9 kbit/s
26.79	Direction II: Running Mean MOS deviation limit of -0.9 MOS/Min exceeded.	No additional consequence, since session cancellations have already been performed for this media direction.
27.10…27.30	Downgrade 3 G.711 session between AN 1 and AN 2 to codecs G.726-40, GSM-EFR, and G.729 --> total saving: 131.8 kbit/s	
30.00…32.00	2 further session requested between AN 1 and AN 2. Granted with G.711 codec (2x90.4 kbit/s)	

Table 8.7: Chronological course of the QoS optimisation test scenario part c)

Time [s]	Event / Observation / Action	Implication
30.10	Direction II: Revision of MOS evolution after downgrading. Running Mean MOS deviation recovered successfully. No further action required.	
34.00…35.00	2 further session requested between AN 1 and AN 3. Granted with G.711 codec (2x90.4 kbit/s)	
35.81	Direction II: Running Mean MOS deviation limit of -0.3 MOS/Min exceeded. Total occupied bitrate: 1168.2 kbit/s	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 116.8 kbit/s
35.85	Direction II: Running Mean MOS deviation limit of -0.9 MOS/Min exceeded.	No additional consequence, since session cancellations have already been performed for this media direction.
36.20…36.40	Downgrade 3 G.711 session between AN 1 and AN 2 to codecs G.726-40, GSM-EFR, and G.729 --> total saving: 131.8 kbit/s	
37.12	Direction IV: Running Mean MOS deviation limit of -0.3 MOS/Min exceeded. Total occupied bitrate: 984.6 kbit/s	Prepare for soft bandwidth reduction through codec downgradings. Downgrade rate = 10% of current total traffic volume = 98.5 kbit/s
37.20	Direction IV: Running Mean MOS deviation limit of -0.9 MOS/Min exceeded.	No additional consequence, since session cancellations have already been performed for this media direction.
37.50…37.70	Downgrade 3 G.711 session between AN 1 and AN 3 to codecs G.726-40, GSM-EFR, and G.729 --> total saving: 131.8 kbit/s	
39.20	Direction II: Revision of MOS evolution after downgrading. Running Mean MOS deviation recovered successfully. No further action required.	
40.50	End of simulation	

Figure 8.4 and Figure 8.5 demonstrate the effect of the performed QoS optimisation action, comparing the QoS characteristics of the reference sessions of all four media stream directions with and without the application of the QoS optimisation framework.

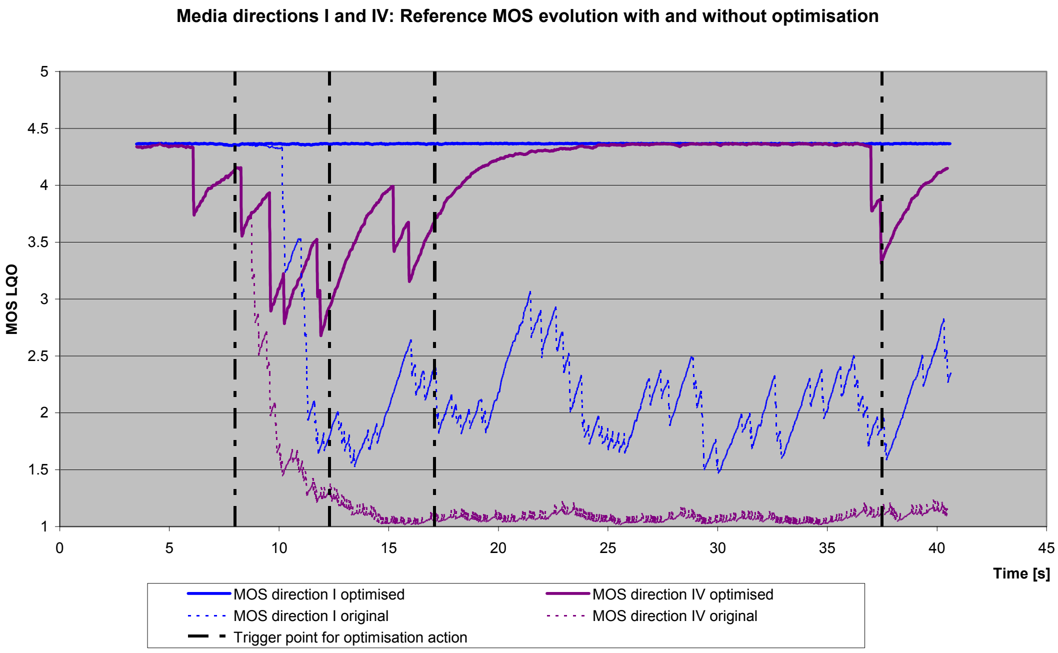


Figure 8.4: MOS evolution of media directions I and IV with and without QoS optimisation

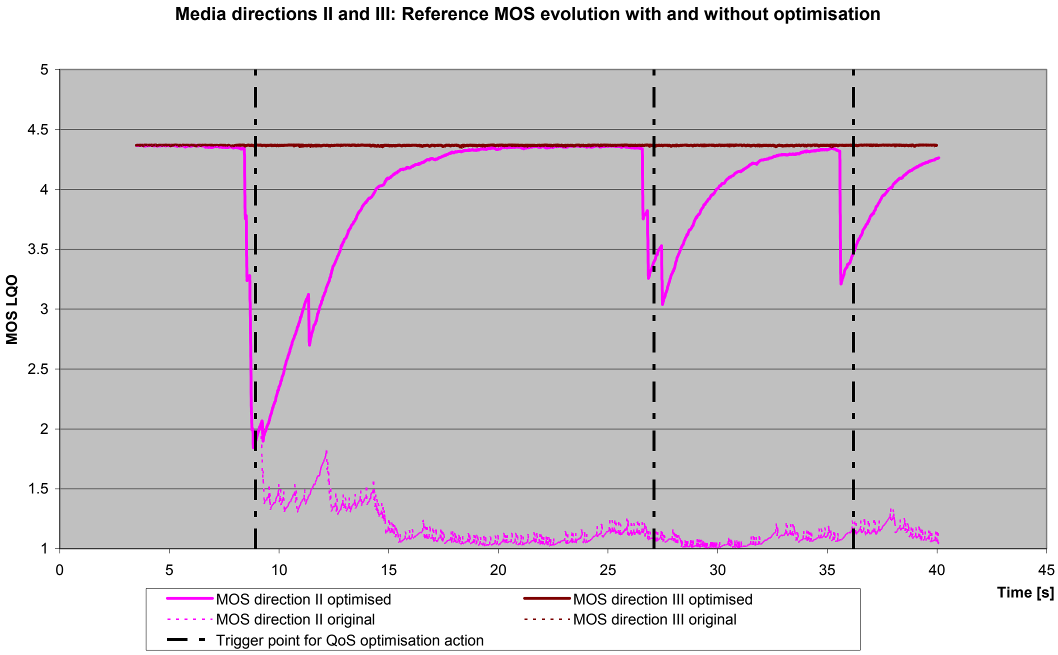


Figure 8.5: MOS evolution of media directions II and III with and without QoS optimisation

It is evident that the MOS evolutions of media stream directions I, II, and IV, which are affected by multiple network overload incidents, resulting in the severe collapse of QoS conditions, recover recurrently once QoS optimisation actions take effect. From Figure 8.4 it is observable that the procedures applied to recover QoS conditions for direction IV do not only result in the regeneration of direction IV QoS, but also prevent the breakdown of QoS conditions in direction I, which would occur without the application of QoS optimisation actions.

Table 8.8 summarises the achieved QoS improvement over all media sessions in this test scenario through the application of the proposed optimisation framework. It is evident that substantial MOS increase (66.40 percent) was especially achieved for media streams sent in direction IV, which was concerned by severe QoS limitations. Also media streams that were sent in directions I and II experience considerable QoS improvement (55.09 and 53.19 percent, respectively).

Table 8.8: Achieved QoS improvement in test scenario

Media stream direction	Mean MOS without optimisation	Mean MOS with optimisation	MOS increase [%]
I	1.94	4.32	**55.09**
II	1.98	4.23	**53.19**
III	4.36	4.33	**-0.69**
IV	1.25	3.72	**66.40**
Overall	2.38	4.15	**42.59**

The average MOS provided to media stream direction III appears to be lightly reduced by 0.69 percent if QoS optimisation is applied. This can be explained by the fact that media streams that were sent in this direction experience excellent QoS conditions both with and without QoS optimisation. Due to required QoS optimisation action for media streams sent in direction II (AN 2 → AN 1), also direction III (AN 1 → AN 2) was concerned by both downgrade and cancellation action, which resulted in the observable minimal MOS decrease. However, with an average MOS of 4.33, media stream sent in direction III are still provided with excellent QoS conditions. Counted over all media streams sent in any direction, a total MOS increase of 42.59 percent was achieved through the application of the proposed QoS optimisation framework.

8.4 Quantitative framework evaluation

This section focuses on a quantitative comparison of the framework for QoS optimisation with the standard NGN QoS architecture. The comparison is based on calculations, comprising both the signalling burden and the call set up delay required for QoS provision. Furthermore, levels of complexity of both approaches are compared in table form.

The quantitative comparison is based on the general network structure of the test scenario intruded in section 8.3.1, consisting of an NGN Core Network and three Access Networks (see Figure 8.6). To allow for varying numbers of parallel sessions and traffic conditions, the numbers of connected user terminals are variable, and network link-related parameters such as bandwidth limits are not predefined. Considered communication directions between subscribers connected to different Access Networks are indicated as dashed arrow lines.

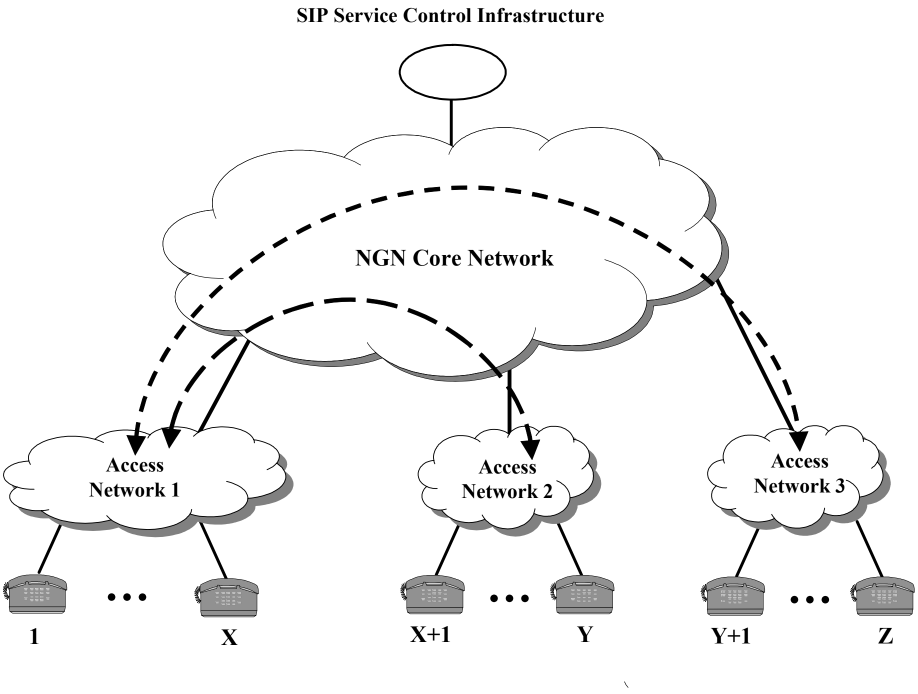


Figure 8.6: General network structure for quantitative framework evaluation

Regarding the following comparisons, for the standardised NGN QoS architecture, the reference layout introduced in section 3.3.3 is considered. In order to adapt to the general network structure shown in Figure 8.6, the third Access Network is assumed to provide the same characteristics as Access Network 2. Figure 8.7 shows the reference layout.

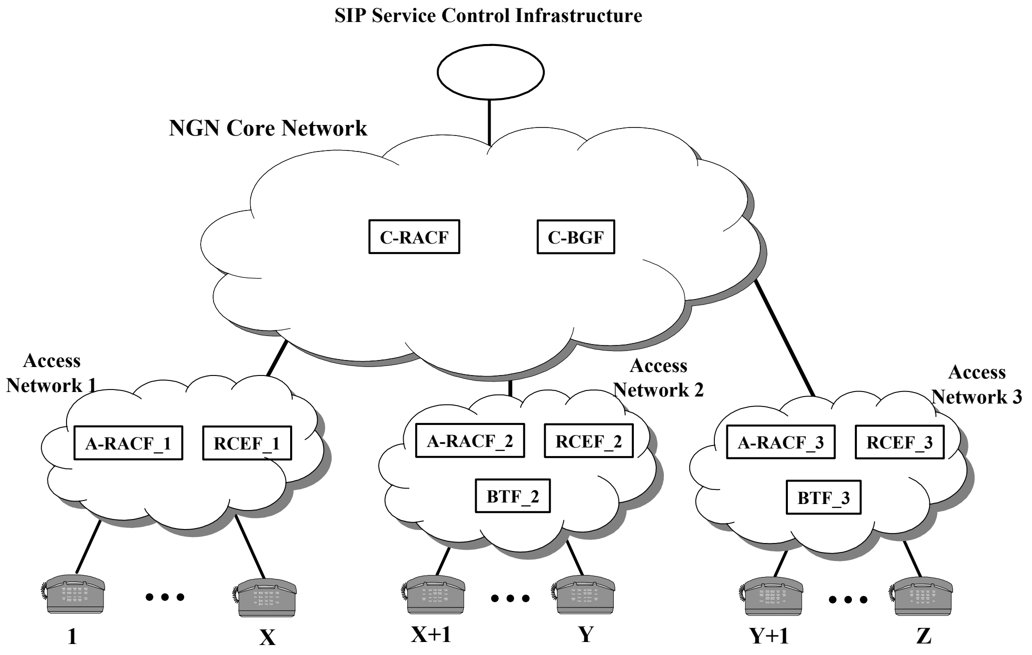


Figure 8.7: Layout of reference NGN QoS architecture for comparison

8.4.1 Signalling effort for QoS provision

This section discusses the signalling burden coming along with the application of the proposed QoS optimisation framework. Therefore, the signalling sequences as introduced in sections 4.4 to 4.6 are considered. To allow for comparability with the reference signalling flow given for standardised NGN QoS provision approach, only those messages are considered which pass the NGN Core Network. It is assumed that the QoS Manager coming with the proposed framework is embedded within the central SIP service control infrastructure, and as such is provided with a direct link to data bases and to the SIP Call Server. Hence, information exchanged between the QoS Manager and any other component of the service control infrastructure is not considered. Table 8.9 provides details on the signalling burden coming along with the application of the proposed QoS optimisation framework, based on the signalling flows introduced in sections 4.4 to 4.6. Regarding the lengths of the exchanged messages, the following assumptions are made.

- SIP message without message body: 100 Byte (assuming compact SIP header form according to (IETF RFC 3261, 2002))
- SIP message body including SDP: 200 Byte
- SIP message body including specific network performance data: 104 Byte (initial RTP timestamp: 32 bit; subsequent 50 timestamps: 16 bit each)

Table 8.9: Framework-related QoS optimisation signalling effort

Framework-specific action	Signalling effort	As described in thesis section
Initial QoS profiling and CAC	16.064 kbit per SIP session initiation attempt	4.4
Recurrent mid-session verification of group affiliations	2.432 kbit per re-verification procedure	4.5.1
Subscription to continuous network performance information	1.6 kbit per subscription action	4.5.2
Continuous transmission of network performance information	2.432 kbit/s per monitored media session	4.5.2
Downgrading of media sessions	12.928 kbit per downgrading action	4.6.1
Cancellation of media sessions	4.8 kbit per cancellation action	4.6.2

For all considerations following within this section, the initial value for re-evaluation of group affiliations has been chosen to be 5 seconds.

In order to allow for the evaluation of the proposed framework subject to different side conditions given in the respective NGN, four reference conditions have been chosen, as shown in Table 8.10.

Table 8.10: Reference conditions for framework evaluation

Parameter	**Reference condition *I***	**Reference condition *II***	**Reference condition *III***	**Reference condition *IV***
No. of recognised virtual groups	6	150	6	6
No. of monitored reference sessions per virtual group	2	3	2	2
Amount of downgraded sessions (% of all sessions)	25%	25%	100%	2%
Amount of cancelled sessions (% of all sessions)	10%	10%	50%	0%

While reference condition *I* approximately corresponds to the outcome of the prototypical evaluation as described in section 8.3.2, conditions *II* and *III* were designed as imaginary worst cases, which do not relate to any framework limitations.

Condition *II* has been designed to assume a significantly higher number of reference sessions whose network performance is monitored. This condition corresponds to a large number of recognised virtual groups. In order to simulate that overall poor network performance conditions are detected that seriously affect the QoS perceived by a large number of subscribers, for condition *III* it has been assumed that all sessions are downgraded during their life time, and every second session has to be cancelled actively in order to radically release bandwidth in the network. In contrast, condition *IV* assumes that only two percent of all sessions have to be downgraded and no session has to be cancelled, which corresponds to satisfying network conditions.

In order to compare the framework-related QoS signalling effort with the standardised NGN QoS architecture, for the latter, the following QoS signalling traffic effort is assumed. The calculatory derivation of these values has been introduced in section 3.3.3.

- Static traffic effort for QoS provision per session initiation attempt: 22.8 kbit
- Periodic path-coupled signalling for refreshing of resource reservation: 0.18 kbit/s per unidirectional media stream

Figure 8.8 compares graphically the QoS-related signalling effort of the standardised NGN QoS architecture and the proposed framework, subject to the mean session duration. A fixed amount of 10,000 parallel sessions is assumed. For the proposed framework, four reference conditions are distinguished as stated in Table 8.10.

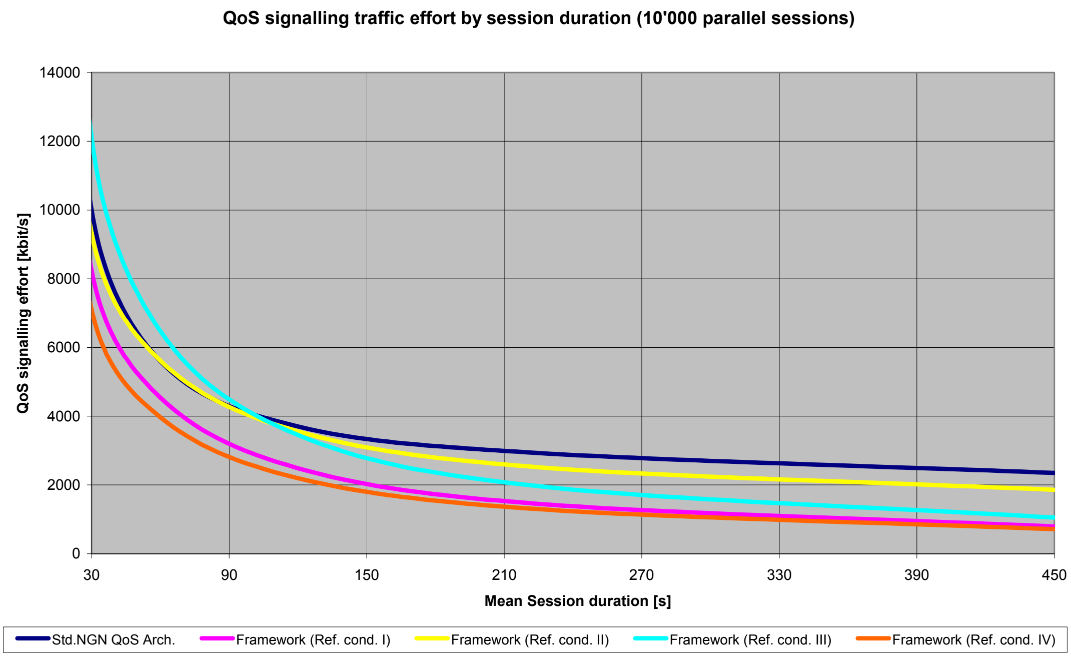


Figure 8.8: Comparison of QoS-related signalling effort subject to mean session duration

From Figure 8.8 it is evident that the proposed framework can help to reduce the QoS signalling effort, compared with the standardised NGN QoS architecture. This is especially true under generally sufficient network performance such as given in reference conditions *I* or *IV*, independent of the considered mean session duration. In contrast, if insufficient network conditions have to be faced (as assumed in reference condition *III*), for session durations below approximately 100 seconds, the QoS signalling effort of the proposed framework is increased compared to the standardised approach. However, for sessions with a longer duration, this effect compensates in favour of the proposed framework.

The increase of the number of monitored sessions results in the overall increase of the QoS signalling effort coming along with the proposed framework. However, as derived from Figure 8.8, even for a relatively large number of monitored sessions (as given in reference condition *II*), the QoS-related traffic amount coming with the framework does not exceed the traffic effort introduced by the standardised approach, if a relatively large number of parallel sessions is considered (here: 10,000 parallel sessions).

Figure 8.9 provides another graphical comparison of the QoS signalling effort. A fixed mean session duration of 120 seconds is considered, and the traffic effort is compared subject to the number of parallel sessions.

Again it is evident that, assuming reference conditions *I* and *IV*, the proposed framework outperforms the standardised NGN QoS architecture, if a mean session duration of 120 seconds is considered. Even if a large amount of sessions has to be downgraded and/or cancelled (as given in reference condition *III*), the amount of QoS optimisation traffic does not exceed the traffic level of the standardised approach.

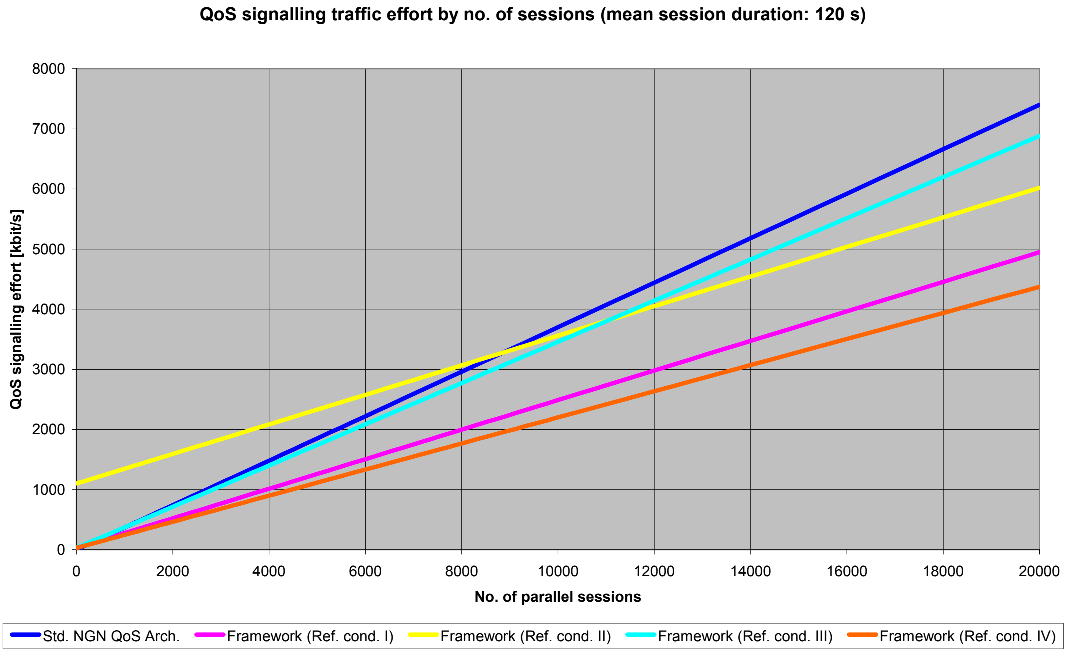


Figure 8.9: QoS-related signalling effort subject to the number of parallel sessions

If a large number of reference sessions have to be monitored (as given in reference condition *II*) and only a relatively small number of sessions exist in parallel, the QoS signalling effort introduced by the framework exceeds the amount of QoS-related signalling coming along with the standardised approach. However, if more than approximately 9000 active sessions exist in parallel, again, this effect compensates in favour of the framework.

Table 8.11 quantitatively compares the signalling effort of the standardised QoS architecture and the proposed framework for a mean session duration of 120 seconds and different numbers of parallel sessions. For this comparison framework reference condition *I* is considered, which corresponds to the outcome of the prototypical framework evaluation. The comparison shows that, considering up to approximately 200 parallel sessions, the QoS-related traffic volume emanating from the standardised NGN QoS architecture is lower compared to the proposed framework.

However, considering higher numbers of parallel sessions that result in significant QoS signalling effort, the application of the proposed framework clearly helps to save QoS-related signalling traffic.

Table 8.11: Quantitative comparison of QoS-related signalling effort

	QoS signalling effort [kbit/s]		
No. of parallel sessions	**Std. NGN QoS Arch.**	**Framework Ref. Cond. I**	**Traffic saving resulting from framework application**
10	3.7	31.8	-
20	7.4	34.3	-
50	18.5	41.6	-
100	37.0	53.9	-
200	74.0	78.5	-
500	185.0	152.3	**17.69%**
1000	370.0	275.2	**25.62%**
2000	740.0	521.1	**29.58%**
5000	1850.0	1258.7	**31.96%**
10000	3700.0	2488.0	**32.76%**

8.4.2 Call set up delay and complexity

This section compares the proposed framework and the standardised NGN QoS architecture regarding both call set up delay and different aspects of complexity.

According to (IETF RFC 6076, 2011) the delay occurring in line with the transport and processing of a session-initiating request is referred to as Session Request Delay (SRD), which corresponds to the time interval between the sending of the initial SIP INVITE message and the receiving of the first status indicative SIP response, measured at the session-originating side only. This metric is also referred to as Post Selection Delay as defined in (ITU-T E.721, 1999).

The mean SRDs occurring in both the application of the standardised NGN QoS architecture and the proposed framework were calculated considering the respective message flows as introduced in section 3.3.3 (standardised architecture) and section 4.4 (framework). For these calculations, as a heuristic example, the estimated time intervals listed in Table 8.12 are assumed for both message transmission and processing.

Table 8.12: Assumed mean time intervals for transmission and processing of messages

Transmission link / Processing action	Mean time interval [ms]
Transmission: User Terminal -- Access Network	20
Transmission: Access Network -- Core Network	10
Processing: traversing network elements (only service+application layer) (such as Call Server, QoS manager, UAG)	1
Processing: Admission Control procedure	100
Processing: Policy Enforcement procedure	100
Processing: AI-based QoS profiling (per existing virtual group)	10

Based on these time intervals, Table 8.13 compares the calculatory mean SRD per call resulting from QoS-relevant signalling and processing for both, the standardised QoS architecture and the proposed framework for given test scenario. For comparative purposes, additional reference target values are included for the mean Post Selection Delay in ISDN-based telecommunication networks under “normal” load (not busy hour-related) as defined in (ITU-T E.721, 1999). The comparison shows that, although the proposed framework introduces an increased call set up delay compared to the standardised NGN QoS architecture, the resulting delay does not exceed target limits which were defined for the ISDN.

Table 8.13: Comparison of mean call set up delay intervals

QoS provision approach / ISDN: Type of connection	Mean Session Request Delay / Mean Post Selection Delay [s]
Standardised NGN QoS architecture	0.717
Proposed QoS optimisation framework (6 virtual groups)	1.418
ISDN: local connection	3.0
ISDN: toll connection	5.0
ISDN: international connection	8.0

Since the framework-specific session set up procedure involves the application of an Artificial Neural Network, attention has also to be paid to the scalability of the set up delay with the rate of new call attempts. Assuming that only one virtual grouping process can be performed at a time in a given Neural Network, due to the queuing effect, further session set up delay would occur if the call attempt rate exceeded a defined number sessions per time. As a consequence, in order to limit the session set up delay introduced by the proposed framework, the QoS Manager must be provided with a sufficient number of Artificial Neural Network instances as well as appropriate processing power. As an example, assuming a time interval of 250 ms required for the profiling process, with a given call rate of 12 session attempts per second, three Artificial Neural Network instances are required in order to avoid additional set up delay due to queuing.

Table 8.14 compares different complexity criteria of the standardised NGN QoS architecture and the proposed QoS optimisation framework. From this comparison it is evident that both QoS provision approaches introduce individual aspects regarding complexity, and hence can not be measured based on a common quantitative scale.

Table 8.14: Comparison of different complexity criteria

Complexity criteria	Std. NGN QoS architecure	Proposed framework
No. of additional messages/processing steps required for initial call setup incl. QoS provision	24	38
No. of logical central entities involved in session setup incl. QoS provision (in both Core and Access Networks)	9 (Call Server, SPDF, RACFs, RCEFs, BGF, etc., in both Access and Core Networks)	2 (Call Server, QoS Manager)
Support of QoS mechanisms required in transport functions (Core and Access Networks)	yes	-
Specific functionalities required at customer side (such as integrated in IAD)	-	yes
Call set up delay [ms] (framework: 6 virtual groups)	717	1418

Obviously, compared to the standardised NGN QoS architecture, the proposed framework introduces an increased signalling and processing effort along with the initial session setup (hence, it also introduces an increased call set up delay). In contrast, the standardised architecture comes with an increased number of logical central entities involved in the QoS provision procedure, each of which runs the risk being a potential point of failure. Additionally, the standardised architecture requires the support of QoS mechanisms in the transport network. On the other hand, in order to apply the proposed QoS optimisation framework, provider-trusted decentralised intermediary network elements (namely UAGs) have to be interconnected logically between each user terminal and the respective connection point at the access network.

Summarising it can be said that, depending on the considered criteria, both the standardised NGN QoS architecture and the proposed QoS optimisation framework come with individual amounts of required effort, introducing complexity in different aspects.

8.5 Summary

Within this chapter, the research prototype developed within this project has been introduced. Its scope is outlined, and the motivation for its sequential processing architecture is described. The schematic functionality architecture of the research prototype is depicted, involving several steps which have to be processed sequentially in order to perform an optimisation run for an NGN communication scenario presented to the prototype. Within each step, a specific action is performed, which might require the application of one or more software or manual processing, respectively. The involved software pieces are named and their functions are briefly described. By following the step-by-step description, the overall prototype functionality is outlined. If it turns out that the given communication scenario has to be adapted in order to include QoS optimisation action, the respective modifications have to be included manually, and the updated scenario is again presented to the research prototype for sequential processing. The research prototype has been successfully adopted for a proof of concept evaluation of the proposed framework, demonstrating its functionalities as well as its general applicability. The chapter concludes with a quantitative comparison of the proposed QoS optimisation framework and the standardised NGN QoS architecture, covering the signalling effort, call set up delay, and the complexity introduced by both QoS provision approaches.

9 Conclusions

This chapter concludes the thesis by summarising the achievements of the research programme (section 9.1). Furthermore, the limitations of the performed research are discussed (section 9.2), and areas for further research are considered (section 9.3).

9.1 Achievements of the research

The research performed within this project was dedicated to the development of a novel approach to facilitate and simplify the provision of Quality of Service in NGN. This has been accomplished through the research-based definition of a comprehensive framework for QoS optimisation in NGN, completed by the elaboration of essential detail aspects and novel techniques. The framework can be applied as both, a stand-alone QoS optimisation solution for NGN providing a sufficiently dimensioned transport infrastructure, or as a supplementary precursor stage for QoS provision, notably facilitating the active control of QoS-related transport network characteristics.

The imperative for the development of alternative or complementary QoS optimisation techniques has been substantiated through extensive review of various existing QoS provision approaches with a main (however, not exclusive) focus on those solutions especially designed for the use in NGN (see chapter 3). Particular attention has been dedicated to an exemplary reference NGN QoS architecture based on ITU-T and ETSI standardisation work, which, on the one hand, offers strict QoS provisioning but, on the other hand, comes along with essential vulnerabilities which could be exposed. The review of conventional NGN QoS provision approaches resulted in the definition of the necessity of supplementary alternative solutions, and the definition of requirements which have to be considered.

Based on these requirements, a general framework has been engineered, providing the optimisation of QoS conditions in SIP-based NGN by pursuing a soft QoS engineering approach, allowing for a simpler QoS provision architecture (see section 4). The framework architecture consists of a centralised unit termed QoS Manager and logical entities called UAGs (User Access Gates). The latter are interconnected between the user terminals and the NGN. These entities are required for QoS monitoring purposes, and to ensure the consideration of orders given by the QoS Manager, which queries and analyses information on the QoS as perceived by the users, and determines appropriate action to optimise QoS conditions. Several mechanisms required for the operation of this framework have been defined, including signalling sequences and operational methodologies for the resource-saving but comprehensive determination and optimisation of the QoS perceived by any active subscriber at any time.

The most essential methodology developed for this framework, providing the fundamental basis for the overall framework functionality, has been termed QoS profiling (see chapter 5). This novel methodology allows for the identification and virtual grouping of media sessions which share network paths essentially influencing the QoS conditions of these sessions. This is performed by matching value sequence patterns which correspond to the network performance affecting the media streams. In order to accomplish this unsupervised pattern recognition task, an ART 2 Artificial Neural Network (ANN) has been identified as a feasible approach. To allow for completely automated und unsupervised pattern recognition operation and, at the same time, consider both, the universal applicability of the framework and the variable character of network performance patterns, has been identified as a further challenge regarding the introduction of QoS profiling. This challenge could be addressed by the development of both, a bootstrap mechanism and an accuracy self-scaling functionality which come as universal add-ons for any application of ART 2 Artificial Neural Networks in need of the respective functionalities.

The comprehensive and resource-saving estimation of QoS conditions as perceived by any subscriber actively involved in a communication session comes as another main function of the developed framework (see chapter 6) to identify the demands for the optimisation of QoS conditions. Since the utilisation of QoS profiling generally allows for the virtual grouping of media sessions which experience similar QoS conditions, the QoS monitoring of only a subset of communication end points is deemed to be sufficient in order to save monitoring traffic. To further minimise the traffic volume generated inline with QoS monitoring, a novel method has been developed to classify media sessions regarding their suitability as QoS reference sessions, based on their expected remaining duration. For the evaluation of QoS conditions from network performance parameters, the ITU-T E model has been chosen as a generally appropriate instrument. In order to adapt the E model for continuous real-time QoS estimation, refinements have been proposed regarding the consideration of jitter and statistical packet loss.

In contrast to the standardised reference NGN QoS architecture, the application of cross-layer signalling was explicitly excluded from the proposed framework, which, hence was not designed to provide guaranteed network performance for the lifetime of a communication session. Therefore an approach has been developed which is based on the dynamic control of the network utilisation on network paths associated with a distinct virtual group (see chapter 7). Since the active control of transport network resources had to be avoided, the network utilisation is adjusted in a passive way by restricting the traffic load introduced to a network path associated with a respective virtual group. For controlling the traffic volume, both the downgrading of sessions to media codecs with lower bitrate demands and the rejection or cancellation of communication sessions were considered. Since especially the latter variant comes as a noticeable inconvenience for the users, concerned sessions have to be choosen carefully. Therefore, a novel function has been developed which classifies concurrent sessions into a ranking list by their objective significances. The objective significance of a session is determined not only by service availability rates as defined within applicable SLA contracts, but also the de facto service availability previously experienced by the concerned users is considered. Furthermore, a simple but effective control scheme has been developed which allows for the adjustment of QoS conditions affecting media sessions assigned to dedicated virtual groups.

For the verification of the overall framework functionalities, a research prototype (see chapter 8) has been developed, providing the most relevant framework functionalities in a semi-automatic manner. The prototype combines an open source packet network simulation software widely accepted within the academic community (ns-2) and an open source Artificial Neural Network simulation software (SNNS) originally developed at the University of Stuttgart, Germany. To allow for the realistic simulation of an NGN communication environment, a rudimentary open source third party SIP add-on for ns-2 was extended to allow for media codec negotiation, media data stream simulation for selected voice codecs, and the ability to perform SIP-based session modification (re-INVITE). In order to combine ns-2 and SNNS, a middleware has been designed and programmed for the off-line read-out and analyses of QoS-relevant network performance data from ns-2 trace files, and for the preprocessing of the obtained data for the unsupervised pattern matching performed by SNNS. Since requiring multiple classification runs, the SNNS pattern recognition process is initiated and managed by a cover software which was also designed in line with this project.

In general, the proposed framework comes as an effective alternative to or extension for the standardised NGN QoS architecture. Its QoS improvement capabilities were verified by the use of simulations, demonstrating that the framework was able to provide an overall MOS of (4.0 $^{+}/_{-}$ 0.3; referring to a "good" voice quality on the MOS scale) to a set of voice telephony sessions which, without any QoS optimisation action, would have been severly affected by insufficient network performance.

Providing reference framework and network conditions, it was demonstrated that the QoS-related signalling effort introduced by the proposed framework underruns the traffic volume coming with the standardised architecture if more than 240 calls/sessions existed in parallel. Considering 500 parallel calls, more than 17 percent of QoS-related traffic volume are saved, which is further improved to more than 30 percent saving for more than 10,000 parallel sessions.

Compared to the standardised QoS architecture, the proposed framework becomes more efficient in terms of QoS signalling effort if the number of media sessions associated with a dedicated virtual group reaches its optimum. This optimum depends on several framework- and network-specific factors, but, however, can be said to be in a range of 47 to 200 media sessions per virtual group, with a number of 80 media sessions per group for the reference setup introduced in chapter 8. Note that the optimum number of media sessions per group is independent of the overall number of parallel existing sessions.

Due to the overall amount of time required for the collection of network performance data throughout the session set up, the specific framework signalling procedures, and the QoS profiling-related processing the proposed framework introduces a session set up delay that generally exceeds the set up delay coming with the application of the standardised NGN QoS architecture. However, for a reference framework setup, a mean session set up delay of 1.418 seconds was obtained, which clearly outperforms the defined target values for the ISDN post-selection delay.

Beside the time required for the transport und processing of messages, the session set up delay introduced by the proposed framework mainly consists of the (fixed) time period required for the collection of network performance data throughout the session initiation process, and the (varying) time period required for the initial QoS profiling, which involves data processing within an Artificial Neural Network (ANN). Generally the session set up delay can be said to scale with the rate of new call attempts if a sufficient number of ANN instances are provided, which are run simultaneously by the central QoS Manager. Note that each ANN instance has to be provided with reasonable processing power. The maximum rate of call attempts served per ANN instance can be calculated as the ratio of *(maximum tolerable delay caused by the QoS profiling process)* over *(time period required per QoS profiling* *process)*, hence the number of required ANN instances can be calculated by dividing the desired call attempt rate by the maximum call attempt rate obtained per ANN.Several papers related to different aspects of this research have been presented at different conferences and have received positive comments from reviewers and delegates.

9.2 Limitations of the research

Although the overall objectives of the research project have been met, some decisions had to be made which resulted in limitations imposed on the work. Those decisions were caused by practical reasons, or were made to delimit the considered research project from related fields of study which could not be fully covered by this research due to generally given time scope limitations for the accomplishment of research degree studies. The key limitations are summarised below.

1. Within the developed framework, if required, the utilisation of distinct network paths is adjusted by controlling the traffic volume caused by media streams of connection-oriented SIP sessions. Therefore a central SIP infrastructure is used, managing the initiation, termination, and codec selection of media sessions between users. However, beside being involved in session-based communication, users could also send and receive arbitrary kinds of data which are typically not managed by the use of SIP. This kind of data traffic comprises most types of Internet traffic, be it elastic or inelastic, such as obtaining web content, downloading files, or using video streaming platforms. The defined framework with its given scope of operation has not been designed to control the amount of traffic caused by applications not

managed through a central SIP infrastructure. However, as this type of traffic is generally considered by the framework regarding its impact on the QoS provided for session-based real-time communication, this limitation does not cause any unsteadiness regarding the overall functionality of the defined framework. Furthermore, an extension of the given framework scope to also consider the control of not session-based traffic for QoS optimisation is generally feasible. However, due to the required effort, developing this extension has been left for further research.

2. The given framework allows for the optimisation of QoS conditions within the scope of an NGN, whose subscribers are attached by the use of intermediary entities (called UAGs) logically located between the user terminal and the IP transport network (note that those intermediary entities might also be integrated either with user terminals or IADs (Integrated Access Devices)). An NGN is typically interconnected with other telecommunication networks such as other NGN or circuit-switched PSTN. In both cases it can be assumed that inter-network media streams are exchanged between the networks via logically terminating interconnection points such as MGWs (Media GateWays) or SBCs (Session Border Controllers). Since those interconnection points can be easily equipped with UAGs representing multiple communication endpoint instances, those points can be considered by the framework processes of both, QoS monitoring and optimisation. However, the scope of influence of the QoS optimisation framework is limited to the inside of a virtual circle spanned among all connected UAGs. Hence, QoS optimisation action for inter-network media sessions is limited to

the considered NGN. However, the developed framework could be extended regarding the awareness of inter-NGN connections, allowing for the integration of QoS negotiation and/or control methods specific to the interconnected NGN.

3. The research prototype designed along with this project has been planned to provide all relevant functionalities required to allow for general proof of concept simulation of the developed framework. However, for the benefit of providing more time to required research, the prototype has not been finalised as a fully integrated software system, but comes as a collection of separate executable software components which have to be run sequentially in order to simulate the full functionality of the framework. Since the operation of this semi-automatic prototype requires human interaction, the prototype is not applicable to run in stand-alone mode. However, since all relevant functions of the framework have been successfully implemented, the prototype allows for framework proof of concept simulation.

Despite these limitations, the research project has made valid contributions to knowledge and provided sufficient proof of concept for the proposed approaches.

9.3 Suggestions and scope for further work

This research project has advanced the field of QoS optimisation for SIP-based NGN. However, a number of areas for future work can be identified, building upon the results of this project. Some of these areas have already been mentioned in previous chapters. These and further ideas are summarised below.

1. Further research work should be performed to investigate the applicability of the defined framework within environments also including mobile access networks. Since the effect of codec downgrading and/or session cancellation on the QoS given in mobile access networks has already been considered by several research teams, a substantial basis is evident regarding the use of the herewith defined framework in environments involving mobile access networks. However, further research is required, which should especially focus on the operability of QoS profiling when mobile access networks are involved in end-to-end communication.

2. An extension to the defined framework could be developed, allowing for the full consideration of traffic which is not based on SIP sessions with the given framework. Regarding its impact on effective QoS conditions, this traffic is already considered by the given framework. However, further research would be required to allow for the comprehension of not session-based traffic even for the moderation of the network utilisation for the improvement of QoS conditions.

3. The given framework could be provided with extended functionality regarding the recognition of specific events or incidents which have occurred in the past under similar conditions, such as recurring QoS impairment in the busy hour. Providing the framework with the ability to memorise previous occurences and learn from the effects of the performed intervention would result in improved history-based QoS optimisation by anticipatory framework behaviour. However, extensive research is required in order to allow for the integration with the required intelligence.

4. In order to integrate the given framework with both, NGN infrastructures and transport networks which support conventional QoS mechanisms, the framework could be extended to allow for the combined consideration of passive and active QoS control. As an alternative several NGN, each being provided with the QoS optimisation framework, could be interconnected, with conventional QoS mechanisms (such as IntServ or DiffServ) used only for inter-network QoS provision. However, further research is required to determine the best trade-off between these procedures and application scenarios.

Especially following the last-mentioned suggestion might result in a promising approach to provide QoS for real-time communication in a less complex and more efficient way. It is therefore the hope of the author of this thesis that the research work performed within this project is considered as a meaningful contribution to the field of Quality of Service provision in SIP-based NGN.

References

1. 3GPP TS 23.228 (2006), Technical Specification, “IP Multimedia Subsystem (IMS); Stage 2 (Release 5)”, 3GPP

2. Abdelzaher, Tarek and Shin, Kang G. (1998), “End-host Architecture for QoS-Adaptive Communication”, *Proceedings of the Fourth IEEE Real-Time Technology and Applications Symposium*, pp. 121-130, IEEE

3. Abu Salah, Stefan; Brack, Stephan; Grebe, Andreas; Marikar, Achim; Trick, Ulrich and Weber, Frank (2008), “BMBF-Forschungsprojekt: Verbesserung der netzeübergreifenden Quality of Service bei SIP-basierter VoIP-Kommunikation (QoSSIP) – Abschlussbericht” (translated title: “Public-funded research project: Improvement of inter-network Quality of Service in SIP-based VoIP communication (QoSSIP) – Final report”), BMBF Project Funding Reference Numbers 1715A05 (FH Köln) und 1715B05 (FH Frankfurt), Universities of Applied Sciences FH Köln, Germany and FH Frankfurt a. M., Germany

4. AL-Akhras, M.; Zedan, H.; John, R. and ALMomani, I. (2009), “Non-intrusive speech quality prediction in VoIP networks using a neural network approach”, *Neurocomputing*, Vol. 72, Issue 10-12, pp. 2595-2608, Elsevier

5. Al-Begain, Khalid; Balakrishna, Chitra; Galindo, Luis Angel and Fernandez, David Moro (2009), *IMS: A Development and Deployment Perspective*, 1st edition, Wiley, ISBN: 978-0-470-74034-7

6. AlQahtani, Salman A. and Mahmoud, Ashraf S. (2006), “Dynamic radio resource allocation for 3G and beyond mobile wireless networks”, *Computer Communications*, Vol. 30, pp. 41-51, Elsevier

7. Aradhye, Hrishikesh B.; Bakshi, Bhavik R.; Davis, James F. and Ahalt, Stanley C. (2004), “Clustering in Wavelet Domain: A Multiresolution ART Network for Anomaly Detection”, *AIChE Journal* (American Institute of Chemical Engineers), Vol. 50, Issue 10, pp. 2455-2466, Wiley InterScience

8. Atzori, L.; Lobina, M. L. and Corona, M. (2006), “Playout buffering of speech packets based on a quality maximization approach”, *IEEE Transactions on Multimedia*, Vol. 8, Issue 2, pp. 420-426, IEEE

9. Bandung, Yoanes; Machbub, Carmadi; Langi, Armein Z. R. and Supangkat, Suhono H. (2008), “Optimizing Voice over Internet Protocol (VoIP) Networks Based-on Extended E-model”, *IEEE Conference on Cybernetics and Intelligent Systems 2008*, pp. 801-805, IEEE

10. Barceló, Francisco and Jordán, Javier (2000), “Channel Holding Time Distribution in Public Telephony Systems (PAMR and PCS)”, *IEEE Transactions on vehicular technology*, Vol. 49, No. 5, pp. 1615-1625, IEEE

11. Bianchi, G.; Borgonovo, F.; Capone, A.; Fratta, L. and Petrioli, C. (2002), "Endpoint Admission Control with Delay Variation Measurements for QoS in IP Networks", *ACM SIGCOMM Computer Communication Review*, Vol. 32, Issue 2, pp. 61-69, ACM

12. Boggia, G.; Camarda, P.; D'Alconzo, A.; De Biasi, A. and Siviero, M. (2005), "Drop Call Probability in Established Cellular Networks: from data Analysis to Modelling", *IEEE 61st Vehicular Technology Conference*, Vol. 5, pp. 2775-2779, IEEE

13. Bohnert, Thomas Michael; Monteiro, Edmundo; Curado, Marilia; Fonte, Alexandre; Ries, Michal; Moltchanov, Dmitri and Koucheryavy, Yevgeni (2007), "Internet Quality of Service: A Bigger Picture", *1st OpenNet Workshop - Service Quality and IP Network Business: Filling the Gap*, Diegem/Brussels, Belgium, March 2007

14. Borella, Michael S.; Swider, Debbie; Uludag, Suleyman and Brewster, Gregory B. (1998), "Internet packet loss: measurement and implications for end-to-end QoS", *Proceedings of the 1998 ICPP Workshops on Architectural and OS Support for Multimedia Applications/Flexible Communication Systems/Wireless Networks and Mobile Computing*, 14 Aug 1998, pp. 3-12, IEEE

15. Boutros, Paul C. and Okey, Allan B. (2005): "Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data", *Briefings in Bioinformatics*, Vol. 6, No. 4, pp. 331-343, Henry Stewart Publications

16. Callejo-Rodríguez, M. A.; Enríquez-Gabeiras, J.; Burakowski, W.; Beben, A.; Sliwinski, J.; Dugeon, O.; Mingozzi, E.; Stea, G.; Diaz, M. and Baresse, L. (2008), "EuQoS: End-to-End QoS over Heterogeneous Networks", *First ITU-T Kaleidoscope Academic Conference, Innovations in NGN: Future Network and Services, 2008, K-INGN 2008*, pp. 177-184, IEEE

17. Calyam, Prasad; Sridharan, Mukundan; Mandrawa, Weiping and Schopis, Paul (2004), "Performance Measurement and Analysis of H.323 Traffic", *Proc. of Passive and Active Measurement Workshop 2004*, pp. 137-146, Springer

18. Carpenter, Gail A. and Grossberg, Stephen (1987), "ART 2: Self-organization of Stable Category Recognition Codes for Analog Input Patterns", *Applied Optics*, Vol. 26, No. 23, pp. 4919-4930, Optical Society of America

19. Carvalho, Leandro; Mota, Edjair; Aguiar, Regeane; Lima, Ana F.; de Souza, José Neuman and Barreto, Anderson (2005), "An E-Model Implementation for Speech Quality Evaluation in VoIP Systems", *Proceedings of the 10th IEEE Symposium on Computers and Communications (ISCC 2005)*, pp. 933-938, IEEE

20. Cho, Eun-Hee; Shin, Kang-Sik and Yoo, Sang-Jo (2006), "SIP-based Qos support architecture and session management in a combined IntServ and

DiffServ networks", *Computer Communications*, Vol. 29, No. 15, pp. 2996-3009, Elsevier Science

21. Cho, Il Kwon and Okamura, Koji (2009), "A Centralized Resource and Admission Control Scheme for NGN Core Networks", *International Conference on Information Networking (ICOIN 2009)*, IEEE

22. Clark, A. D. (2001), "Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality", *IP Telephony Workshop 2001*, Columbia University

23. Cochennec, Jean-Yves (2002), "Activities on next-generation networks under Global Information Infrastructure in ITU-T", *Communications Magazine*, Vol. 40, Issue 7, pp. 98-101, IEEE

24. David, Josheff C.; Sanmartin, Paul M. and Márquez, José D. (2010), "Model of QoS on NGN: An Analysis of Performance", *Electronics, Robotics and Automotive Mechanics Conference (CERMA 2010)*, pp. 271-276, IEEE

25. Ding, Lijing and Goubran, Rafik A. (2003), "Speech Quality Prediction in VoIP Using the Extended E-Model", *Global Telecommunications Conference (GLOBECOM '03)*, IEEE

26. Ding, Yu-Xin; Shi, Yan; Shi, Yong and Jiang, Jun-Qing (2008), "A hybrid clustering algorithm based on ART2 and its application in anomaly detection", *Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition*, pp.282-286, IEEE

27. Do Valle, Raquel F.; de Carvalho, Leandro S. G.; Aguiar, Regeane B.; Mota, Edjair S. and Freitas, Diógenes (2010), "Dynamical Management of Dejitter Buffers Based on Speech Quality", *2010 IEEE Symposium on Computers and Communications (ISCC)*, pp. 56-61, IEEE

28. DOCSIS 1.1 CM-SP-RFIv1.1-C01-050907 (2005), "Radio Frequency Interface Specification", Cable Television Laboratories, Inc.

29. ETSI ES 282 001 V2.0.0 (2008), ETSI Standard, "NGN Functional Architecture", ETSI TISPAN

30. ETSI ES 282 003 V2.0.0 (2008), ETSI Standard, "Resource and Admission Control Sub-System (RACS): Functional Architecture", ETSI TISPAN

31. ETSI TR 180 000 V1.1.1 (2006), Technical Report, "NGN Terminology", ETSI TISPAN

32. ETSI TR 182 022 V2.0.0 (2007), Technical Report, "Architectures for QoS handling", ETSI TISPAN

33. ETSI TS 185 001 V1.1.1 (2005), Technical Specification, "NGN Quality of Service (QoS) Framework and Requirements", ETSI TISPAN

34. European Regulators Group (ERG) (2008a), "ERG Common Statement on Regulatory Principles of IP-IC / NGN Core - A work program towards a Common Position", ERG

35. European Regulators Group (ERG) (2008b), “Supplementary Document to the ERG Common Statement on Regulatory Principles of IP-IC / NGN Core - A work program towards a Common Position”, ERG

36. Freriks, L. W.; Cluitmans, P. J. M. and van Gils, M. J. (1992), “The Adaptive Resonance Theory Network: (Clustering-) Behaviour in Relation With Brainstem Auditory Evoked Potential Patterns“, *EUT report 92-E-264*, Eindhoven University of Technology, Netherlands, ISBN 90-6144-264-8

37. Ghazel, Cherif and Saïdane, Leila (2009), “Achieving a QoS Target in NGN Networks via an Efficient Admission Control Strategy“, *Eighth International Conference on Networks (ICN '09)*, pp. 318-323, IEEE

38. Giordano, Silvia; Salsano, Stefano; Van den Berghe, Steven; Ventre, Giorgio and Giannakopoulos, Dimitrios (2003), “Advanced QoS Provisioning in IP Networks: The European Premium IP Projects“, *Communications Magazine*, Vol. 41, Issue 1, pp. 30-36, IEEE

39. Gozdecki, Janusz; Jajszczyk, Andrzej and Stankiewicz, Rafal (2003), “Quality of service terminology in IP networks”, *Communications Magazine*, Vol. 41, Issue 3, pp. 153-159, IEEE

40. Grossberg, Stephen (1976a), “Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors”, *Biological Cybernetics*, Vol. 23, No. 3, pp. 121-134, Springer

41. Grossberg, Stephen (1976b), “Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions”, *Biological Cybernetics*, Vol. 23, No. 3, pp. 187-202, Springer

42. Hagsand, Olof; Más, Ignacio; Marsh, Ian and Karlsson, Gunnar (2004), “Self-Admission Control for IP Telephony Using Early Quality Estimation”, *NETWORKING 2004, Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications, LNCS*, Vol. 3042, pp. 381-391, Springer

43. Henning, Karina (2002), “Interaktive Einführung in neuronale Netze (Teilgebiet: ART-Architekturen)” (translated title: “Interactive introduction in neural networks (subdiscipline: ART architectures)”), Exam Thesis, University of Münster, Germany

44. Herrmann, Kai-Uwe (1992), “ART (Adaptive Resonance Theory) - Architekturen, Implementierung und Anwendung” (translated title: “ART (Adaptive Resonance Theory) - architectures, implementation and application”), Diploma Thesis, University of Stuttgart, Germany

45. Hole, David P. and Tobagi, Fouad A. (2004), “Capacity of an IEEE 802.11b wireless LAN supporting VoIP”, *IEEE International Conference on Communications*, pp. 196-201, IEEE

46. IETF RFC 1633 (1994), Request For Comments, “Integrated Services in the Internet Architecture: an Overview”, IETF

47. IETF RFC 2205 (1997), Request For Comments, “Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification”, IETF

48. IETF RFC 2474 (1998), Request For Comments, “Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers”, IETF

49. IETF RFC 2475 (1998), Request For Comments, “An Architecture for Differentiated Services”, IETF

50. IETF RFC 3031 (2001), Request For Comments, “Multiprotocol Label Switching Architecture”, IETF

51. IETF RFC 3261 (2002), Request For Comments, “SIP: Session Initiation Protocol”, IETF

52. IETF RFC 3262 (2002), Request For Comments, “Reliability of Provisional Responses in Session Initiation Protocol (SIP)”, IETF

53. IETF RFC 3264 (2002), Request For Comments, “An Offer/Answer Model with the Session Description Protocol (SDP), IETF

54. IETF RFC 3265 (2002), Request For Comments, “Session Initiation Protocol (SIP) Specific Event Notification”, IETF

55. IETF RFC 3311 (2002), Request For Comments, “The Session Initiation Protocol (SIP) UPDATE Method”, IETF

56. IETF RFC 3312 (2002), Request For Comments, “Integration of Resource Management and Session Initiation Protocol (SIP)”, IETF

57. IETF RFC 3428 (2002), Request For Comments, “Session Initiation Protocol (SIP) Extension for Instant Messaging”, IETF

58. IETF RFC 3515 (2003), Request For Comments, “The Session Initiation Protocol (SIP) Refer Method”, IETF

59. IETF RFC 3550 (2003), Request For Comments, “RTP: A Transport Protocol for Real-Time Applications”, IETF

60. IETF RFC 3581 (2003), Request For Comments, “An Extension to the Session Initiation Protocol (SIP) for Symmetric Response Routing”, IETF

61. IETF RFC 3588 (2003), Request For Comments, “Diameter Base Protocol”, IETF

62. IETF RFC 3611 (2003), Request For Comments, “RTP Control Protocol Extended Reports (RTCP XR)”, IETF

63. IETF RFC 3725 (2004), Request For Comments, “Best Current Practices for Third Party Call Control (3pcc) in the Session Initiation Protocol (SIP)”, IETF

64. IETF RFC 3856 (2004), Request For Comments, “A Presence Event Package for the Session Initiation Protocol (SIP)”, IETF

65. IETF RFC 3951 (2004), Request For Comments, “Internet Low Bit Rate Codec (iLBC)”, IETF

66. IETF RFC 4032 (2005), Request For Comments, “Update to the Session Initiation Protocol (SIP) Preconditions Framework”, IETF

67. IETF RFC 4080 (2005), Request For Comments, “Next Steps in Signaling (NSIS): Framework”, IETF

68. IETF RFC 4124 (2005), Request For Comments, “Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering”, IETF

69. IETF RFC 4566 (2006), Request For Comments, “SDP: Session Description Protocol”, IETF

70. IETF RFC 5027 (2007), Request For Comments, “Security Preconditions for Session Description Protocol (SDP) Media Streams”, IETF

71. IETF RFC 5905 (2010), Request For Comments, “Network Time Protocol Version 4: Protocol and Algorithms Specification”, IETF

72. IETF RFC 6035 (2010), Request For Comments, “Session Initiation Protocol Event Package for Voice Quality Reporting”, IETF

73. IETF RFC 6076 (2011), Request For Comments, “Basic Telephony SIP End-to-End Performance Metrics”, IETF

74. IETF RFC 6241 (2011), Request For Comments, “Network Configuration Protocol (NETCONF)”, IETF

75. IETF RFC 6337 (2011), Request For Comments, “Session Initiation Protocol (SIP) Usage of the Offer/Answer Model”, IETF

76. iLocus (2010), “41 percent of voice equipment in wireless operator networks is IP”, Blog on the *iLocus 11th annual VoIP industry report*, http://www.ilocus.com/content/blog/41-percent-voice-equipment-wireless-operator-networks-ip (last visited: 2011-12-22)

77. ITU-T E.721 (1999), Recommendation, “Network grade of service parameters and target values for circuit-switched services in the evolving ISDN”, ITU-T

78. ITU-T E.800 (2008), Recommendation, “Definitions of terms related to quality of service”, ITU-T

79. ITU-T G.107 (2009), Recommendation, “The E-model: a computational model for use in transmission planning”, ITU-T

80. ITU-T G.113 (2007), Recommendation, “Transmission impairments due to speech processing”, ITU-T

81. ITU-T G.114 (2003), Recommendation, “One-way transmission time”, ITU-T

82. ITU-T G.711 (1988), Recommendation, “Pulse code modulation (PCM) of voice frequencies”, ITU-T

83. ITU-T G.723.1 (2006), Recommendation, “Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s”, ITU-T

84. ITU-T G.729 (2007), Recommendation, “Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)”, ITU-T

85. ITU-T P.10/G.100 Amd. 2 (2008), Amendment 2 to ITU-T Recommendation P.10/G.100 (2006), “Vocabulary for performance and quality of service – Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100”, ITU-T

86. ITU-T P.563 (2004), Recommendation, “Single-ended method for objective speech quality assessment in narrow-band telephony applications”, ITU-T

87. ITU-T P.800 (1996), Recommendation, “Methods for the subjective determination of transmission quality”, ITU-T

88. ITU-T P.800.1 (2006), Recommendation, “Mean Opinion Score (MOS) terminology”, ITU-T

89. ITU-T P.862 (2001), Recommendation, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, ITU-T

90. ITU-T Y.1291 (2004), Recommendation, “An architectural framework for support of Quality of Service in packet networks”, ITU-T

91. ITU-T Y.1541 (2006), Recommendation, “Network performance objectives for IP-based services”, ITU-T

92. ITU-T Y.2001 (2004), Recommendation, “General overview of NGN”, ITU-T

93. ITU-T Y.2011 (2004), Recommendation, “Next Generation Networks – Frameworks and functional architecture models”, ITU-T

94. ITU-T Y.2012 (2010), Recommendation, “Functional requirements and architecture of next generation networks”, ITU-T

95. ITU-T Y.2111 (2006), Recommendation, “Resource and admission control functions in Next Generation Networks”, ITU-T

96. Jain, A. K.; Mao, Jianchang and Mohiuddin, K. M. (1996), “Artificial neural networks: a tutorial”, *Computer*, Vol. 29, No. 3, pp. 31-44, IEEE

97. Jambunathan, K.; Fontama, V. N.; Hartle, S. L. and Ashforth-Frost, S. (1997), “Using ART2 networks to deduce flow velocities”, *Artificial Intelligence in Engineering*, 11, pp. 135-141, Elsevier

98. Jedrzycki, Chris and Leung, Victor C.M. (1996), “Probability Distribution of Channel Holding Time in Cellular Telephony Systems”, *IEEE 46th Vehicular Technology Conference*, Vol. 1, pp. 247-251, IEEE

99. Jyoti, Jeewan; El-Tawab, Samy; El-Derini, M.Nazih; Aboelela, Emad and Aly, Hussein (2006), “Improving Quality of Service for Voice-Over-IP Using Routing Diversity”, *23rd Biennial Symposium on Communications*, pp. 364-367, IEEE

100. Kim, Anbin and Jeong, Seong-Ho (2011), “Adaptive QoS Control in Advanced Networks”, *Third International Conference on Ubiquitous and Future Networks (ICUFN) 2011*, pp. 264-267, IEEE

101. Kos, Anton; Klepec, Borut and Tomazic, Saso (2002), “Techniques for performance improvement of VoIP applications”, *11th Mediterranean Electrotechnical Conference 2002 (MELECON 2002)*, pp. 250-254, IEEE

102. Lewcio, Blazej; Wältermann, Marcel; Möller, Sebastian and Vidales, Pablo (2009), “E-model supported switching between narrowband and wideband speech quality”, *International Workshop on Quality of Multimedia Experience 2009 (QoMEx 2009)*, pp. 98-103, IEEE

103. Li, Zhongbo; Zhao, Shenghui; Xie, Xiang and Kuang, Jingming (2008), “An Improved Speech Playout Buffering Algorithm Based on a New Version of E-Model in VoIP”, *Third International Conference on Communications and Networking in China 2008 (ChinaCom 2008)*, pp. 122-126, IEEE

104. Mani, Mehdi and Crespi, Noël (2005), “New QoS Control Mechanism Based on Extension to SIP for Access to UMTS Core Network over Hybrid Access Networks”, *IEEE Wireless and Mobile Computing, Networking and Communications*, WiMob 2005, Vol. 2, pp. 150-157, IEEE

105. Martini, B.; Baroncelli, F.; Martini, V.; Torkman, K. and Castoldi, P. (2009), “ITU-T RACF implementation for application-driven QoS control in MPLS networks”, *IFIP/IEEE International Symposium on Integrated Network Management (IM'09)*, pp. 422-429, IEEE

106. Mase, Kenichi and Toyama, Yuichiro (2002), “End-to-End Measurement Based Admission Control for VoIP Networks”, *IEEE International Conference on Communications (ICC 2002)*, Vol. 2, pp. 1194-1198, IEEE

107. Matsumoto, N.; Hayashi, M. and Tanaka, H. (2009), “Network Middleware Design for Bridging Legacy Infrastructures and NGN”, *Next Generation Internet Networks (NGI '09)*, IEEE

108. Mautner, Pavel; Rohlik, Ondrej; Matousek, Vaclav and Kempf, Juergen (2002), “Signature verification using ART-2 neural network”, *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP '02)*, Vol. 2, pp. 636-639, IEEE

109. Mei, R. D. van der; Meeuwissen, H. B. and Phillipson, F. (2005), “User Perceived Quality-of-Service for Voice-over-IP in a Heterogeneous Multi-Domain Network Environment”, *Proceedings of the 19th International Teletraffic Congress (ITC-19)*, Beijing, September 2005, pp. 1109-1120

110. Menth, Michael (2006), “Das echtzeitfähige und ausfallsichere Internet der nächsten Generation – Next Generation Networks” (translated title: “The real-time capable and fail-proof Internet of the next generation – Next Generation Networks”), *Lecture series on 'The Internet and its applications'*, 25 April 2006, Brandenburg University of Technology, Cottbus, Germany

111. Mohajerzadeh, Amir Hossein; Monsefi, Reza; Yaghmaee, Mohammad Hossein and Farzaneh, Nazbanoo (2010), “An Efficient and Class based Active Queue Management for Next Generation Networks”, *5th International Symposium on Telecommunications (IST'2010)*, pp. 255-260, IEEE

112. Ohuchi, Toshiya; Tanabe, Shiro and Kawanishi, Hidehiko (1994), “A network control architecture for advanced personal communications”, *Global Telecommunications Conference (GLOBECOM '94), Communications: The Global Bridge*, Vol. 3, pp. 1707-1711, IEEE

113. Park, Juyoung and Kang, Shin Gak (2005), “QoS Architecture for NGN”, *Advanced Communication Technology (ICACT 2005)*, pp. 1064-1067, IEEE

114. Prior, Rui Pedro de Magalhães Claro (2007), “Scalable Network Architectures Supporting Quality of Service”, PhD thesis, Faculty of Sciences of the University of Porto

115. Qiu, Jianming; Shao, Huai-Rong; Zhu, Wenwu and Zhang, Ya-Qin (2001), “An end-to-end probing-based admission control scheme for multimedia applications”, *IEEE International Conference on Multimedia and Expo (ICME 2001)*, pp. 665-668, IEEE

116. Raja, Adil; Atif Azad, R. Muhammad; Flanagan, Colin and Ryan, Conor (2007), “Real-Time, Non-intrusive Evaluation of VoIP”, *Proceedings of the 10th European conference on Genetic programming (EuroGP'07), LNCS*, Vol. 4445, pp. 217-228, Springer

117. Rani, B. S. and Renganathan, S. (2003), “Wavelet based Texture Classification with Evolutionary Clustering Networks”, *Conference on Convergent Technologies for Asia-Pacific Region (TENCON 2003)*, Vol. 1, pp. 239-243, IEEE

118. Rayón Villela, P. and Sossa Azuela, J. H. (2000), “A Procedure to Select the Vigilance Threshold for the ART2 for Supervised and Unsupervised Training”, *LNCS*, Vol. 1793, pp. 389-400, Springer

119. Ren, Jiuchun; Zhang, ChongMing; Huang, WeiChao and Mao, Dilin (2010), “Enhancement to E-MODEL on Standard deviation of Packet Delay”, *3rd International Conference on Information Sciences and Interaction Sciences (ICIS 2010)*, pp. 256-259, IEEE

120. Ripley, Brian D. (2005), *Pattern Recognition and Neural Networks*, 8th print., Cambridge University Press, ISBN: 0-521-46086-7

121. Sá, J. P. Marques de (2001), *Pattern recognition: concepts, methods, and applications*, Springer, ISBN: 3-540-42297-8

122. Saika, A.; El Kouch, R.; Bellafkih, M. and Raouyane, B. (2011), “Functioning and Management of MPLS/QOS in the IMS architecture”, *2011 International Conference on Multimedia Computing and Systems (ICMCS)*, IEEE

123. Schnaidt, Konstantin (2009), “Realisierung eines Künstlichen Neuronalen Netzes zur Mustererkennung und -zuordnung von QoS-Parametern in SIP-basierten NGN” (translated title: “Realisation of an Artificial Neural Network

for pattern recognition and classification of QoS parameters in SIP-based NGN"), Diploma Thesis, University of Applied Sciences Frankfurt a. M., Germany

124. Sengupta, S.; Chatterjee, M.; Ganguly, S. and Izmailov, R. (2006), "Improving R-Score of VoIP Streams over WiMax", *IEEE International Conference on Communications (ICC '06)*, Vol. 2, pp. 866-871, IEEE

125. Senthilkumar, L. and Sankaranarayanan, V. (2006), "Comparison of Endpoint Admission Control Algorithms by probing at Exponential Interval and at Constant Interval", *International Journal of Computer Science and Network Security (IJCSNS)*, Vol. 6, No. 12, pp. 192-196

126. Sfairopoulou, Anna; Bellalta, Boris and Macián, Carlos (2008), "How to Tune VoIP Codec Selection in WLANs?", *IEEE communications letters*, Vol. 12, No. 8, pp. 551-553, IEEE

127. Sfairopoulou, A.; Bellalta, B.; Macián, C. and Oliver, M. (2011), "A Comparative Survey of Adaptive Codec Solutions for VoIP over Multirate WLANs: A Capacity versus Quality Performance Trade-Off", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2011, Article ID 534520, Hindawi Publishing Corporation

128. Shih, Frank Y. (2010), *Image Processing and Pattern Recognition – Fundamentals and Techniques*, John Wiley & Sons, ISBN: 978-0-470-40461-4

129. Simmonds, Andrew and Nanda, Priyadarsi (2002), "Resource Management in Differentiated Services Networks", IFIP Interworking 2002, *Proceedings in 'Converged Networking: Data and Real-time Communications over IP'*, pp. 313-323, ed. C McDonald, pub. Kluwer Academic Publishers, ISBN: 1-4020-7379-8

130. Skorin-Kapov, L. and Matijasevicy, M. (2009), "A QoS Negotiation and Adaptation Framework for Multimedia Services in NGN", *10th International Conference on Telecommunications (ConTEL 2009)*, pp. 249-256, IEEE

131. Smoreda, Zbigniew and Licoppe, Christian (2000), "Gender-Specific Use of the Domestic Telephone", *Social Psychology Quarterly*, 2000, Vol. 63, No. 3, pp. 238-252, Sage Journals

132. Solís, M.; Benítez-Pérez, H.; Rubio, E.; Medina-Gómez, L.; Moreno, E.; Gonzalez, G. and Leija, L. (2008), "Pattern Classification of Decomposed Wavelet Information using ART2 Networks for echoes Analysis", *Journal of Applied Research and Technology*, Vol. 6, No. 1, pp. 33-44, National Autonomous University of Mexico, Mexico

133. Tebbani, Badis and Haddadou, Kamel (2008), "Codec-based Adaptive QoS Control for VoWLAN with Differentiated Services", *1st IFIP Wireless Days WD 2008*, IEEE

134. Tebbani, Badis; Haddadou, Kamel and Pujolle, Guy (2008), "Session-Based QoS Management Architecture for Wireless Local Area Networks", *LNCS*, Vol. 5275, pp. 117-126, Springer

135. Tebbani, Badis; Haddadou, Kamel and Pujolle, Guy (2009), "A Session-based Management Architecture for QoS Assurance to VoIP Applications on Wireless Access Networks", *6th IEEE Consumer Communications and Networking Conference (CCNC 2009)*, IEEE

136. Theodoridis, Sergios and Koutroumbas, Konstantinos (2009), *Pattern recognition*, 4th edition, Academic Press, ISBN: 978-1-59749-272-0

137. Toral, H.; Torres, D.; Hernández, C. and Estrada, L. (2008), "Self-Similarity, Packet Loss, Jitter, and Packet Size: Empirical Relationships for VoIP", *18th International Conference on Electronics, Communications and Computers*, pp. 11-16, IEEE

138. Trick, Ulrich and Weber, Frank (2004), *SIP, TCP/IP und Telekommunikationsnetze*, 1st edition, Oldenbourg, Munich, Germany, ISBN: 3-486-27529-1

139. Trick, Ulrich and Weber, Frank (2009), *SIP, TCP/IP und Telekommunikationsnetze*, 4th edition, Oldenbourg, Munich, Germany, ISBN: 978-3-486-59000-5

140. Tüysüz, Mehmet Fatih and Mantar, Hacı Ali (2010), "A Cross Layer QoS Algorithm to Improve Wireless Link Throughput and Voice Quality over Multi-rate WLANs", *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference (IWCMC '10)*, pp. 209-213, ACM

141. University of Southern California (2011), "The Network Simulator – ns-2", http://www.isi.edu/nsnam/ns/ (last visited: 2012-04-25)

142. University of Tübingen (2008), "SNNS – Stuttgart Neural Network Simulator", http://www.ra.cs.uni-tuebingen.de/SNNS/ (last visited: 2012-04-25)

143. Vautier, M.; Fromentoux, G.; Hartrisse, X. and Vinel, R. (2002), "Resources control and QoS implementation in a NGN DSL access network", *2nd European Conference on Universal Multiservice Networks (ECUMN),* pp. 305-314, IEEE

144. Vidal, I.; Garcia, J.; Valera, F.; Soto, I. and Azcorra, A. (2007), "Integration of a QoS Aware End User Network within the TISPAN NGN Solutions", *Fourth European Conference on Universal Multiservice Networks (ECUMN '07)*, pp. 152-162, IEEE

145. W3C SOAP specifications (2007), "SOAP Version 1.2", accessable via http://www.w3.org/TR/soap (last visited at 2011-08-22)

146. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2007), "QoS in SIP-based NGN – state of the art and new requirements", *Proceedings of the third collaborative research symposium on Security, E-learning, Internet*

and Networking (SEIN 2007), pp. 201-214, Information Security & Network Research Group – University of Plymouth, Plymouth, UK, ISBN: 978-1-8410-2173-7

147. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2008), "A framework for improved QoS evaluation and control in SIP-based NGN", *Proceedings of the Seventh International Network Conference (INC 2008)*, pp. 27-37, University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-188-1

148. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2009a), "AI-based QoS profiling for NGN user terminals", *Proceedings of the third International conference on Internet Technologies and Applications (ITA 09)*, pp. 539-548, Centre for Applied Internet Research – Glyndŵr University, Wrexham, UK, ISBN: 978-0-946881-65-9

149. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2009b), "Applying and validating AI-based QoS profiling for NGN user terminals", *Proceedings of the fifth collaborative research symposium on Security, E-learning, Internet and Networking (SEIN 2009)*, pp. 205-215, Centre for Security, Communications & Network Research – University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-236-9

150. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2010), "A Bootstrap Mechanism for NGN QoS Profiling", *Proceedings of the eighth International Network Conference (INC 2010)*, pp. 61-70, University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-259-8

151. Weber, Frank and Trick, Ulrich (2008), "Optimizing and simplifying SIP-based NGNs' QoS architecture", *International SIP 2008 Conference*, upperside, Paris, France

152. Welzl, Michael and Mühlhäuser, Max (2003), "Scalability and Quality of Service: A Trade-off?", *Communications Magazine*, Vol. 41, Issue 6, pp. 32-36, IEEE

153. White, Christopher M.; Raymond, J. and Teague, K.A. (2004), "A real-time network simulation application for multimedia over IP", *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, Vol. 2, pp. 2245-2249, IEEE

154. Xu, Rui and Wunsch, Don (2008): *Clustering*, Wiley-IEEE Press, ISBN: 978-0470276808

155. Yang, Xu; Bigham, John and Cuthbert, Laurie (2005), "Resource Management for Service Providers in Heterogeneous Wireless Networks", *Wireless Communications and Networking Conference*, Vol. 3, pp. 1305-1310, IEEE

156. Zhang, Hongli; Gu, Zhimin and Tian, Zhenqing (2011), "QoS Evaluation Based on Extend E-Model in VoIP", *13th International Conference on Advanced Communication Technology (ICACT 2011)*, pp. 852-854, IEEE

157. Zheng, Li; Zhang, Liren and Xu, Dong (2001), "Characteristics of network delay and delay jitter and its effect on voice over IP (VoIP)", *IEEE*

International Conference on Communications (ICC 2001), Vol. 1, pp. 122-126, IEEE

158. Zhou, Yong; Xiao, Xiaojun; Du, Chunsheng and Zhou, Jing A. (2006), “Field trial of end-to-end QoS control based on RACS”, *10th IEEE Singapore International Conference on Communication systems (ICCS 2006)*, IEEE

Appendix A – Abbreviations

3GPP	Third Generation Partnership Project
3SQM	Single Sided Speech Quality Measure
A	
a	Answerer
A-	Access
ACELP	Algebraic Code-Excited Linear Prediction
AF	Application Functions
ALG	Application Layer Gateway
AN	Access Network
ANN	Artificial Neural Network
ART	Adaptive Resonance Theory
ATM	Asynchronous Transfer Mode
B	
B2BUA	Back-to-Back User Agent
BE	Best Effort
BGF	Border Gateway Function
BTF	Basic Transport Function
B/W	Bandwidth
C	
C-	Core
CAC	Call Admission Control
CN	Core Network
CND	Customer Network Device
CQE	Conversational Quality Estimated
CS	Call Server

D

DiffServ	Differentiated Services
DOCSIS	Data-Over-Cable Service Interface Specifications
DSL	Digital Subscriber Line
DS-TE	DiffServ-aware MPLS Traffic Engineering

E

EAC	Endpoint Admission Control
EFR	Enhanced Full Rate
ETNO	European Telecommunications Network Operators' Association
ETSI	European Telecommunications Standards Institute
EuQoS	End-to-end Quality of Service support over heterogeneous networks

F

FR	Full Rate

G

GoS	Grade of Service
GP	Genetic Programming
GSM	Global System for Mobile communications

H

HFC	Hybrid Fiber Coax
HR	Half Rate

I

i	Initiator
IAD	Integrated Access Device
IETF	Internet Engineering Task Force
iLBC	Internet Low Bitrate Codec
IMS	IP Multimedia Subsystem
IntServ	Integrated Services

IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
L	
LDAP	Lightweight Directory Access Protocol
LQO	Listening Quality Objective
LTM	Long Term Memory
M	
MAC	Medium Access Control
MGW	Media Gateway
Min	Minute
m.l.e.	Maximum Likelihood Estimator
MOS	Mean Opinion Score
MPLS	Multi-Protocol Label Switching
MP-MLQ	Multi-Pulse-Maximum Likelihood Quantization
N	
NAPT	Network Address and Port Translation
nC	New Codec
NETCONF	Network Configuration Protocol
NGN	Next Generation Networks
NP	Network Performance
NS-2	Network Simulator 2
NSIS	Next Steps In Signalling
NTP	Network Time Protocol
O	
oC	Original Codec
OCR	Optical Character Recognition

P

PCMA	Pulse Code Modulation a-law
PCMU	Pulse Code Modulation μ-law
PESQ	Perceptual Evaluation of Speech Quality
PR	Pattern Recognition
PSTN	Public Switched Telephone Network
PUT	Pattern Under Test

Q

Q	QoS Manager
QoE	Quality of Experience
QoS	Quality of Service

R

RACF	Resource and Admission Control Function
RACS	Resource and Admission Control Subsystem
RCEF	Resource Control Enforcement Function
Ref.	Reference
RGW	Residential Gateway
RSVP	Resource Reservation Protocol
RTCP	RTP Control Protocol
RTP	Real-time Transport Protocol

S

SBC	Session Border Controller
SCTP	Stream Control Transmission Protocol
SDP	Session Description Protocol
SF	Service and call control Functions
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SNNS	Stuttgart Neural Network Simulator
SOA	Service-Oriented Architecture

SOM	Self-Organising Maps
SPDF	Service-based Policy Decision Function
STM	Short Term Memory

T

TCP	Transmission Control Protocol
TLS	Transport Layer Security
tu	Time Unit

U

UA	User Agent
UAG	User Access Gate
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
URI	Uniform Resource Identifier
UTRAN	Universal Terrestrial Radio Access Network

V

VLAN	Virtual Local Area Network
VoD	Video on Demand
VoIP	Voice over IP

W

WLAN	Wireless Local Area Network

X

x-	(generic)
XR	Extended Reports

Appendix B – Further SIP extensions and mechanisms

- The SIP preconditions framework, as specified by (IETF RFC 3312, 2002) and generalised by (IETF RFC 4032, 2005), was defined to allow for the negotiation of session preconditions in line with the SIP session initiation process. The original goal of this framework was to ensure that the called party user is alerted only if both involved user terminals have successfully reserved resources for media exchange over the transport network, after they have agreed that this was a desired or mandatory condition for the session. The reservation could have been performed by utilising IntServ/RSVP (see section 3.2.2). In any case, the agreement for and the success of the reservation of network resources are announced by respective SDP attributes inline with SDP offer/answer exchange between the involved user terminals. Since the SIP preconditions framework has been generalised by (IETF RFC 4032, 2005), it is able to handle arbitrary preconditions, such as properties of media encryption, as defined in (IETF RFC 5027, 2007). The SIP messages PRACK (IETF RFC 3262, 2002) and UPDATE (IETF RFC 3311, 2002) are associated with the SIP preconditions framework.

- Instant Messaging (IETF RFC 3428, 2002): The SIP MESSAGE method is used to carry short text messages in a simple way from one SIP entity to another. Note that instant messaging is not based on connection-oriented communication states and, therefore, does not rely on SIP session establishment procedures.

- Symmetric SIP Response Routing (IETF RFC 3581, 2003): This extension provides improved NAPT (Network Address and Port Translation) and Firewall traversal for SIP signalling.

- Security functions: Amongst others SIP supports digest authentication, TLS-based (Transport Layer Security) hop-by-hop encryption for SIP messages, and end-to-end encryption for SIP message bodies. The most relevant security functionalities for SIP have been basically described in (IETF RFC 3261, 2002). Several extensions and clarifications to SIP security functions have been defined in additional standards.

Appendix C – Publications and Presentations

The following list includes publications and presentations related to the area of this research, to which the author of this thesis has contributed during the course of research.

1. Trick, Ulrich and Weber, Frank (2009), *SIP, TCP/IP und Telekommunikationsnetze*, 4th edition, Oldenbourg, Munich, Germany, ISBN: 978-3-486-59000-5

2. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2007), "QoS in SIP-based NGN – state of the art and new requirements", *Proceedings of the third collaborative research symposium on Security, E-learning, Internet and Networking (SEIN 2007)*, pp. 201-214, Information Security & Network Research Group – University of Plymouth, Plymouth, UK, ISBN: 978-1-8410-2173-7

3. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2008), "A framework for improved QoS evaluation and control in SIP-based NGN", *Proceedings of the Seventh International Network Conference (INC 2008)*, pp. 27-37, University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-188-1

4. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2009a), "AI-based QoS profiling for NGN user terminals", *Proceedings of the third International conference on Internet Technologies and Applications (ITA 09)*, pp. 539-548, Centre for Applied Internet Research – Glyndŵr University, Wrexham, UK, ISBN: 978-0-946881-65-9

5. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2009b), "Applying and validating AI-based QoS profiling for NGN user terminals", *Proceedings of the fifth collaborative research symposium on Security, E-learning, Internet and Networking (SEIN 2009)*, pp. 205-215, Centre for

Security, Communications & Network Research – University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-236-9

6. Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U. and Ghita, B. (2010), "A Bootstrap Mechanism for NGN QoS Profiling", *Proceedings of the eighth International Network Conference (INC 2010)*, pp. 61-70, University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-259-8

7. Weber, Frank and Trick, Ulrich (2007), Poster: "QoS in SIP-based NGN – introducing fundamental requirements and a new approach", Joint EuroFGI and ITG Workshop on 'Visions of Future Generation Networks', 7th Wuerzburg Workshop on IP, University of Wuerzburg, Germany

8. Weber, Frank and Trick, Ulrich (2008), "Optimizing and simplifying SIP-based NGNs' QoS architecture", *International SIP 2008 Conference*, upperside, Paris, France

9. Abu Salah, Stefan; Brack, Stephan; Grebe, Andreas; Marikar, Achim; Trick, Ulrich and Weber, Frank (2008), "BMBF-Forschungsprojekt: Verbesserung der netzeübergreifenden Quality of Service bei SIP-basierter VoIP-Kommunikation (QoSSIP) – Abschlussbericht" (translated title: "Public-funded research project: Improvement of inter-network Quality of Service in SIP-based VoIP communication (QoSSIP) – Final report"), BMBF Project Funding Reference Numbers 1715A05 (FH Köln) und 1715B05 (FH Frankfurt), Universities of Applied Sciences FH Köln, Germany and FH Frankfurt a. M., Germany

Copies of the papers most closely related to the research described are enclosed within this appendix.

Published in *Proceedings of the Seventh International Network Conference (INC 2008)*, pp. 27-37, University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-188-1

A framework for improved QoS evaluation and control in SIP-based NGN

F.Weber[1,2], W.Fuhrmann[3], U.Trick[2], U.Bleimann[3], B.Ghita[1]

[1] Network Research Group, University of Plymouth, Plymouth, United Kingdom
[2] Research Group for Telecommunication Networks, University of Applied Sciences Frankfurt/M., Frankfurt/M., Germany
[3] University of Applied Sciences Darmstadt, Darmstadt, Germany
e-mail: weber@e-technik.org

Abstract

Today's standardised approaches for the control of QoS (Quality of Service) in NGN (Next Generation Networks) come along with a high volume of additional, unscalable traffic for the allocation and reservation of network resources within the IP transport network. This paper describes a new framework for comprehensive QoS control in SIP-based (Session Initiation Protocol) NGN, addressing the shortcomings of standardised NGN QoS provision and thus, leading to a more efficient QoS provision model. Because this framework approach does not rely on traditional IP QoS mechanisms it can be applied to arbitrary combinations of IP transport technologies. In order to save up network resources and, at the same time, provide appropriate service performance to the subscribers multiple factors (such as individual QoS requirements of different media and codecs) are considered to perform scalable end-to-end QoS monitoring, rating, and control.

Keywords

NGN, SIP, QoS, IP, admission control, framework for comprehensive QoS control

1. Introduction

The concept of NGN as defined by ITU-T NGN GSI (International Telecommunication Union – Telecommunication Standardization Sector NGN Global Standards Initiative) and ETSI TISPAN (European Telecommunications Standards Institute Telecoms & Internet converged Services & Protocols for Advanced Network) can be outlined by several main key features (Trick and Weber, 2007); (ITU-T Y.2001, 2004), one of which is the provision of Quality of Service (QoS).

Traditional ways to provide QoS in IP transport networks are usually based on mechanisms such as IntServ/RSVP (Integrated Services / Resource Reservation Protocol) (IETF RFC 1633, 1994) (IETF RFC 2205, 1997), DiffServ (Differentiated Services) (IETF RFC 2474, 1998) (IETF RFC 2475, 1998), or MPLS (Multi-Protocol Label Switching) (IETF RFC 3031, 2001). However, all these mechanisms come with individual characteristics that, depending on the overall design and dimension of a respective NGN architecture, potentially lead to inefficient or insufficient QoS provision, respectively. IntServ/RSVP, while supporting precise end-to-end per-flow service provisioning, lacks of scalability (Simmonds and Nanda, 2002), (Welzl and Mühlhäuser, 2003), is considered potentially insecure (Giordano *et al.*, 2003), and, as mentioned in (Bohnert *et al.*, 2007), is not an adequate solution for the use in complex scenarios. On the other hand, the more scalable DiffServ mechanism, by definition providing only relative prioritisation for packets of selected data flows, is not efficient in dealing with network overload and is considered impractical for the use in networks that have to deal with a relative high volume of high priority traffic (Menth, 2006). Also the application of MPLS, which offers potentially beneficial properties to network operators in order to provide QoS in NGN, is limited related to scalability and efficiency (Bohnert *et al.*, 2007). Even though the combined application of at least two of the above-mentioned mechanisms generally leads to an improved relationship between scalability and efficiency, the individual issues of the mechanisms used can still have an effect (Welzl and Mühlhäuser, 2003), (Giordano *et al.*, 2003). Additionally, typical IP QoS approaches are (by definition) not aware of communication sessions (e.g., a Voice over IP session) established by higher layer protocols such as SIP. Therefore, within the standardised NGN QoS concept according to (ETSI ES 282 001, 2005) and (ITU-T Y.2001, 2004), a logical relationship between the SIP service control layer and the IP transport layer is created in order to provide QoS for the exchange of media session data between user end systems. Unfortunately, this technique comes along with a high volume of additional, unscalable traffic within the IP transport network (Park and Kang, 2005). Thus, today's standardised NGN approaches for QoS control led to inefficient resource management traffic.

2. Standardised NGN architecture and QoS provision

The following sections describe the standardised general NGN architecture, according to ETSI TISPAN and ITU-Ts NGN GSI, and its QoS provision concept.

2.1. Standardised NGN architecture

The standardised NGN architectures defined by ETSI TISPAN in (ETSI ES 282 001, 2005) and ITU-T in (ITU-T Y.2001, 2004) correspond to each other in all fundamental aspects. Both architectures can generally be divided into two strata, the service stratum (layered on-top) and the transport stratum below.

An NGN's service stratum consists of logical functional components, allowing a subscriber to use services and applications provided by the NGN, such as (in the simplest case) initiating a SIP telephony session with another subscriber. In terms of SIP, the service stratum comprises of a SIP server infrastructure.

The NGN's transport stratum provides IP connectivity and QoS-based IP transport to the user. The IP transport functions consist of any arbitrary IP transport infrastructure, including both access and core networks.

The user equipment (e.g., a SIP end system) is connected to an interface (e.g., a DSL-based access network (Digital Subscriber Line)) of the respective NGN's transport functions. By transmitting IP packets over this interface, the user equipment uses SIP to communicate with the NGN's service stratum (e.g., to setup media sessions to other users' end systems).

2.2. QoS provision in NGN

Once a service is requested by an NGN subscriber a top-down resource handling is performed in order to provide QoS for the respective service (such as a VoIP session). That is, if a service request originating from a user equipment enters the service stratum, resource and QoS requirements for the respective service are identified and handed down into the transport stratum where network resources are authorised, allocated, and reserved within the respective NGN's IP transport network. Depending on the IP transport technology (such as Ethernet, ATM, MPLS), the QoS mechanisms supported by the IP infrastructure (such as DiffServ or IntServ), and the network's dimension and topology, the steps required in order to provide QoS to media sessions implicate the use of a variety of different signalling protocols (such as RSVP for resource reservation), in conjunction with potentially required gateways.

3. Identified issues of NGN QoS provision and resulting requirements

As denoted in chapter 2, because of the fact that both the NGN's service stratum and the transport stratum are involved in the QoS provision process, signalling between these two strata is compulsory. Combined with the application of traditional IP QoS mechanisms (see chapter 1), this leads to an unpredictable volume of resource and QoS management traffic.

Note that the absolute amount of signalling effort in order to provide per-session QoS finally depends on the subscribers' session behaviour, because the standardised NGN QoS architecture works on a per-session basis. That is, a subscriber establishing and cancelling many sessions in a short period of time will require a substantial amount of resources and QoS management traffic.

Based on our earlier research work, the following main requirements for the provision of QoS in SIP-based NGN have been identified (Weber *et al.*, 2007).

- Functions and mechanisms, leading to a trustworthy QoS for each established session and, at the same time, do not occupy resources on a per-session basis themselves.
- Simple and resource saving QoS control should be preferred. If possible, approaches should rely on already standardised protocols (such as SIP) and architectures (such as NGN according to (ITU-T Y.2012, 2006) and (ETSI ES 282 001, 2005)).
- NGN QoS control has to be aware of a certain amount of traffic that is not session-based (such as TCP web traffic).
- The QoS provision in NGN should be independent of underlying transport technologies such as MPLS, ATM, and VLAN. Arbitrary IP network architectures should be supported, regardless their specific integrated QoS mechanisms.

4. Integrated framework for comprehensive QoS control in SIP-based NGN

Based on our research a framework has been created that fulfils the requirements stated in chapter 3 and, hence, solves the identified issues of the standardised NGN QoS. The framework combines several QoS provision requirements within telecommunication networks. As a result it leads to a comprehensive solution that can be implemented into the standardised NGN architecture according to (ETSI ES 282 001, 2005) and (ITU-T Y.2012, 2006) with no or minor changes, depending on the arrangement of the respective NGN's logical components.

4.1. Framework tasks

This subsection specifies the tasks to be fulfilled by the integrated framework for comprehensive QoS control in SIP-based NGN.

The main prerequisite to satisfy the requirements stated in section 3 is that any task fulfilled by the integrated framework for comprehensive QoS control causes only a minimum of additional traffic (such as QoS signalling traffic). Additionally the fulfilment of the tasks must lead to scalability of the QoS control framework with the number of subscribers and their behaviour, and with network characteristics (such as IP transport technology, IP QoS mechanisms, and network topology).

In order to satisfy the requirements stated in chapter 3, the integrated framework for comprehensive QoS control in SIP-based NGN is designed to fulfil the following tasks.

- Task 1: QoS measurement - Periodical evaluation (and near future prediction) of QoS conditions on every relevant segment of the transport network (e.g. between two access networks)
- Task 2: Advanced Admission Control - Integration with admission control for SIP-based services (Advanced Admission Control)
- Task 3: Manipulation of QoS conditions within the IP network (QoS control)

4.2. Framework overview and elements

Figure 1 gives an overview of the integrated framework for comprehensive QoS control in SIP-based NGN.

The transport network of the NGN, used as an example for the framework implementation shown in figure 1, consists of an IP-based NGN core network and several IP access networks. Note that the framework implementation is independent of the IP transport technology (such as Ethernet or WLAN in the access and ATM or MPLS in the core, respectively). Further on, the framework is independent of any arbitrary QoS mechanism supported by core or access networks.

For session control, in the simplest case, a SIP call server is provided by a SIP service provider. Note that SIP session control could also be performed by an IMS (IP Multimedia Subsystem) as defined by 3GPP (Third Generation Partnership Project) (3GPP TS 23.228, 2006). The SIP session control functions are usually connected to the IP core network.

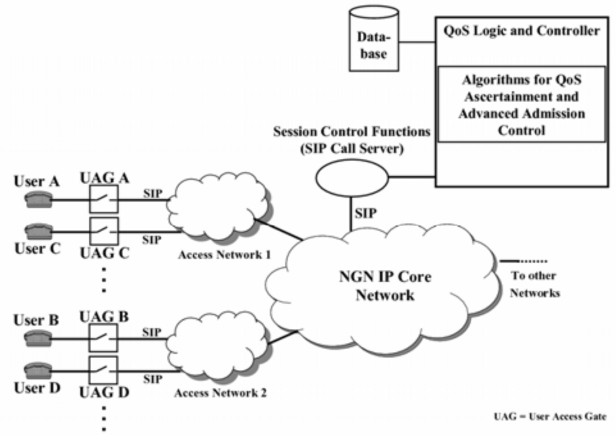


Figure 1: Integrated framework for comprehensive QoS control in SIP-based NGN (Weber and Trick, 2008)

To provide a basic NGN with the integrated framework for comprehensive QoS control, the following elements have to be added.

- User Access Gates (UAGs): This entity can be located either at the subscribers' residence or at the border of the access network. It works as a mediation device for all data exchanged between a user end system and the network and covers the functionalities shown in figure 2. Note that the UAG must be trusted by the SIP service provider.

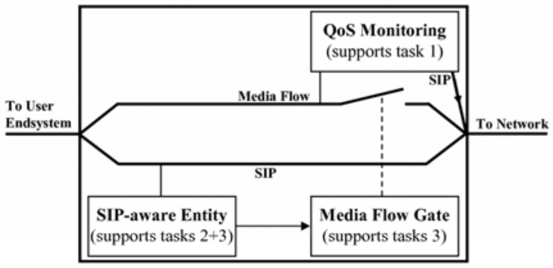

QoS Monitoring (supports task 1)
SIP
Media Flow
To User Endsystem
To Network
SIP
SIP-aware Entity (supports tasks 2+3)
Media Flow Gate (supports tasks 3)

Figure 2: User Access Gate (UAG) block diagram (Weber and Trick, 2008)

- QoS Logic and Controller (QoS L&C): This centralised entity has to be provided with interfaces to the SIP session control function and to a database. It covers the functionalities shown in figure 3.

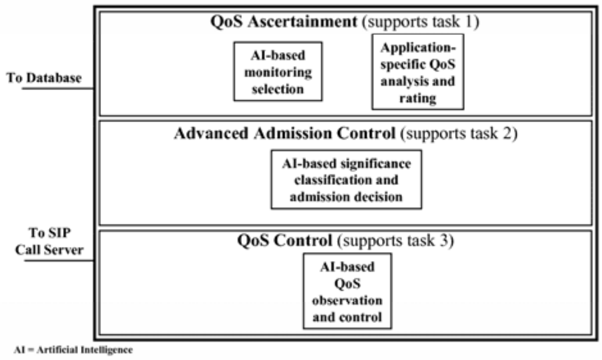

QoS Ascertainment (supports task 1)
AI-based monitoring selection
Application-specific QoS analysis and rating
To Database
Advanced Admission Control (supports task 2)
AI-based significance classification and admission decision
To SIP Call Server
QoS Control (supports task 3)
AI-based QoS observation and control
AI = Artificial Intelligence

Figure 3: QoS Logic and Controller (QoS L&C) block diagram (Weber and Trick, 2008)

Within the following chapters, the UAG and QoS L&C entities are described in the context of the overall framework.

4.3. Framework basic principle

As denoted in chapter 2, the standardised NGN QoS architecture makes use of QoS and resource control within and below the IP layer, applying a top-down QoS control scheme. This leads to the inefficiencies and the lack of scalability mentioned in chapter 3.

The integrated framework for comprehensive QoS control does not deal with active QoS control on or below the IP layer. It is assumed that the arbitrary IP core and access networks within the NGN may provide best effort IP transport. Therefore, two subscribers A and C, with the same connectivity and access network (see figure 1), will experience the same QoS when connecting to two other subscribers, B and D, from a remote network, providing that both sessions use the same parameters (i.e. both sessions are based on the same medium and use the same codecs). In this case, in order to collect information about the QoS experienced by subscribers A and C, it is sufficient to obtain QoS information for one of these two sessions.

4.4. Framework functionality fulfilling task 1

In the simplest case, the information provided by the QoS Monitoring functionality included in the UAG (see figure 2) comprises jitter and packet loss rates for the IP packets carrying the media data to the user end system. This information can be obtained by the QoS L&C, prompting the SIP Call Server to query the UAG for the QoS information of an ongoing media session.

To save IP transport capacity the QoS Ascertainment functionality included within the QoS L&C (see figure 3) keeps track of all ongoing sessions in order to select carefully the UAGs that have to be queried for QoS information (note that comparable sessions under comparable conditions between comparable subscribers will experience the same QoS). The QoS Ascertainment functionality makes use of artificial intelligence (AI) in order to learn which UAGs are queried best in order to obtain a sufficient overview of the QoS between several access networks, considering several different media and codecs, and to minimise the number of queried UAGs.

The relative QoS information retrieved from each queried UAG is matched with the QoS L&C through QoS matrixes, providing information about QoS values (such as jitter and packet loss rates) tolerated by the respective medium and codec. This matching leads to the information about the absolute QoS experienced by the respective subscriber. The QoS matrixes are stored within a database that can be accessed by the QoS L&C.

By continuously keeping record of the QoS experienced for several sessions of different media between different subscribers, the QoS Ascertainment functionality of the QoS L&C creates timeline-based QoS profiles for every network segment (e.g. between two different access networks). The timeline-based QoS profiles can be used for near-future QoS prediction by other sub-functionalities of the QoS L&C.

4.5. Framework functionality fulfilling task 2

When a session is requested with or by a subscriber of a NGN, a SIP request is sent to the SIP session control functions. In order to consider the given QoS situation for the admission control, the SIP Call Server provides all relevant information about the requested session (such as the involved subscribers' identities, type of media and codecs, affected network segments) to the Advanced Admission Control functionality of the QoS L&C (see figure 3). The Advanced Admission Control functionality performs the following steps in order to decide how the session request has to be handled.

- Step 1: Media session significance classification. The objective significance of each media session is classified considering the following parameters, specific to the involved subscribers.
  - General subscriber policy criteria (e.g. premium versus basic customer)
  - Formerly experienced session availability ratio (policy-assured/granted) per medium per subscriber

  If applicable, further parameters may be taken into account for significance classification. A ranking list of significance for all ongoing and requested sessions is continuously kept and updated.
- Step 2: Identification of QoS conditions required for media session. By matching the requested media and codecs with QoS matrixes stored within the database, the Advanced Admission Control identifies QoS parameter values (such as jitter and packet loss rate) that will be tolerated by the respective media and codec.

- Step 3: Identification of network segments of the path between the endpoints.
- Step 4: Analysis of current and prospective near-future QoS conditions within the affected network segments. The Advanced Admission Control functionality obtains this information from the QoS Ascertainment functionality.
- Step 5: Admission Decision. Taking into account all relevant information derived from the performance of the steps 1 to 4, the Advanced Admission Control functionality makes a decision about the requested media sessions. The following results are possible.
    - The session will be rejected
    - The session will be granted as-is
    - The session will be granted under QoS downgrade conditions (e.g. a low-bitrate codec has to be used)

  In order to consider the ranking list of significance (see step 1), in case that a concerning incoming session request potentially would not experience satisfying QoS conditions if granted (due to high traffic volume in the network) the Advanced Admission Control can also decide to reject or downgrade competing media sessions that are objectively less relevant. For the achievement of the goal to cancel or reject as less (objectively least crucial) sessions as possible in order to maintain or recover sufficient QoS conditions for as much (objectively more crucial) sessions as possible, Artificial Intelligence (AI) will be investigated as a potential alternative for mathematical calculation as a basis for the Advanced Admission Control functionality of the QoS L&C.

In any case the admission decision is provided to the SIP Call Server in order to react adequately on the SIP session request and, if applicable, cancel or downgrade concurrent sessions. On its way to user end systems, the SIP signalling originated from the SIP Call Server is identified and recognised by the SIP-aware Entity within the UAGs (see figure 2) of the respective subscribers. The UAGs take responsibility that the subscribers' user end systems follow the signalled directives (e.g. codec downgrade or session rejection/cancellation). The SIP-aware Entity controls the UAGs' Media Flow Gate functionalities (see figure 2) (e.g. for bandwidth limitation or media flow cut-off) in order to enforce the QoS L&C's decision.

4.6. Framework functionality fulfilling task 3

As stated in chapter 4.4 the QoS Ascertainment functionality of the QoS L&C collects, derives and arranges particular information of the QoS experienced by subscribers connected to several access networks, exchanging several kinds of media coded with several codecs. This information can be accessed by the QoS Control functionality of the QoS L&C (see figure 3) in order to continuously observe the overall QoS provided for several network segments. Also the timeline-based QoS profiles are considered by the QoS Control functionality in order to predict near-future trends of the overall QoS.

In case a potential or yet existing shortage of QoS is detected for one or several ongoing media sessions (e.g. in a high traffic period), the QoS Control functionality has to react in order to maintain or recover satisfying QoS conditions for as much high priority media sessions as possible. Hence, the QoS Control functionality can decide to cancel or downgrade concurrent media sessions that are objectively less relevant. Therefore, the ranking list of media session significance (see chapter 4.5, step 1) is taken into account. The action that has to be performed to make decisions, and the involvement of SIP call server and UAGs in order to execute and control codec downgrades and SIP session cancellations are similar as described within step 5 of chapter 4.5, and the paragraph below step 5, respectively.

5. Conclusions and Future Work

This paper proposed an integrated framework for comprehensive QoS control in SIP-based NGN. In contrast to the standardised NGN QoS architecture, this framework does not rely on top-down network resource allocation, but makes use of algorithms and artificial intelligence (AI) in order to provide sufficient QoS conditions for as much objectively crucial sessions as possible. This is achieved without any impact on the IP transport network or on any QoS mechanisms provided by the IP layer or the underlying IP transport technology.

The next step in order to bring the framework to implementation state will be the definition of the basis (such as neural networks) for the required artificial intelligence, and define the respective algorithms. Further, a prototype of the framework will be planned and implemented into an existing NGN under laboratory conditions. In future the framework may be extended to the active manipulation of the volume of non-session-based traffic (such as TCP web traffic) carried over the network in order to provide sufficient QoS to media sessions. For this the SIP-based control of the UAGs may be a key functionality.

6. Annotation

The research project providing the basis for this publication was partially funded by the Federal Ministry of Education and Research (BMBF) of the Federal Republic of Germany under grant number 1715B05. The authors of this publication are in charge of its content.

7. References

3GPP TS 23.228 (2006), Technical Specification, “IP Multimedia Subsystem (IMS); Stage 2 (Release 5)”, 3GPP

Thomas Michael Bohnert, Edmundo Monteiro, Marilia Curado, Alexandre Fonte, Michal Ries, Dmitri Moltchanov, and Yevgeni Koucheryavy (2007), "Internet Quality of Service: A Bigger Picture", 1st OpenNet Workshop - Service Quality and IP Network Business: Filling the Gap, Diegem/Brussels, Belgium, March 2007

ETSI ES 282 001 V1.1.1 (2005), ETSI Standard, “NGN Functional Architecture Release 1”, ETSI TISPAN

ETSI TS 185 001 V1.1.1 (2005), Technical Specification, “NGN Quality of Service (QoS) Framework and Requirements”, ETSI TISPAN

Silvia Giordano, Stefano Salsano, Steven Van den Berghe, Giorgio Ventre, and Dimitrios Giannakopoulos (2003), „Advanced QoS Provisioning in IP Networks: The European Premium IP Projects", *Communications Magazine*, IEEE, Volume 41, Issue 1, January 2003, pp. 30 - 36

IETF RFC 1633 (1994), Request For Comments, "Integrated Services in the Internet Architecture: an Overview", IETF

IETF RFC 2205 (1997), Request For Comments, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", IETF

IETF RFC 2474 (1998), Request For Comments, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", IETF

IETF RFC 2475 (1998), Request For Comments, "An Architecture for Differentiated Services", IETF

IETF RFC 3031 (2001), Request For Comments, "Multiprotocol Label Switching Architecture", IETF

ITU-T Y.2001 (2004), Recommendation, "General overview of NGN", ITU-T

ITU-T Y.2012 (2006), Recommendation, "Functional requirements and architecture of the NGN release 1", ITU-T

Michael Menth (2006), „Das echtzeitfähige und ausfallsichere Internet der nächsten Generation – Next Generation Networks" (The real-time capable and fail-proof Internet of the next generation – Next Generation Networks), lecture series on ‚The Internet and its applications', BTU Cottbus, Germany, 25 April 2006

Juyoung Park and Shin Gak Kang (2005), "QoS Architecture for NGN", *Advanced Communication Technology, ICACT 2005*, pp. 1064-1067, IEEE, July 2005

Andrew Simmonds and Priyadarsi Nanda (2002), „Resource Management in Differentiated Services Networks", IFIP Interworking 2002, *Proceedings in 'Converged Networking: Data and Real-time Communications over IP'*, pp. 313 - 323, ed. C McDonald, pub. Kluwer Academic Publishers, ISBN: 1-4020-7379-8, 2003

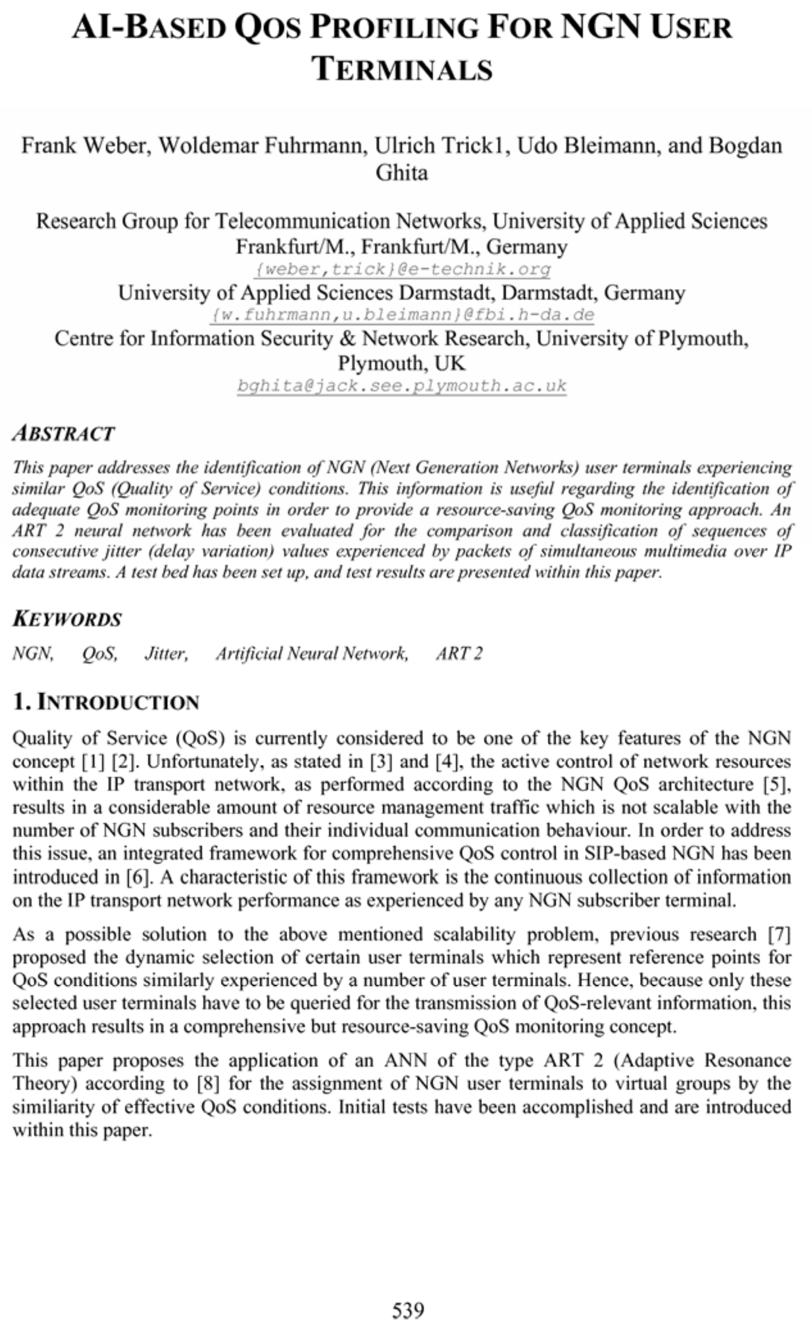
U. Trick and F. Weber, (2007), *SIP, TCP/IP und Telekommunikationsnetze*, Oldenbourg, Munich, Germany, ISBN: 978-3-486-58228-4

F. Weber, W. Fuhrmann, U. Trick, U. Bleimann, B. Ghita, (2007), "QoS in SIP-based NGN – state of the art and new requirements", *Proceedings of the third collaborative research symposium on Security, E-learning, Internet and Networking SEIN 2007*, pp. 201-214, Information Security & Network Research Group – University of Plymouth, Plymouth, ISBN: 978-1-8410-2173-7

F. Weber and U. Trick, (2008), "Optimizing and simplifying SIP-based NGNs' QoS architecture", International SIP 2008 Conference, Paris, France

Michael Welzl and Max Mühlhäuser (2003), „Scalability and Quality of Service: A Trade-off?", *Communications Magazine*, IEEE, Volume 41, Issue 6, June 2003, pp. 32 - 36

Published in *Proceedings of the third International conference on Internet Technologies and Applications (ITA 09)*, pp. 539-548, Centre for Applied Internet Research – Glyndŵr University, Wrexham, UK, ISBN: 978-0-946881-65-9



AI-Based QoS Profiling For NGN User Terminals

Frank Weber, Woldemar Fuhrmann, Ulrich Trick1, Udo Bleimann, and Bogdan Ghita

Research Group for Telecommunication Networks, University of Applied Sciences Frankfurt/M., Frankfurt/M., Germany
{weber,trick}@e-technik.org
University of Applied Sciences Darmstadt, Darmstadt, Germany
{w.fuhrmann,u.bleimann}@fbi.h-da.de
Centre for Information Security & Network Research, University of Plymouth, Plymouth, UK
bghita@jack.see.plymouth.ac.uk

Abstract

This paper addresses the identification of NGN (Next Generation Networks) user terminals experiencing similar QoS (Quality of Service) conditions. This information is useful regarding the identification of adequate QoS monitoring points in order to provide a resource-saving QoS monitoring approach. An ART 2 neural network has been evaluated for the comparison and classification of sequences of consecutive jitter (delay variation) values experienced by packets of simultaneous multimedia over IP data streams. A test bed has been set up, and test results are presented within this paper.

Keywords

NGN, QoS, Jitter, Artificial Neural Network, ART 2

1. Introduction

Quality of Service (QoS) is currently considered to be one of the key features of the NGN concept [1] [2]. Unfortunately, as stated in [3] and [4], the active control of network resources within the IP transport network, as performed according to the NGN QoS architecture [5], results in a considerable amount of resource management traffic which is not scalable with the number of NGN subscribers and their individual communication behaviour. In order to address this issue, an integrated framework for comprehensive QoS control in SIP-based NGN has been introduced in [6]. A characteristic of this framework is the continuous collection of information on the IP transport network performance as experienced by any NGN subscriber terminal.

As a possible solution to the above mentioned scalability problem, previous research [7] proposed the dynamic selection of certain user terminals which represent reference points for QoS conditions similarly experienced by a number of user terminals. Hence, because only these selected user terminals have to be queried for the transmission of QoS-relevant information, this approach results in a comprehensive but resource-saving QoS monitoring concept.

This paper proposes the application of an ANN of the type ART 2 (Adaptive Resonance Theory) according to [8] for the assignment of NGN user terminals to virtual groups by the similiarity of effective QoS conditions. Initial tests have been accomplished and are introduced within this paper.

2. Next Generation Networks (NGN) and Quality of Service (QoS)

In 2004, the ITU-T (International Telecommunication Union – Telecommunication Standardization Sector) released its definition of NGN in [1]. According to [1], [2], and [9] the term NGN stands for a telecommunication network concept that can be characterised by a number of key features including, amongst others, "Packet-based data transport" and "Quality of Service support". Although the term "Packet-based data transport" does not refer to any particular technology or protocol, IP (Internet Protocol) is the most likely network protocol choice for an NGN environment according to [9], [10]. The use of SIP (Session Initiation Protocol) for NGN service provisioning and signalling is widely accepted, and also suggested in [11].

2.1. The NGN architecture

Figure 1 shows the principle structure of an NGN.

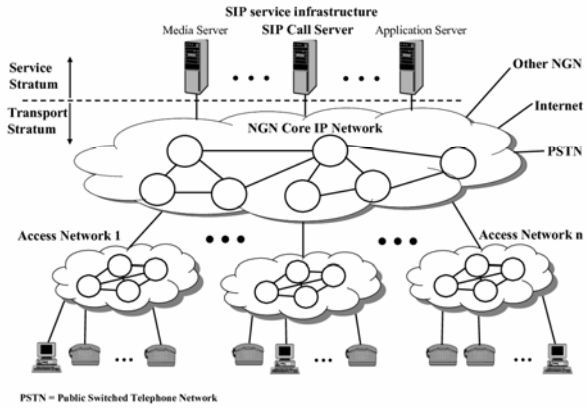
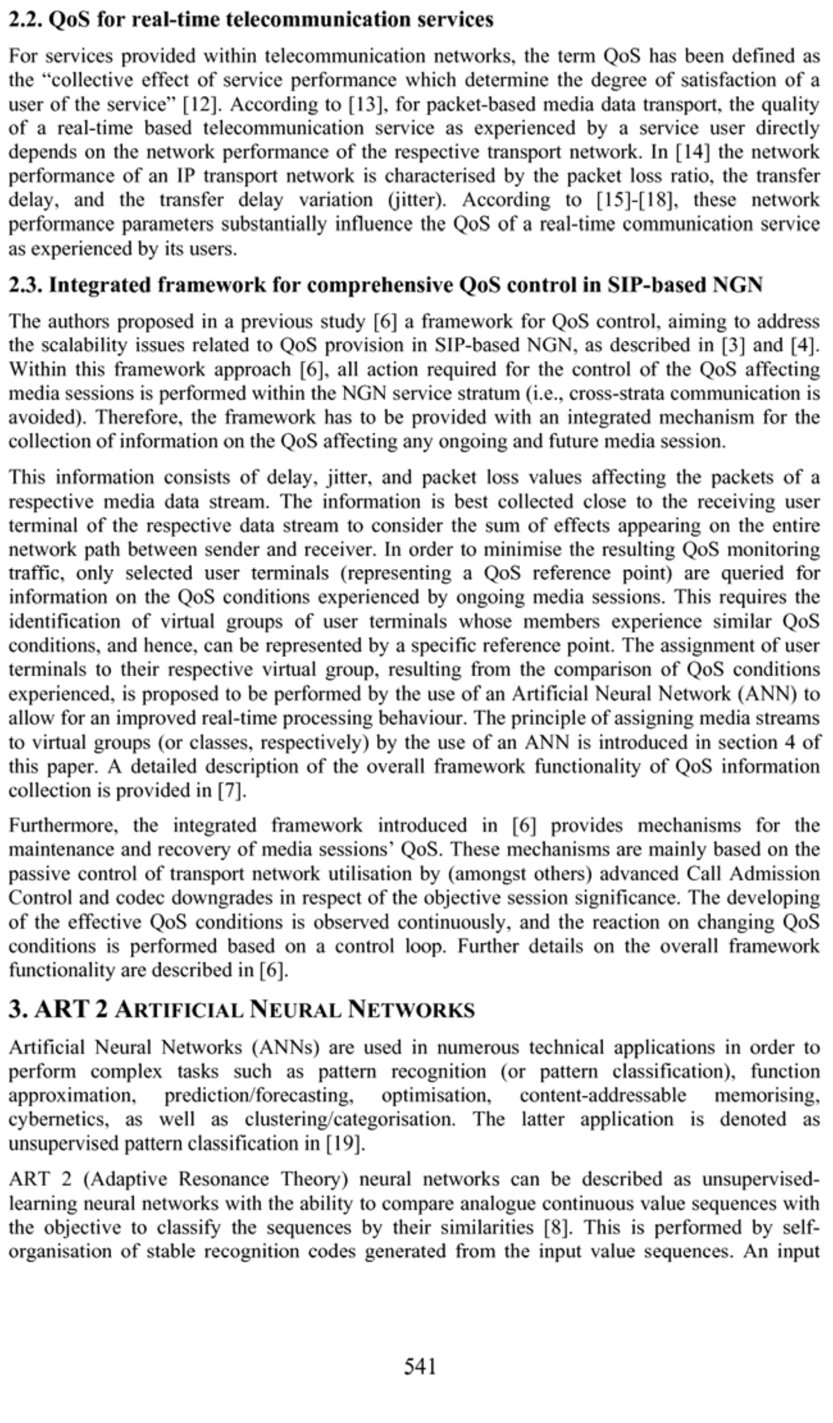


Figure 1. Principle structure of a SIP-based NGN

In the NGN architecture considered within this paper, SIP was chosen as the service signalling protocol. An NGN can be logically divided into a service stratum and a transport stratum. Utilising SIP for service control and signalling, an NGN's service stratum provides services and applications to the NGN subscribers, such as (in the simplest case) the initiation of a VoIP (Voice over IP) session. In terms of SIP, the service stratum comprises a SIP server infrastructure. The NGN's transport stratum provides IP connectivity and IP transport to the user terminals (such as VoIP phones). It consists of any arbitrary IP transport infrastructure, including both access and core networks. The user terminals are connected to interfaces (such as a DSL interface (Digital Subscriber Line)) provided by several access networks. By transmitting IP packets over this interface, the user equipment uses SIP to communicate with the NGN service stratum (e.g., to setup media sessions to other users' end systems). Once a media session is established, media data are exchanged peer-to-peer between the involved NGN user terminals. Hence, after session initiation, the service stratum is not involved in the media data exchange.



2.2. QoS for real-time telecommunication services

For services provided within telecommunication networks, the term QoS has been defined as the "collective effect of service performance which determine the degree of satisfaction of a user of the service" [12]. According to [13], for packet-based media data transport, the quality of a real-time based telecommunication service as experienced by a service user directly depends on the network performance of the respective transport network. In [14] the network performance of an IP transport network is characterised by the packet loss ratio, the transfer delay, and the transfer delay variation (jitter). According to [15]-[18], these network performance parameters substantially influence the QoS of a real-time communication service as experienced by its users.

2.3. Integrated framework for comprehensive QoS control in SIP-based NGN

The authors proposed in a previous study [6] a framework for QoS control, aiming to address the scalability issues related to QoS provision in SIP-based NGN, as described in [3] and [4]. Within this framework approach [6], all action required for the control of the QoS affecting media sessions is performed within the NGN service stratum (i.e., cross-strata communication is avoided). Therefore, the framework has to be provided with an integrated mechanism for the collection of information on the QoS affecting any ongoing and future media session.

This information consists of delay, jitter, and packet loss values affecting the packets of a respective media data stream. The information is best collected close to the receiving user terminal of the respective data stream to consider the sum of effects appearing on the entire network path between sender and receiver. In order to minimise the resulting QoS monitoring traffic, only selected user terminals (representing a QoS reference point) are queried for information on the QoS conditions experienced by ongoing media sessions. This requires the identification of virtual groups of user terminals whose members experience similar QoS conditions, and hence, can be represented by a specific reference point. The assignment of user terminals to their respective virtual group, resulting from the comparison of QoS conditions experienced, is proposed to be performed by the use of an Artificial Neural Network (ANN) to allow for an improved real-time processing behaviour. The principle of assigning media streams to virtual groups (or classes, respectively) by the use of an ANN is introduced in section 4 of this paper. A detailed description of the overall framework functionality of QoS information collection is provided in [7].

Furthermore, the integrated framework introduced in [6] provides mechanisms for the maintenance and recovery of media sessions' QoS. These mechanisms are mainly based on the passive control of transport network utilisation by (amongst others) advanced Call Admission Control and codec downgrades in respect of the objective session significance. The developing of the effective QoS conditions is observed continuously, and the reaction on changing QoS conditions is performed based on a control loop. Further details on the overall framework functionality are described in [6].

3. ART 2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are used in numerous technical applications in order to perform complex tasks such as pattern recognition (or pattern classification), function approximation, prediction/forecasting, optimisation, content-addressable memorising, cybernetics, as well as clustering/categorisation. The latter application is denoted as unsupervised pattern classification in [19].

ART 2 (Adaptive Resonance Theory) neural networks can be described as unsupervised-learning neural networks with the ability to compare analogue continuous value sequences with the objective to classify the sequences by their similarities [8]. This is performed by self-organisation of stable recognition codes generated from the input value sequences. An input sequence, also referred to as a pattern, is interpreted as an n-dimensional vector by an ART network, where n is the number of values comprised by the respective input pattern.

An ART 2 ANN provides n input units and m output units, the latter of which represent m individual output classes. If an arbitrary number of n-dimensional patterns is presented to an ART 2 network, after a predefined number of learning cycles, the network tries to map each pattern to one of m output classes by accomplishing a multi-step comparison process for each pattern. Patterns showing typical similarities are to be assigned to the same output class.

A number of setup parameters are provided by ART 2 ANN, of which the vigilance parameter ρ is the most effective. Depending on the degree of deviation in value patterns providing questionable similarities, the exactness of the assignment process can be influenced by this parameter. Further details on the theory of ART 2 neural networks can be found in [8].

4. Test Scenario

The following subsections provide the descriptions of the test bed used for the evaluation of an ART 2 neural network for the assignment of VoIP streams to classes by the similarity of their experienced QoS characteristics. From the classification of data streams, the classification of user terminals can be derived within an NGN service stratum.

4.1. Test bed setup and extraction of QoS characteristics

The purpose of the tests performed was to evaluate whether concurrent multimedia over IP packet streams exchanged within an NGN can be assigned to different classes, each representing certain QoS characteristics. The QoS characteristics are assumed to be mainly influenced by the QoS conditions effective within the access networks of both, the sending and the receiving host of a respective multimedia stream. In our experiments, the QoS characteristics of packet streams were found to be best represented by jitter values (see section 2.2) because of their susceptibility to changes within the IP network load utilisation.

For the accomplishment of the tests, an NGN test bed has been set up, in which concurrent VoIP (Voice over IP) data streams (each providing constant packet sending intervals) were established. Figure 2 shows the test network layout, emulating a simplified NGN scenario.

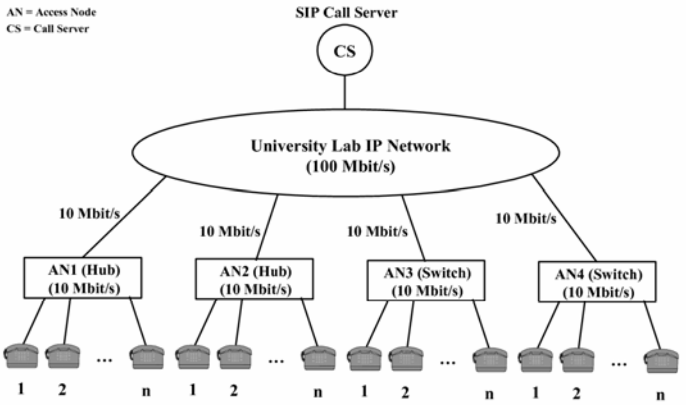
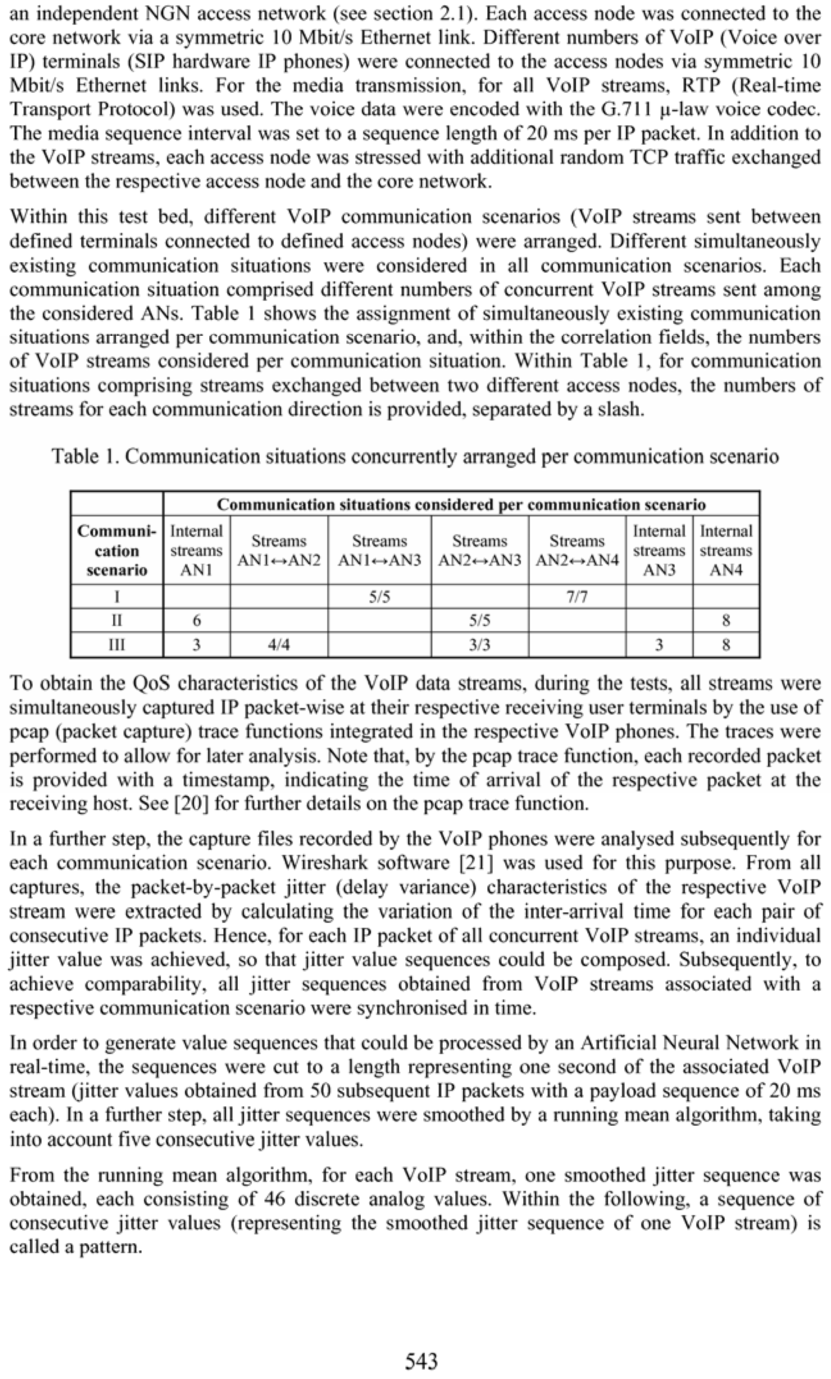


Figure 2. Test network layout

The NGN service stratum is represented by a SIP Proxy/Registrar Server. The transport stratum consists of an IP core network (University Lab IP Network, 100 Mbit/s) and four relocated access nodes (AN) (2 hubs (AN1 and AN2) and 2 switches (AN3 and AN4)), each representing



an independent NGN access network (see section 2.1). Each access node was connected to the core network via a symmetric 10 Mbit/s Ethernet link. Different numbers of VoIP (Voice over IP) terminals (SIP hardware IP phones) were connected to the access nodes via symmetric 10 Mbit/s Ethernet links. For the media transmission, for all VoIP streams, RTP (Real-time Transport Protocol) was used. The voice data were encoded with the G.711 µ-law voice codec. The media sequence interval was set to a sequence length of 20 ms per IP packet. In addition to the VoIP streams, each access node was stressed with additional random TCP traffic exchanged between the respective access node and the core network.

Within this test bed, different VoIP communication scenarios (VoIP streams sent between defined terminals connected to defined access nodes) were arranged. Different simultaneously existing communication situations were considered in all communication scenarios. Each communication situation comprised different numbers of concurrent VoIP streams sent among the considered ANs. Table 1 shows the assignment of simultaneously existing communication situations arranged per communication scenario, and, within the correlation fields, the numbers of VoIP streams considered per communication situation. Within Table 1, for communication situations comprising streams exchanged between two different access nodes, the numbers of streams for each communication direction is provided, separated by a slash.

Table 1. Communication situations concurrently arranged per communication scenario

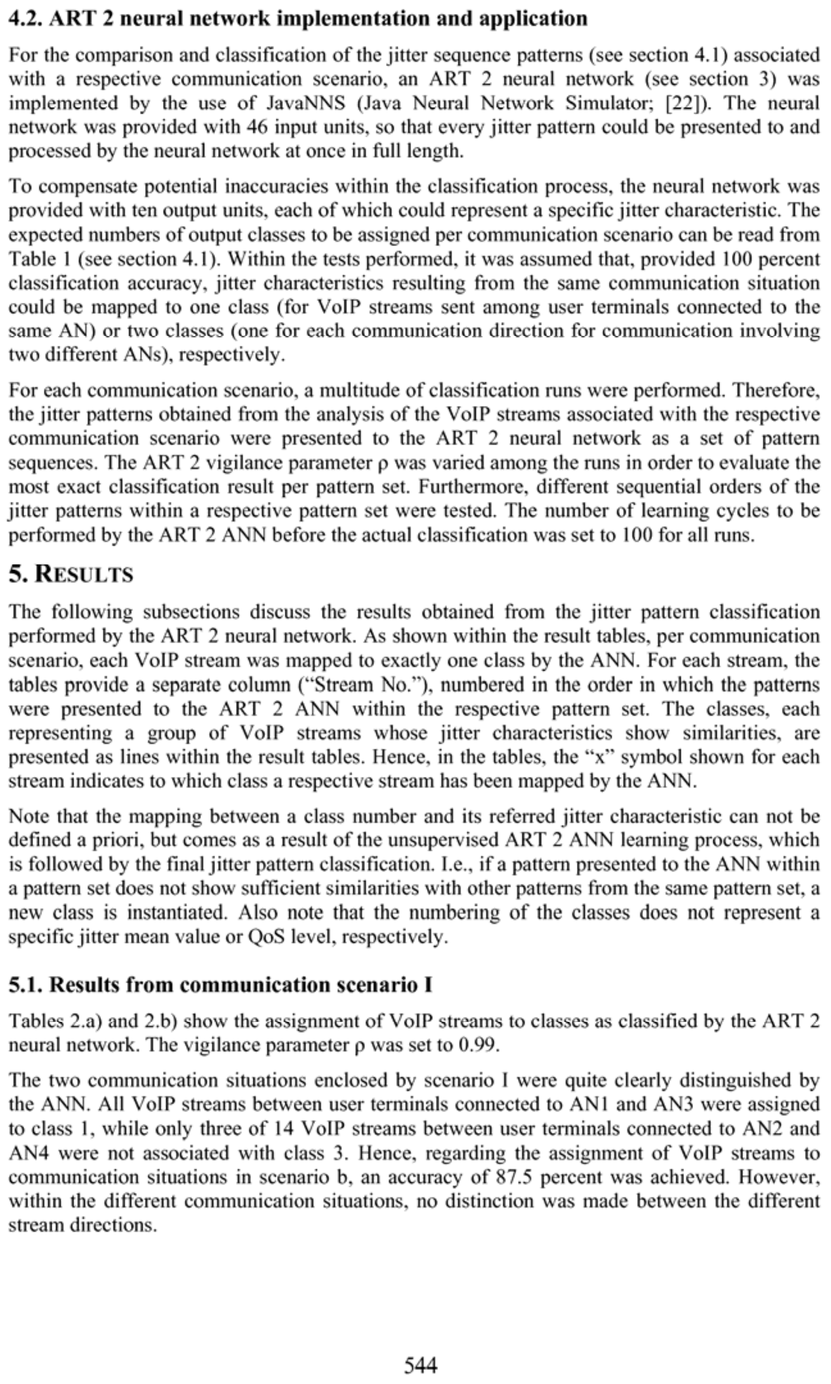
	Communication situations considered per communication scenario						
Communi-cation scenario	Internal streams AN1	Streams AN1↔AN2	Streams AN1↔AN3	Streams AN2↔AN3	Streams AN2↔AN4	Internal streams AN3	Internal streams AN4
I			5/5		7/7		
II	6			5/5			8
III	3	4/4		3/3		3	8

To obtain the QoS characteristics of the VoIP data streams, during the tests, all streams were simultaneously captured IP packet-wise at their respective receiving user terminals by the use of pcap (packet capture) trace functions integrated in the respective VoIP phones. The traces were performed to allow for later analysis. Note that, by the pcap trace function, each recorded packet is provided with a timestamp, indicating the time of arrival of the respective packet at the receiving host. See [20] for further details on the pcap trace function.

In a further step, the capture files recorded by the VoIP phones were analysed subsequently for each communication scenario. Wireshark software [21] was used for this purpose. From all captures, the packet-by-packet jitter (delay variance) characteristics of the respective VoIP stream were extracted by calculating the variation of the inter-arrival time for each pair of consecutive IP packets. Hence, for each IP packet of all concurrent VoIP streams, an individual jitter value was achieved, so that jitter value sequences could be composed. Subsequently, to achieve comparability, all jitter sequences obtained from VoIP streams associated with a respective communication scenario were synchronised in time.

In order to generate value sequences that could be processed by an Artificial Neural Network in real-time, the sequences were cut to a length representing one second of the associated VoIP stream (jitter values obtained from 50 subsequent IP packets with a payload sequence of 20 ms each). In a further step, all jitter sequences were smoothed by a running mean algorithm, taking into account five consecutive jitter values.

From the running mean algorithm, for each VoIP stream, one smoothed jitter sequence was obtained, each consisting of 46 discrete analog values. Within the following, a sequence of consecutive jitter values (representing the smoothed jitter sequence of one VoIP stream) is called a pattern.



4.2. ART 2 neural network implementation and application

For the comparison and classification of the jitter sequence patterns (see section 4.1) associated with a respective communication scenario, an ART 2 neural network (see section 3) was implemented by the use of JavaNNS (Java Neural Network Simulator; [22]). The neural network was provided with 46 input units, so that every jitter pattern could be presented to and processed by the neural network at once in full length.

To compensate potential inaccuracies within the classification process, the neural network was provided with ten output units, each of which could represent a specific jitter characteristic. The expected numbers of output classes to be assigned per communication scenario can be read from Table 1 (see section 4.1). Within the tests performed, it was assumed that, provided 100 percent classification accuracy, jitter characteristics resulting from the same communication situation could be mapped to one class (for VoIP streams sent among user terminals connected to the same AN) or two classes (one for each communication direction for communication involving two different ANs), respectively.

For each communication scenario, a multitude of classification runs were performed. Therefore, the jitter patterns obtained from the analysis of the VoIP streams associated with the respective communication scenario were presented to the ART 2 neural network as a set of pattern sequences. The ART 2 vigilance parameter ρ was varied among the runs in order to evaluate the most exact classification result per pattern set. Furthermore, different sequential orders of the jitter patterns within a respective pattern set were tested. The number of learning cycles to be performed by the ART 2 ANN before the actual classification was set to 100 for all runs.

5. Results

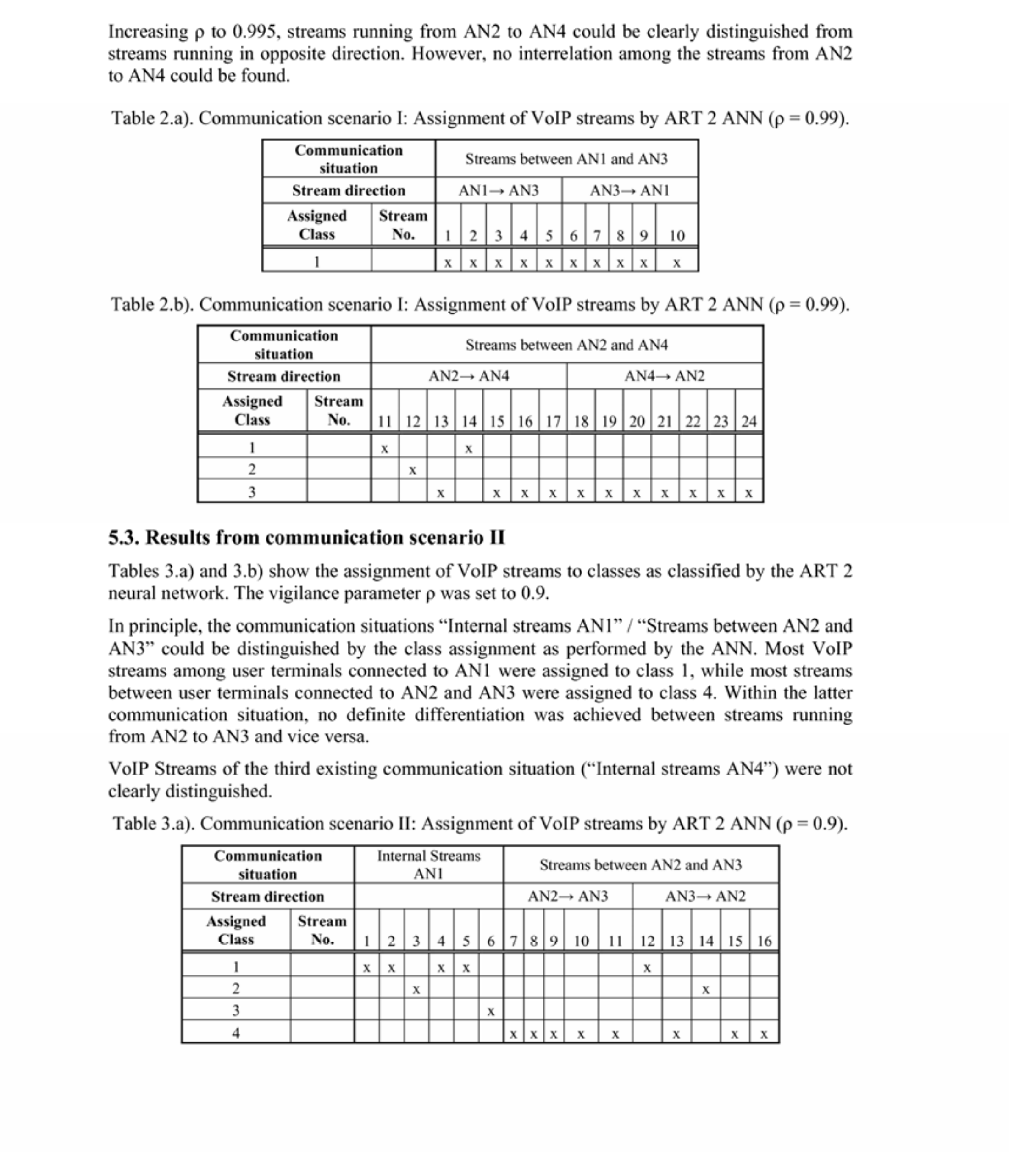
The following subsections discuss the results obtained from the jitter pattern classification performed by the ART 2 neural network. As shown within the result tables, per communication scenario, each VoIP stream was mapped to exactly one class by the ANN. For each stream, the tables provide a separate column (“Stream No.”), numbered in the order in which the patterns were presented to the ART 2 ANN within the respective pattern set. The classes, each representing a group of VoIP streams whose jitter characteristics show similarities, are presented as lines within the result tables. Hence, in the tables, the “x” symbol shown for each stream indicates to which class a respective stream has been mapped by the ANN.

Note that the mapping between a class number and its referred jitter characteristic can not be defined a priori, but comes as a result of the unsupervised ART 2 ANN learning process, which is followed by the final jitter pattern classification. I.e., if a pattern presented to the ANN within a pattern set does not show sufficient similarities with other patterns from the same pattern set, a new class is instantiated. Also note that the numbering of the classes does not represent a specific jitter mean value or QoS level, respectively.

5.1. Results from communication scenario I

Tables 2.a) and 2.b) show the assignment of VoIP streams to classes as classified by the ART 2 neural network. The vigilance parameter ρ was set to 0.99.

The two communication situations enclosed by scenario I were quite clearly distinguished by the ANN. All VoIP streams between user terminals connected to AN1 and AN3 were assigned to class 1, while only three of 14 VoIP streams between user terminals connected to AN2 and AN4 were not associated with class 3. Hence, regarding the assignment of VoIP streams to communication situations in scenario b, an accuracy of 87.5 percent was achieved. However, within the different communication situations, no distinction was made between the different stream directions.



Increasing ρ to 0.995, streams running from AN2 to AN4 could be clearly distinguished from streams running in opposite direction. However, no interrelation among the streams from AN2 to AN4 could be found.

Table 2.a). Communication scenario I: Assignment of VoIP streams by ART 2 ANN (ρ = 0.99).

Communication situation		Streams between AN1 and AN3									
Stream direction		AN1→ AN3					AN3→ AN1				
Assigned Class	Stream No.	1	2	3	4	5	6	7	8	9	10
1		x	x	x	x	x	x	x	x	x	x

Table 2.b). Communication scenario I: Assignment of VoIP streams by ART 2 ANN (ρ = 0.99).

Communication situation		Streams between AN2 and AN4													
Stream direction		AN2→ AN4							AN4→ AN2						
Assigned Class	Stream No.	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1		x			x										
2			x												
3				x		x	x	x	x	x	x	x	x	x	x

5.3. Results from communication scenario II

Tables 3.a) and 3.b) show the assignment of VoIP streams to classes as classified by the ART 2 neural network. The vigilance parameter ρ was set to 0.9.

In principle, the communication situations "Internal streams AN1" / "Streams between AN2 and AN3" could be distinguished by the class assignment as performed by the ANN. Most VoIP streams among user terminals connected to AN1 were assigned to class 1, while most streams between user terminals connected to AN2 and AN3 were assigned to class 4. Within the latter communication situation, no definite differentiation was achieved between streams running from AN2 to AN3 and vice versa.

VoIP Streams of the third existing communication situation ("Internal streams AN4") were not clearly distinguished.

Table 3.a). Communication scenario II: Assignment of VoIP streams by ART 2 ANN (ρ = 0.9).

Communication situation		Internal Streams AN1						Streams between AN2 and AN3									
Stream direction								AN2→ AN3					AN3→ AN2				
Assigned Class	Stream No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		x	x		x	x							x				
2				x											x		
3							x										
4								x	x	x	x	x		x		x	x

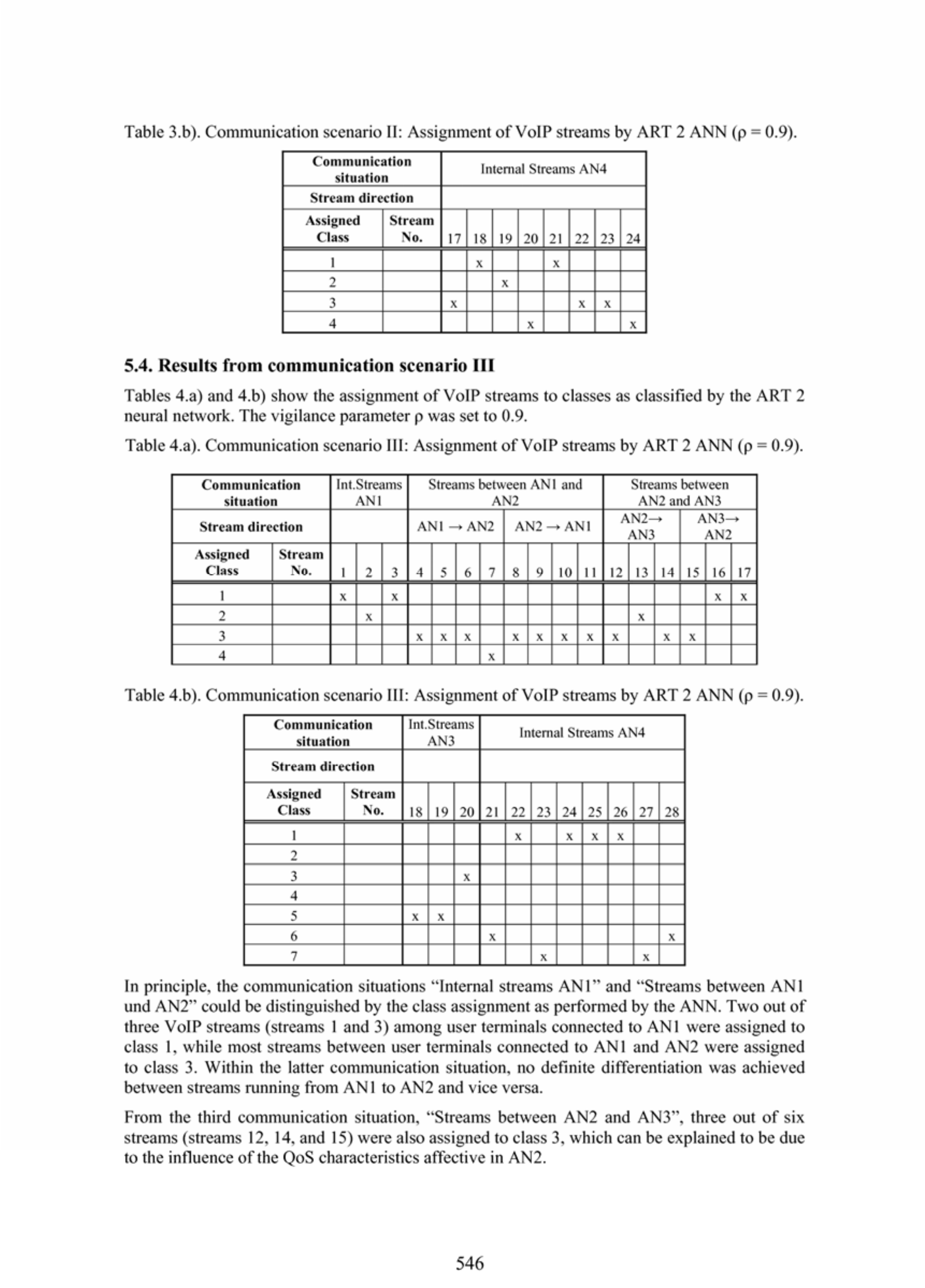


Table 3.b). Communication scenario II: Assignment of VoIP streams by ART 2 ANN (ρ = 0.9).

Communication situation		Internal Streams AN4							
Stream direction									
Assigned Class	Stream No.	17	18	19	20	21	22	23	24
1			x			x			
2				x					
3		x					x	x	
4					x				x

5.4. Results from communication scenario III

Tables 4.a) and 4.b) show the assignment of VoIP streams to classes as classified by the ART 2 neural network. The vigilance parameter ρ was set to 0.9.

Table 4.a). Communication scenario III: Assignment of VoIP streams by ART 2 ANN (ρ = 0.9).

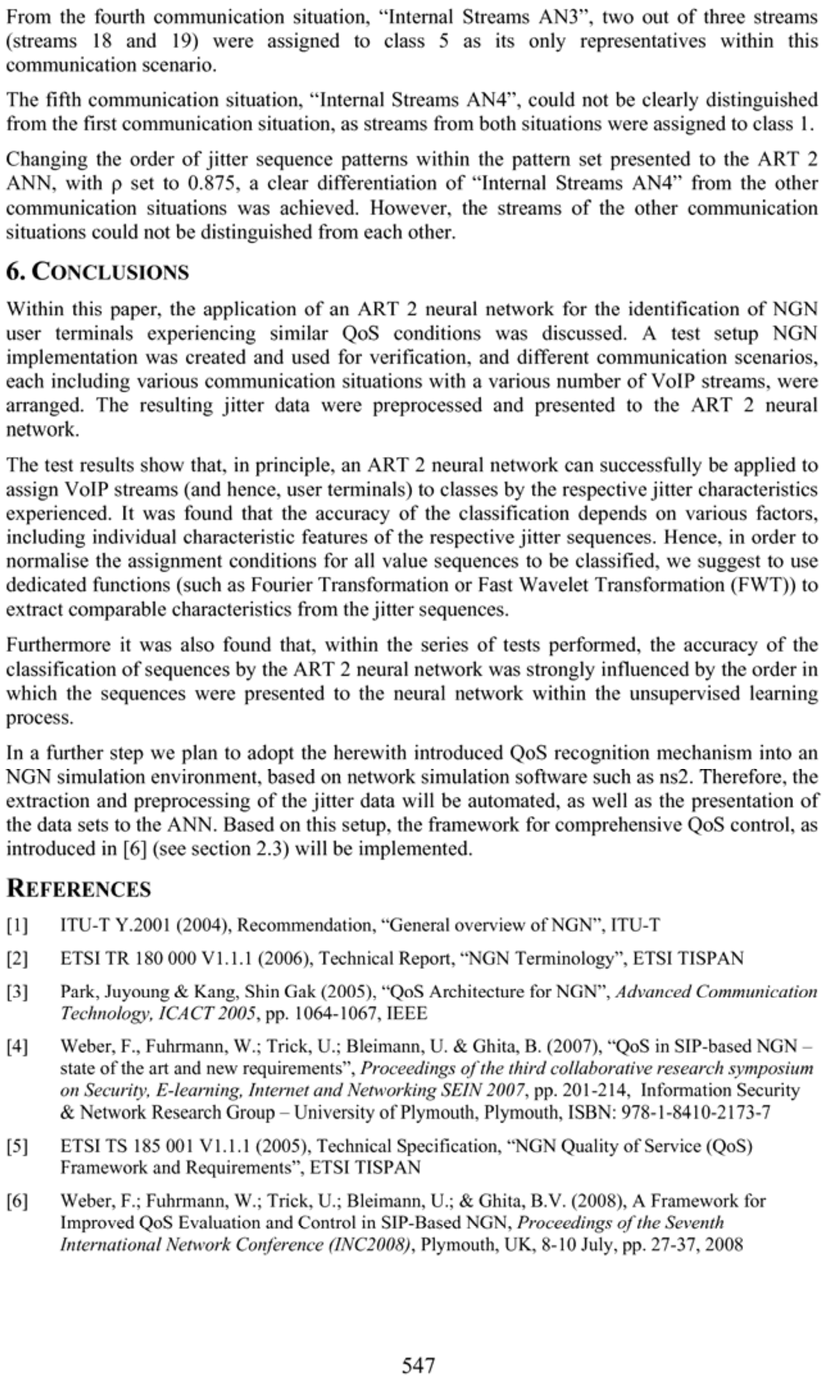
Communication situation		Int.Streams AN1			Streams between AN1 and AN2								Streams between AN2 and AN3					
Stream direction					AN1 → AN2				AN2 → AN1				AN2→ AN3			AN3→ AN2		
Assigned Class	Stream No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1		x		x													x	x
2			x											x				
3					x	x	x		x	x	x	x	x		x	x		
4								x										

Table 4.b). Communication scenario III: Assignment of VoIP streams by ART 2 ANN (ρ = 0.9).

Communication situation		Int.Streams AN3			Internal Streams AN4							
Stream direction												
Assigned Class	Stream No.	18	19	20	21	22	23	24	25	26	27	28
1						x		x	x	x		
2												
3				x								
4												
5		x	x									
6					x							x
7							x				x	

In principle, the communication situations "Internal streams AN1" and "Streams between AN1 und AN2" could be distinguished by the class assignment as performed by the ANN. Two out of three VoIP streams (streams 1 and 3) among user terminals connected to AN1 were assigned to class 1, while most streams between user terminals connected to AN1 and AN2 were assigned to class 3. Within the latter communication situation, no definite differentiation was achieved between streams running from AN1 to AN2 and vice versa.

From the third communication situation, "Streams between AN2 and AN3", three out of six streams (streams 12, 14, and 15) were also assigned to class 3, which can be explained to be due to the influence of the QoS characteristics affective in AN2.



From the fourth communication situation, "Internal Streams AN3", two out of three streams (streams 18 and 19) were assigned to class 5 as its only representatives within this communication scenario.

The fifth communication situation, "Internal Streams AN4", could not be clearly distinguished from the first communication situation, as streams from both situations were assigned to class 1.

Changing the order of jitter sequence patterns within the pattern set presented to the ART 2 ANN, with ρ set to 0.875, a clear differentiation of "Internal Streams AN4" from the other communication situations was achieved. However, the streams of the other communication situations could not be distinguished from each other.

6. Conclusions

Within this paper, the application of an ART 2 neural network for the identification of NGN user terminals experiencing similar QoS conditions was discussed. A test setup NGN implementation was created and used for verification, and different communication scenarios, each including various communication situations with a various number of VoIP streams, were arranged. The resulting jitter data were preprocessed and presented to the ART 2 neural network.

The test results show that, in principle, an ART 2 neural network can successfully be applied to assign VoIP streams (and hence, user terminals) to classes by the respective jitter characteristics experienced. It was found that the accuracy of the classification depends on various factors, including individual characteristic features of the respective jitter sequences. Hence, in order to normalise the assignment conditions for all value sequences to be classified, we suggest to use dedicated functions (such as Fourier Transformation or Fast Wavelet Transformation (FWT)) to extract comparable characteristics from the jitter sequences.

Furthermore it was also found that, within the series of tests performed, the accuracy of the classification of sequences by the ART 2 neural network was strongly influenced by the order in which the sequences were presented to the neural network within the unsupervised learning process.

In a further step we plan to adopt the herewith introduced QoS recognition mechanism into an NGN simulation environment, based on network simulation software such as ns2. Therefore, the extraction and preprocessing of the jitter data will be automated, as well as the presentation of the data sets to the ANN. Based on this setup, the framework for comprehensive QoS control, as introduced in [6] (see section 2.3) will be implemented.

References

[1] ITU-T Y.2001 (2004), Recommendation, "General overview of NGN", ITU-T

[2] ETSI TR 180 000 V1.1.1 (2006), Technical Report, "NGN Terminology", ETSI TISPAN

[3] Park, Juyoung & Kang, Shin Gak (2005), "QoS Architecture for NGN", *Advanced Communication Technology, ICACT 2005*, pp. 1064-1067, IEEE

[4] Weber, F., Fuhrmann, W.; Trick, U.; Bleimann, U. & Ghita, B. (2007), "QoS in SIP-based NGN – state of the art and new requirements", *Proceedings of the third collaborative research symposium on Security, E-learning, Internet and Networking SEIN 2007*, pp. 201-214, Information Security & Network Research Group – University of Plymouth, Plymouth, ISBN: 978-1-8410-2173-7

[5] ETSI TS 185 001 V1.1.1 (2005), Technical Specification, "NGN Quality of Service (QoS) Framework and Requirements", ETSI TISPAN

[6] Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U.; & Ghita, B.V. (2008), A Framework for Improved QoS Evaluation and Control in SIP-Based NGN, *Proceedings of the Seventh International Network Conference (INC2008)*, Plymouth, UK, 8-10 July, pp. 27-37, 2008

[7] Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U.; & Ghita, B.V. (2008), Selection of QoS Monitoring Points in a New QoS Control Framework for SIP-Based NGN, *Proceedings of the Fourth Collaborative Research Symposium on Security, E-learning, Internet and Networking (SEIN 2008)*, Wrexham, UK, ISBN: 978-1-84102-196-6, pp. 176-185, 2008

[8] Carpenter, G.A. & Grossberg, S. (1987). ART 2: "Self-organization of Stable Category Recognition Codes for Analog Input Patterns", *Applied Optics*, 26 , pp. 4919-4930.

[9] Trick, Ulrich & Weber, Frank (2007), *SIP, TCP/IP und Telekommunikationsnetze (3rd edition)*, Oldenbourg, Munich, Germany, ISBN: 978-3-486-58228-4

[10] ITU-T Y.2011 (2004), Recommendation, "Next Generation Networks – Frameworks and functional architecture models", ITU-T

[11] ETSI ES 282 001 V2.0.0 (2008), ETSI Standard, "NGN Functional Architecture", ETSI TISPAN

[12] ITU-T E.800 (1994), Recommendation, "Terms and definitions related to Quality of Service and Network Performance including dependability", ITU-T

[13] ITU-T Y.1291 (2004), Recommendation, "An architectural framework for support of Quality of Service in packet networks", ITU-T

[14] Gozdecki, Janusz; Jajszczyk, Andrzej & Stankiewicz, Rafal (2003), "Quality of service terminology in IP networks", *Communications Magazine*, Volume 41, Issue 3, pp. 153-159, IEEE

[15] Zheng, Li; Zhang, Liren & Xu, Dong (2001), "Characteristics of network delay and delay jitter and its effect on voice over IP (VoIP)", *IEEE International Conference on Communications, 2001, ICC 2001*, Volume 1, pp. 122-126, IEEE

[16] Kos, Anton; Klepec, Borut & Tomazic, Saso (2002), "Techniques for performance improvement of VoIP applications", *11th Mediterranean Electrotechnical Conference, 2002. MELECON 2002*, pp. 250-254, IEEE

[17] Borella, Michael S.; Swider, Debbie; Uludag, Suleyman; Brewster, & Gregory B. (1998), "Internet packet loss: measurement and implications for end-to-end QoS", *Proceedings of the 1998 ICPP Workshops on Architectural and OS Support for Multimedia Applications/Flexible Communication Systems/Wireless Networks and Mobile Computing*, 14 Aug 1998, pp. 3-12, IEEE

[18] White, Christopher M.; Raymond, J. & Teague, K.A. (2004), "A real-time network simulation application for multimedia over IP", *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, 2004, Volume 2, pp. 2245-2249, IEEE

[19] Jain, A.K.; Jianchang Mao & Mohiuddin, K.M., (1996), "Artificial neural networks: a tutorial", *Computer*, vol.29, no.3, pp. 31-44, Mar 1996

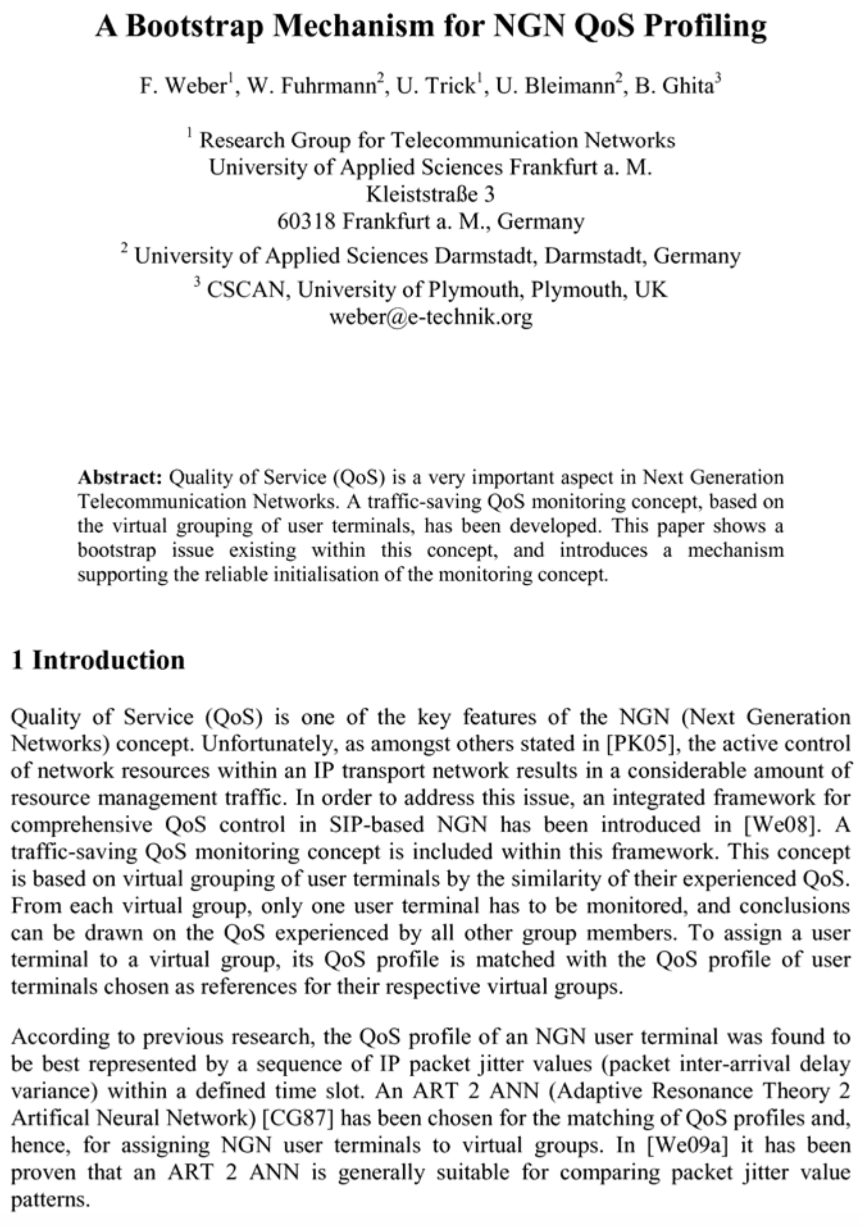
[20] http://www.tcpdump.org/ (as visited on 27 February 2009)

[21] http://www.wireshark.org/ (as visited on 04 February 2009)

[22] http://www.ra.cs.uni-tuebingen.de/software/JavaNNS/welcome_e.html (as visited on 04 February 2009)

548

Published in *Proceedings of the eighth International Network Conference (INC 2010)*, pp. 61-70, University of Plymouth, Plymouth, UK, ISBN: 978-1-84102-259-8



A Bootstrap Mechanism for NGN QoS Profiling

F. Weber[1], W. Fuhrmann[2], U. Trick[1], U. Bleimann[2], B. Ghita[3]

[1] Research Group for Telecommunication Networks
University of Applied Sciences Frankfurt a. M.
Kleiststraße 3
60318 Frankfurt a. M., Germany

[2] University of Applied Sciences Darmstadt, Darmstadt, Germany

[3] CSCAN, University of Plymouth, Plymouth, UK
weber@e-technik.org

Abstract: Quality of Service (QoS) is a very important aspect in Next Generation Telecommunication Networks. A traffic-saving QoS monitoring concept, based on the virtual grouping of user terminals, has been developed. This paper shows a bootstrap issue existing within this concept, and introduces a mechanism supporting the reliable initialisation of the monitoring concept.

1 Introduction

Quality of Service (QoS) is one of the key features of the NGN (Next Generation Networks) concept. Unfortunately, as amongst others stated in [PK05], the active control of network resources within an IP transport network results in a considerable amount of resource management traffic. In order to address this issue, an integrated framework for comprehensive QoS control in SIP-based NGN has been introduced in [We08]. A traffic-saving QoS monitoring concept is included within this framework. This concept is based on virtual grouping of user terminals by the similarity of their experienced QoS. From each virtual group, only one user terminal has to be monitored, and conclusions can be drawn on the QoS experienced by all other group members. To assign a user terminal to a virtual group, its QoS profile is matched with the QoS profile of user terminals chosen as references for their respective virtual groups.

According to previous research, the QoS profile of an NGN user terminal was found to be best represented by a sequence of IP packet jitter values (packet inter-arrival delay variance) within a defined time slot. An ART 2 ANN (Adaptive Resonance Theory 2 Artifical Neural Network) [CG87] has been chosen for the matching of QoS profiles and, hence, for assigning NGN user terminals to virtual groups. In [We09a] it has been proven that an ART 2 ANN is generally suitable for comparing packet jitter value patterns.

An ART 2 ANN provides a number of setup parameters, of which the vigilance parameter ρ is the most effective. It strongly influences the degree of similarity that patterns have to exhibit in order to be assigned to the same group. The most suitable value for ρ directly depends on the characteristics of the patterns to be matched. In [We09b] a mechanism has been introduced to automatically determine the most suitable ρ value for arbitrary numbers of jitter patterns. Note that this mechanism is based on the distinguishability of at least two reference jitter patterns which are known to represent different virtual groups. This mechanism is periodically applied within the continuing grouping process of the QoS control framework introduced in [We08].

Regarding the initial start-up of the grouping process, a general bootstrapping issue occurs. The most suitable ρ value for a set of jitter patterns can only be reliably identified if the set consists of at least two reference patterns, each representing a different virtual group. This is also true for the classification of the initial patterns available after the framework start-up. However, in order to reliably identify the group affiliations of the first jitter patterns available, a most suitable ρ value is required. Hence, without being aware of the group affiliations of at least the first two jitter patterns available, any further classification will be potentially inaccurate. This paper proposes a mechanism to solve this bootstrapping issue. Initial tests have been accomplished and are introduced within this paper.

2 Next Generation Networks (NGN) and Quality of Service (QoS)

In 2004, the ITU-T (International Telecommunication Union – Telecommunication Standardization Sector) released its definition of NGN in [IT04a]. According to [IT04a], [ET06], and [TW09] the term NGN stands for a telecommunication network concept that can be characterised by a number of key features including, amongst others, "Packet-based data transport" and "Quality of Service support". Although the term "Packet-based data transport" does not refer to any particular technology or protocol, IP (Internet Protocol) is the most likely network protocol choice for an NGN environment according to [TW09]. The use of SIP (Session Initiation Protocol) for NGN service provisioning and signalling is widely accepted, and also suggested in [ET08].

2.1 QoS for real-time telecommunication services

For services provided within telecommunication networks, the term QoS has been defined as the "collective effect of service performance which determines the degree of satisfaction of a user of the service" [IT94]. According to [IT04b], for packet-based media data transport (which is given in NGN), the quality of a real-time based telecommunication service as experienced by a service user directly depends on the network performance of the respective transport network. In [Go03] the network performance of an IP transport network is characterised by the packet loss ratio, the transfer delay, and the transfer delay variation (jitter). These network performance parameters substantially influence the QoS of a real-time communication service as experienced by its users.

2.2 Integrated framework for comprehensive QoS control in SIP-based NGN

The authors previously proposed a framework for QoS control, aiming to address the scalability issues related to QoS provision in SIP-based NGN, described in [PK05] and [We08]. This framework (see Figure 1) is provided with an integrated mechanism for the collection of information on the QoS affecting any ongoing and future media session.

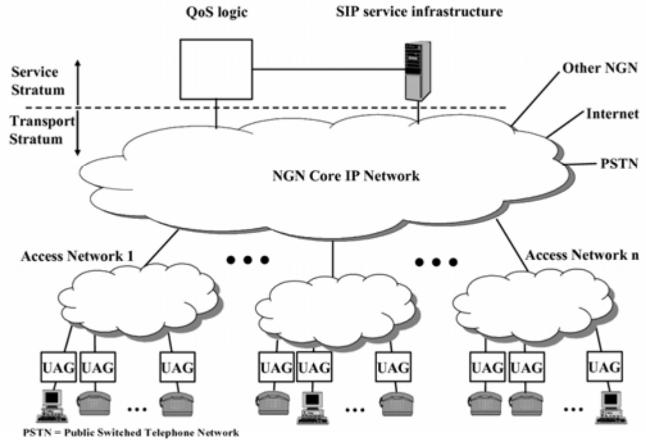


Figure 1: Integrated framework for comprehensive QoS control in SIP-based NGN

In terms of this framework, QoS information consists of delay, jitter, and packet loss values affecting the packets of a respective media data stream. The information is collected in so called user access gates (UAGs) close to the user terminal receiving the respective data stream. In order to minimise the resulting QoS monitoring traffic, only UAGs of selected user terminals (representing a QoS reference point) are queried for information on the QoS conditions experienced by ongoing media sessions. Note that the querying is initiated by the provider's SIP service infrastructure, and triggered by the provider's QoS logic. This requires the identification of virtual groups of user terminals whose members experience similar QoS conditions, and hence, can be represented by a specific reference point. The assignment of user terminals to their respective virtual group is proposed to be performed by the QoS logic entity within the provider infrastructure. The grouping results from the comparison of QoS conditions experienced, and is performed by the support of an Artificial Neural Network (ANN) to allow for an improved real-time processing behaviour. The principle of assigning media streams to virtual groups (or classes, respectively) by the use of an ANN is introduced in section 4 of this paper. A detailed description of the overall framework functionality of QoS information collection is provided in [We08].

3 ART 2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are used in numerous technical applications in order to perform complex tasks such as pattern recognition (or pattern classification), function approximation, prediction/forecasting, optimisation, content-addressable memorising, cybernetics, as well as clustering/categorisation. The latter application is denoted as unsupervised pattern classification in [Ja96].

ART 2 (Adaptive Resonance Theory 2) neural networks can be described as unsupervised-learning neural networks with the ability to compare analogue continuous value sequences with the objective to classify the sequences by their similarities according to [CG87]. This is performed by self-organisation of stable recognition codes generated from the input value sequences. An input sequence, also referred to as a pattern, is interpreted as an n-dimensional vector by an ART network, where n is the number of values comprised by the respective input pattern.

An ART 2 ANN provides n input units and m output units, the latter of which represent m individual output classes. If an arbitrary number of n-dimensional patterns is presented to an ART 2 network, after a predefined number of learning cycles, the network tries to map each pattern to one of m output classes by accomplishing a multi-step comparison process for each pattern. Patterns showing typical similarities are assigned to the same output class. For the comparison and classification of the patterns, the ART 2 ANN interprets every pattern as an n-dimensional vector, where n refers to the number of input values per pattern.

A number of setup parameters are provided by ART 2 ANN, of which the vigilance parameter ρ is the most effective. ρ represents a selectable threshold for the deviation $||\mathbf{r}||$ of two n-dimensional vectors u and cp (see equation (1)), where u represents the candidate pattern to be currently classified. Vector cp represents the ART2-internal pattern image of a certain class. With e=0, it is obvious from equation (2) that a reset event is triggered when $||\mathbf{r}|| <$ ρ. The reset event causes the ART2-internal resumption of the classification process of the respective pattern represented by u_i, excluding the respective class represented by cp_i.

$$r_i = \frac{u_i + cp_i}{e + ||\mathbf{u}|| + ||c\mathbf{p}||} \qquad (1) \qquad\qquad \text{Reset} \Leftrightarrow \frac{\rho}{e + ||\mathbf{r}||} > 1 \qquad (2)$$

Further details on the theory of ART 2 neural networks can be found in [CG87].

4 NGN QoS profiling

The term 'QoS profiling' refers to the virtual grouping of NGN user terminals by QoS conditions encountered. The reason for applying QoS profiling is the reduction of network traffic resulting from comprehensive QoS monitoring.

The integrated NGN QoS control framework briefly introduced in section 2.2 provides a centralised unit for the rating of QoS conditions. It is assumed that all NGN user terminals associated with a virtual group encounter similar QoS conditions. Hence, it is sufficient to choose one group member as the group's reference point and, subsequently, gather and rate QoS condition information from this reference point only. This information is assumed to represent the QoS experienced by any member of the respective virtual group.

Note that at least one reference point has to be chosen per existing virtual group by an arbitrary selection process. Each group's selected reference point is queried to continuously provide QoS information to the framework's centralised QoS rating unit.

4.1 Virtual grouping of NGN user terminals

In order to assign new NGN user terminals to existing virtual groups, their QoS characteristics have to be matched with the QoS characteristics gathered from the reference points. According to our experiments introduced in [We09a], the QoS characteristics of packet streams were found to be best represented by jitter values (see section 2.1) because of their susceptibility to changes within the IP network load utilisation.

In [We09b] a general mechanism was introduced to utilise an ART 2 ANN (see section 3) for the virtual grouping of NGN user terminals by their jitter characteristics. For this purpose, sets of jitter patterns are sequentially presented to the ANN. Figure 2 shows the principles of this mechanism.

After the mechanism start-up, a pattern set is arranged. It comprises all reference patterns (each representing an autonomous virtual group) plus one Pattern Under Test (PUT). Note that all patterns must be synchronised in time. An initial value for the ART 2 vigilance parameter ρ is set and the pattern set is presented to the ART 2 ANN and several classification runs are performed sequentially. After each run, ρ is adapted for the next run, until a ρ value is found so that all reference patterns are identified and assigned to different classes by the ANN (the finally determined ρ value is considered the most suitable ρ value for this specific pattern set). If the PUT matches one of the reference patterns, the ANN assigns the PUT automatically to the class represented by the reference. If the PUT could not be assigned to any class, a new virtual group is established and the PUT is the first member of the new group (and, hence, automatically becomes the reference of this group). This mechanism is periodically applied within the continuing grouping process of the QoS control framework introduced in [We08].

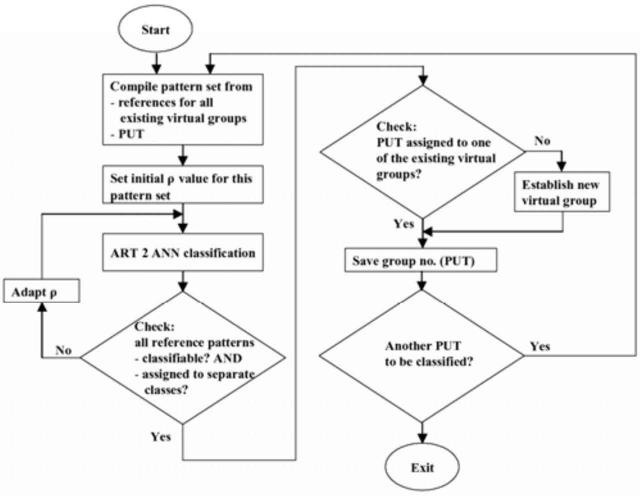


Figure 2: Mechanism for the virtual grouping of NGN user terminals

4.2 Bootstrap issue

As stated in section 4.1, the mechanism for the virtual grouping of NGN user terminals is based on the distinguishability of reference value patterns, each representing an autonomous virtual group. This distinguishability of references is considered mandatory in order to determine the most suitable ρ value for a specific pattern set. In turn, the most suitable ρ value is required to reliably identify the group affiliation of the respective PUT.

Reference patterns originate from ordinary NGN user terminals. These user terminals might be selected as representatives of their virtual groups by any arbitrary algorithm. Note that the assignment of a user terminal to its virtual group is performed before this user terminal might be selected as a group representative.

Also note that, like any other user terminal, a user terminal nominated as a group representative originally was assigned to its group, which typically requires the application of the grouping procedure described in section 4.1. However, the application of this procedure requires the reliable knowledge of the group affiliations of the reference patterns. Hence, the introduced QoS profiling approach lacks of a start-up mechanism providing the required information.

5 A bootstrap mechanism for NGN QoS profiling

As previously mentioned, the NGN QoS profiling approach introduced in [We09b] shows a bootstrap issue. This issue results from the unawareness of the group memberships of the first NGN user terminals who initially exchange media streams after the start-up of the QoS control framework briefly described in section 2.2.

Figure 3 shows a mechanism that solves this bootstrap issue. Note that at least three jitter patterns (monitored synchronously at different NGN user terminals) must be available in order to apply this mechanism. These jitter patterns might be derived from the first three media data streams exchanged after the framework start-up. The idea of this bootstrap mechanism is to detect mutual similarity and discrimination features among those three patterns and, hence, identify their group affiliations. No previous knowledge is required regarding any relationship of the pattern sources.

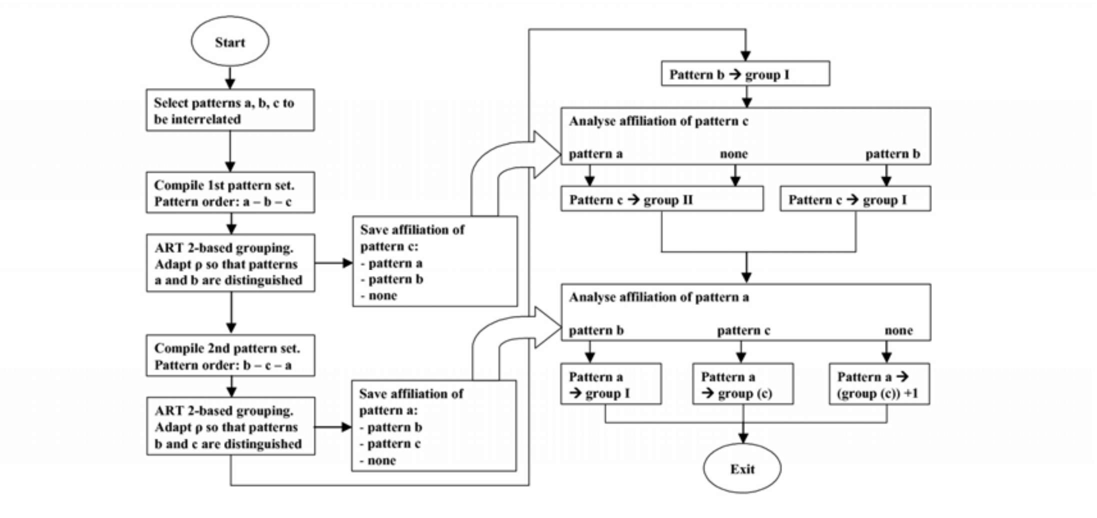
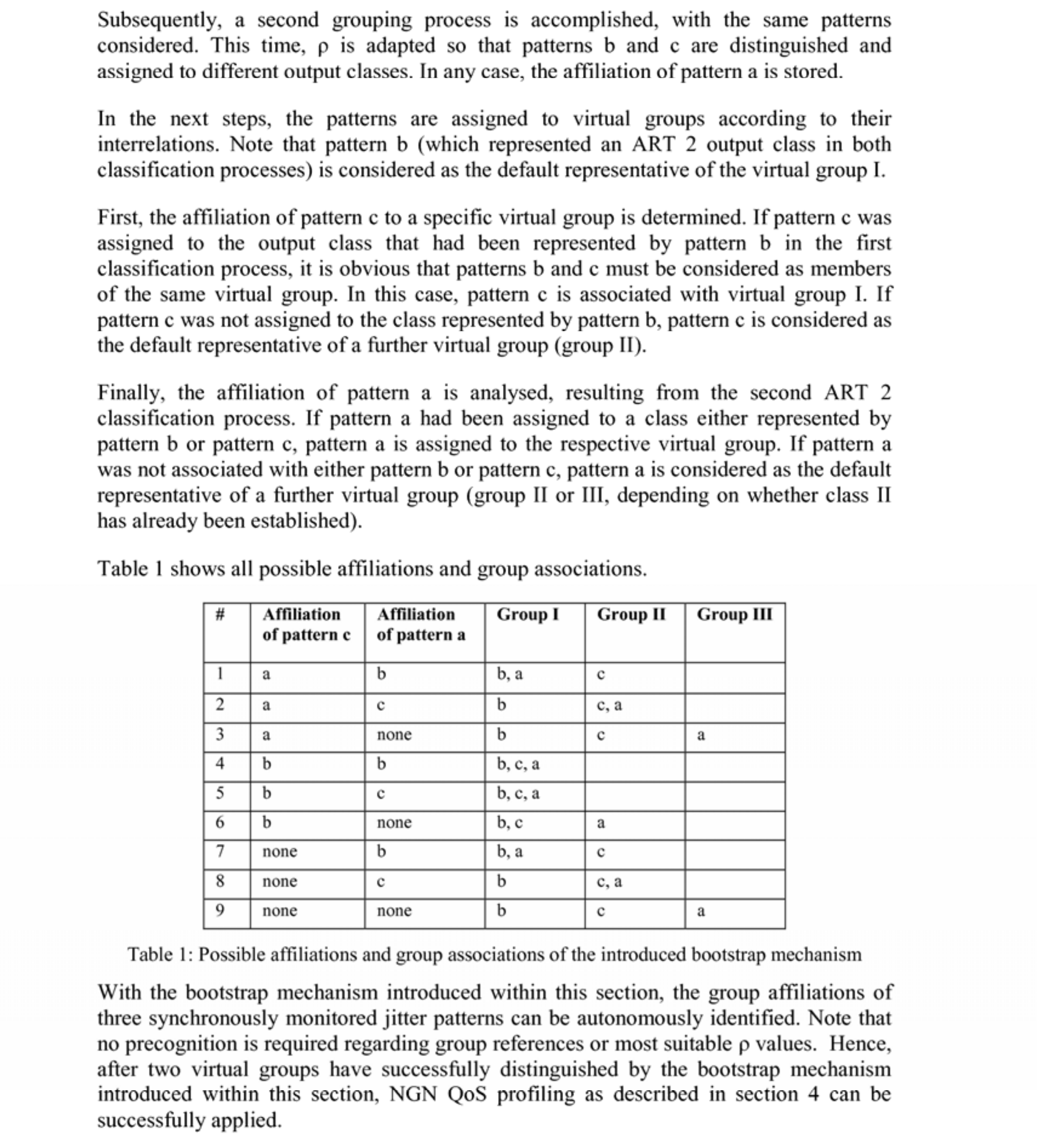

Start
Select patterns a, b, c to be interrelated
Compile 1st pattern set. Pattern order: a – b – c
ART 2-based grouping. Adapt ρ so that patterns a and b are distinguished
Save affiliation of pattern c:
- pattern a
- pattern b
- none
Compile 2nd pattern set. Pattern order: b – c – a
ART 2-based grouping. Adapt ρ so that patterns b and c are distinguished
Save affiliation of pattern a:
- pattern b
- pattern c
- none
Pattern b → group I
Analyse affiliation of pattern c
pattern a
none
pattern b
Pattern c → group II
Pattern c → group I
Analyse affiliation of pattern a
pattern b
pattern c
none
Pattern a → group I
Pattern a → group (c)
Pattern a → (group (c)) +1
Exit

Figure 3: Bootstrap mechanism for NGN QoS profiling

Three jitter value patterns (named a, b, and c) are arranged as a pattern set. According to the mechanism described in section 4.1, a first ART 2-based grouping process is performed. During this grouping process, the pattern set is presented to an ART 2 ANN equipped with two output units (hence, two different classes can be distinguished). The ART 2 vigilance parameter ρ is adapted so that patterns a and b are distinguished, and each of them is assigned to a different output class. Pattern c, as a member of the same pattern set, is also considered within the grouping procedure. Depending on similarity characteristics, it might be assigned to the same output class as pattern a or to the same output class as pattern b. However, if no sufficient similarity is given, pattern c will not be assigned to any existing output class. In any case, the grouping result for pattern c is stored for further processing.



Subsequently, a second grouping process is accomplished, with the same patterns considered. This time, ρ is adapted so that patterns b and c are distinguished and assigned to different output classes. In any case, the affiliation of pattern a is stored.

In the next steps, the patterns are assigned to virtual groups according to their interrelations. Note that pattern b (which represented an ART 2 output class in both classification processes) is considered as the default representative of the virtual group I.

First, the affiliation of pattern c to a specific virtual group is determined. If pattern c was assigned to the output class that had been represented by pattern b in the first classification process, it is obvious that patterns b and c must be considered as members of the same virtual group. In this case, pattern c is associated with virtual group I. If pattern c was not assigned to the class represented by pattern b, pattern c is considered as the default representative of a further virtual group (group II).

Finally, the affiliation of pattern a is analysed, resulting from the second ART 2 classification process. If pattern a had been assigned to a class either represented by pattern b or pattern c, pattern a is assigned to the respective virtual group. If pattern a was not associated with either pattern b or pattern c, pattern a is considered as the default representative of a further virtual group (group II or III, depending on whether class II has already been established).

Table 1 shows all possible affiliations and group associations.

#	Affiliation of pattern c	Affiliation of pattern a	Group I	Group II	Group III
1	a	b	b, a	c	
2	a	c	b	c, a	
3	a	none	b	c	a
4	b	b	b, c, a		
5	b	c	b, c, a		
6	b	none	b, c	a	
7	none	b	b, a	c	
8	none	c	b	c, a	
9	none	none	b	c	a

Table 1: Possible affiliations and group associations of the introduced bootstrap mechanism

With the bootstrap mechanism introduced within this section, the group affiliations of three synchronously monitored jitter patterns can be autonomously identified. Note that no precognition is required regarding group references or most suitable ρ values. Hence, after two virtual groups have successfully distinguished by the bootstrap mechanism introduced within this section, NGN QoS profiling as described in section 4 can be successfully applied.

6 Test and conclusion

The bootstrap mechanism for NGN QoS profiling introduced in section 5 has been exemplarily tested in a proof-of-concept manner. Initial result trends are introduced within this section, and a conclusion is given.

6.1 Test and result trends

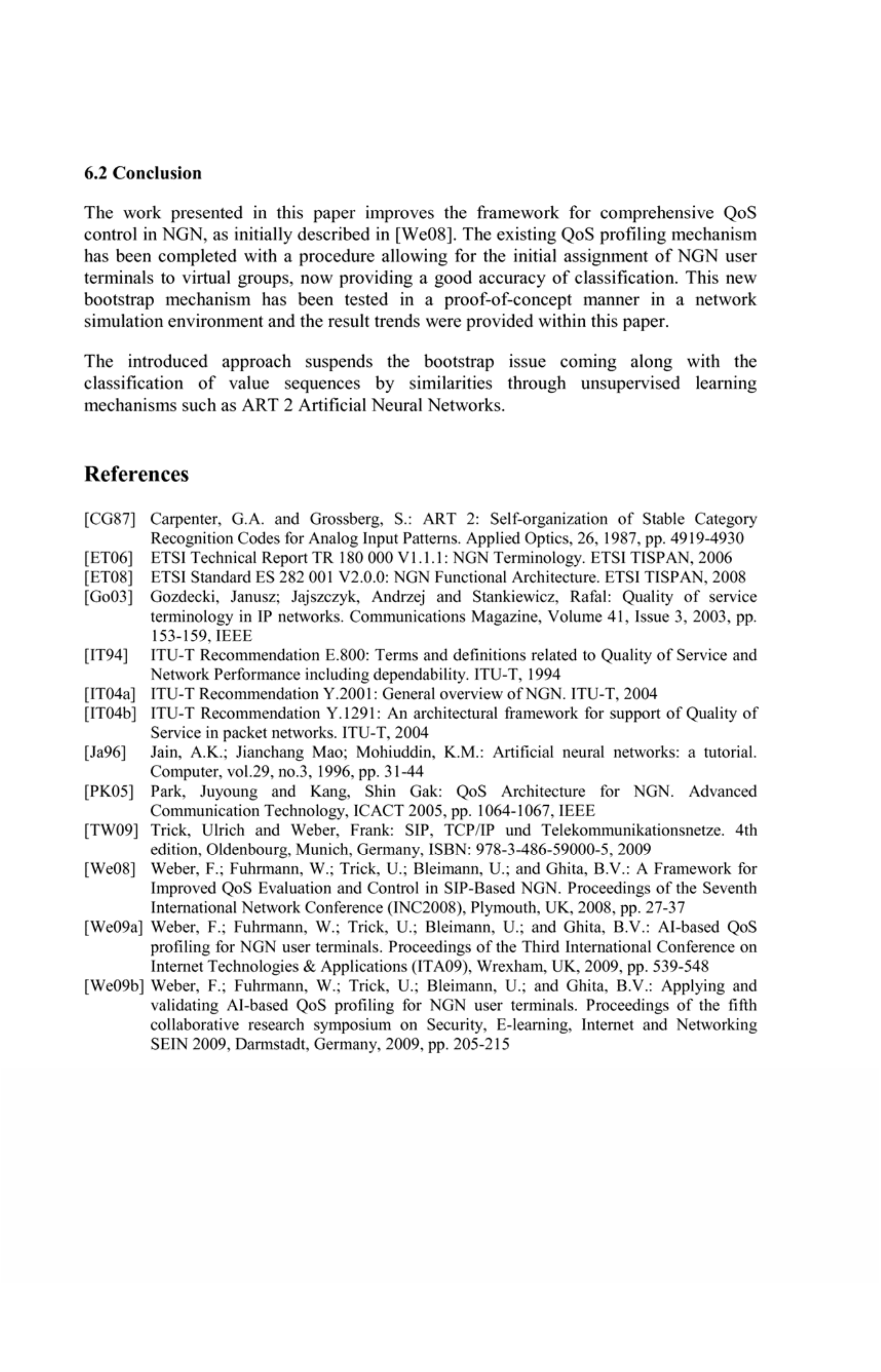
Based on the ns-2 network simulator, a SIP-based NGN architecture has been set up, allowing for the simulation of different communication scenarios. Upon session initiation, media flow packets (simulating VoIP calls with G.711 codec) were bidirectional exchanged in a peer-to-peer manner between the user terminals. By the use of the ns-2 trace function, all media packets were recorded and time-stamped at their respective receiving user terminals. The collected data were synchronised and post-processed to extract the per-packet inter-arrival jitter of each media flow. Several pattern sets were arranged, each comprising three patterns to emulate a bootstrap scenario.

Table 2 shows several scenarios considered. The scenarios differ in the number of comprised virtual groups and in the order of the patterns representing the groups. The accuracy stated in Table 2 is related to the correct assignment of patterns to virtual groups by the bootstrap mechanism introduced in section 5 of this paper, with ten different pattern sets (each comprising 3 patterns) tested per scenario.

Scenario No.	**No. of virtual groups included**	**Order of patterns within sets (by group numbers)**	**Accuracy of group assignment achieved**
1	3	I - II - III	100%
2	1	I - I - I	67%
3	2	I - II - II	100%
4	2	I - II - I	100%
5	2	I - I - II	100%

Table 2: Test scenarios for QoS profiling bootstrap mechanism

Table 2 shows that the bootstrap mechanism introduced within this paper provides an excellent assignment accuracy for those scenarios in which the considered jitter patterns comprise more than one virtual group. However, in scenario 2, in which all three patterns included within a pattern set belong to the same virtual group, a limited accuracy is experienced. This is due to the fact that the effectiveness of the bootstrap mechanism introduced within this paper depends on the distinguishability of patterns. However, the distinguishability of patterns belonging to the same virtual group is naturally limited. Those patterns must provide significant similarities in order to be defined as members of the same virtual groups.



6.2 Conclusion

The work presented in this paper improves the framework for comprehensive QoS control in NGN, as initially described in [We08]. The existing QoS profiling mechanism has been completed with a procedure allowing for the initial assignment of NGN user terminals to virtual groups, now providing a good accuracy of classification. This new bootstrap mechanism has been tested in a proof-of-concept manner in a network simulation environment and the result trends were provided within this paper.

The introduced approach suspends the bootstrap issue coming along with the classification of value sequences by similarities through unsupervised learning mechanisms such as ART 2 Artificial Neural Networks.

References

[CG87] Carpenter, G.A. and Grossberg, S.: ART 2: Self-organization of Stable Category Recognition Codes for Analog Input Patterns. Applied Optics, 26, 1987, pp. 4919-4930

[ET06] ETSI Technical Report TR 180 000 V1.1.1: NGN Terminology. ETSI TISPAN, 2006

[ET08] ETSI Standard ES 282 001 V2.0.0: NGN Functional Architecture. ETSI TISPAN, 2008

[Go03] Gozdecki, Janusz; Jajszczyk, Andrzej and Stankiewicz, Rafal: Quality of service terminology in IP networks. Communications Magazine, Volume 41, Issue 3, 2003, pp. 153-159, IEEE

[IT94] ITU-T Recommendation E.800: Terms and definitions related to Quality of Service and Network Performance including dependability. ITU-T, 1994

[IT04a] ITU-T Recommendation Y.2001: General overview of NGN. ITU-T, 2004

[IT04b] ITU-T Recommendation Y.1291: An architectural framework for support of Quality of Service in packet networks. ITU-T, 2004

[Ja96] Jain, A.K.; Jianchang Mao; Mohiuddin, K.M.: Artificial neural networks: a tutorial. Computer, vol.29, no.3, 1996, pp. 31-44

[PK05] Park, Juyoung and Kang, Shin Gak: QoS Architecture for NGN. Advanced Communication Technology, ICACT 2005, pp. 1064-1067, IEEE

[TW09] Trick, Ulrich and Weber, Frank: SIP, TCP/IP und Telekommunikationsnetze. 4th edition, Oldenbourg, Munich, Germany, ISBN: 978-3-486-59000-5, 2009

[We08] Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U.; and Ghita, B.V.: A Framework for Improved QoS Evaluation and Control in SIP-Based NGN. Proceedings of the Seventh International Network Conference (INC2008), Plymouth, UK, 2008, pp. 27-37

[We09a] Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U.; and Ghita, B.V.: AI-based QoS profiling for NGN user terminals. Proceedings of the Third International Conference on Internet Technologies & Applications (ITA09), Wrexham, UK, 2009, pp. 539-548

[We09b] Weber, F.; Fuhrmann, W.; Trick, U.; Bleimann, U.; and Ghita, B.V.: Applying and validating AI-based QoS profiling for NGN user terminals. Proceedings of the fifth collaborative research symposium on Security, E-learning, Internet and Networking SEIN 2009, Darmstadt, Germany, 2009, pp. 205-215