An IP QoS architecture for 4G networks

Janusz Gozdecki[1], Piotr Pacyna[1], Victor Marques[2], Rui L. Aguiar[3], Carlos Garcia[4],
Jose Ignacio Moreno[4], Christophe Beaujean[5], Eric Melin[5], Marco Liebsch[6]

[1] AGH University of Technology, Kraków, Poland ({pacyna, gozdecki}@kt.agh.edu.pl)
[2] Portugal Telecom Inovação, 3810-106 Aveiro Portugal (victor-m-marques@ptinovacao.pt)
[3] Instituto de Telecomunicações/Universidade de Aveiro, 3810 Aveiro, Portugal
(ruilaa@det.ua.pt)
[4] Universidad Carlos III de Madrid, Spain ({cgarcia, jmoreno}@it.uc3m.es)
[5] Motorola Labs, Paris, France ({Christophe.Beaujean@crm.mot.com, erik@motorola.com)
[6] NEC Laboratories, Heidelberg, Germany (marco.liebsch@ccrle.nec.de)

Abstract: This paper describes an architecture for differentiation of Quality of Service in heterogeneous wireless-wired networks. This architecture applies an "all-IP" paradigm, with embedded mobility of users. The architecture allows for multiple types of access networks, and enables user roaming between different operator domains. The architecture is able to provide quality of service per-user and per-service An integrated service and resource management approach is presented based on the cooperative association between Quality of Service Brokers and Authentication, Authorisation, Accounting and Charging systems. The different phases of QoS-operation are discussed. The overall QoS concepts are presented with some relevant enhancements that address specifically voice services. In particular, EF simulations results are discussed in this context.

1 INTRODUCTION

Availability of the network services anywhere, at anytime, can be one of the key factors that attract individuals and institutions to the new network infrastructures, stimulate the development of telecommunications, and propel economies. This bold idea has already made its way into the telecommunication community bringing new requirements for network design, and envisioning a change of the current model of providing services to customers. The emerging new communications paradigm assumes a user to be able to access services independently of her or his location, in an almost transparent way, with the terminal being able to pick the preferred access technology at current location (ad-hoc, wired, wireless LAN, or cellular), and move between technologies seamlessly i.e. without noticeable disruption.

Unified, secure, multi-service, and multiple-operator network architectures are now being developed in a context commonly referenced to as networks Beyond-3G or, alternatively, 4G networks [1]. The 4G concept supports the provisioning of multiple types of services, ranging from simple network access to complex multimedia virtual reality, including voice communication services, which are themselves a challenge in packet-based mobile communications environments.

Due to the heterogeneity of the access technologies, the Internet Protocol version 6 (IPv6) is being targeted as the common denominator across multiple access technologies, and make the solution basically independent of the underlying technology - and therefore future-proof. However, fitting such important concepts as support for Quality of Service (QoS), Authentication, Authorisation, Accounting and Charging (AAAC) and mobility into the native Internet architecture poses numerous difficulties and is a real challenge.

Therefore, the primary target of this paper is to present a solution for QoS support in mobile environments[1]. In order to do so, we make frequent references to the problem of integration of QoS, AAAC and mobility. In the course of the paper we discuss the methods that let us create and exploit the intrinsic associations between the service level agreements expressed in user profiles, and the network control mechanisms capable to monitor network usage per service and per user, in order to provide these services while the user moves and the terminal changes access technologies. The proposed architecture supports network services, in a secure and auditable way. Both user-to-network interfaces and inter-operator interfaces are defined, so that multiple service providers can interoperate. The architecture is able to support multimedia services, and has been further optimised for voice services. Voice services are now among the most demanding in terms of network design, imposing hard limits on network performance. In order to handle these services we will use the Expedited Forward (EF) concept of the differentiated services framework.

In the next section we briefly describe the network environment. Section 3 describes the overall QoS architecture, while section 4 details the signalling flow of end-to-end QoS support in the architecture and presents a simulation study that allows an optimised configuration of the access routers. Finally section 5 recaps our key conclusions.

2 AN ALL-IP 4G NETWORK ARCHITECTURE

The overall 4G architecture discussed in this paper is IPv6-based, supporting seamless mobility between different access technologies. Mobility is a substantial problem in such environment, because inter-technology handovers have to be supported. In our case, we targeted Ethernet (802.3) for wired access; Wi-Fi (802.11b) for wireless LAN access; and W-CDMA - the radio interface of UMTS - for cellular access (Fig. 1). With this diversity, mobility cannot be simply handled by the lower layers, but needs to be implemented at the network layer. An "IPv6-based" mechanism has to be used for interworking, and no technology-internal mechanisms for handover, neither on the wireless LAN nor on other technology, can be used. So, in fact no mobility mechanisms are supported in the W-CDMA cells, but instead the same IP protocol supports the movement between cells. Similarly, the 802.11 nodes are only in BSS modes, and will not create an ESS: IPv6 mobility will handle handover between cells.

[1] The concepts that are presented in this paper have been developed and tested in controlled environments in the IST project Moby Dick [2] and are currently being refined.

The users/terminals may handover between any of these technologies without breaking their network connection, and sustaining voice connections. The users can further roam between administrative domains, being able to use their contracted services across domains if only appropriate agreements between those domains exist. The service providers are be able to keep track of the services being used by their costumers, both inside their own network, and while roaming. This is essential, e.g. for voice calls charging.

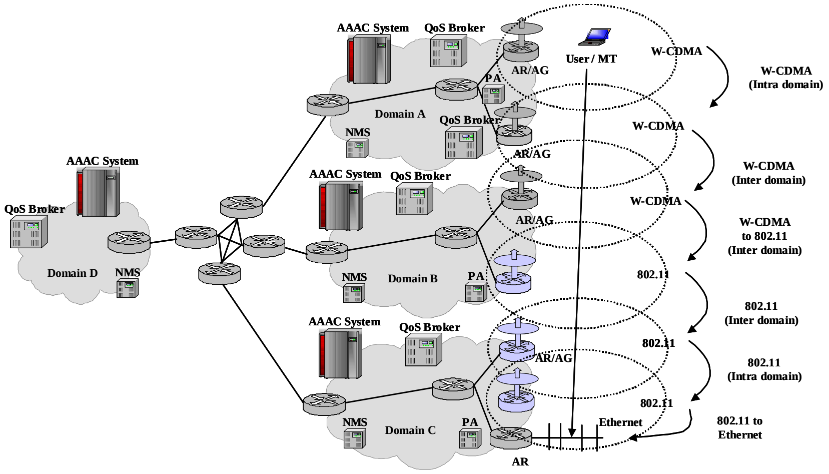

Figure 1: General Network Architecture

Figure 1 depicts the conceptual network architecture, illustrating some of the handover possibilities in such network with a moving user. Four administrative domains are shown in the figure with different types of access technologies. Each administrative domain is managed by an AAAC system. At least one network access control entity, the QoS Broker, is required per domain. Due to the requirements of full service control by the provider, all the handovers are explicitly handled by the management infrastructure through IP-based protocols, even when they are intra-technology, such as between two different Access Points in 802.11, or between two different Radio Network Controllers in WCDMA. All network resources are managed by the network provider, while the user only controls its local network, terminal, and applications.

Summarising Figure 1, the key entities are:

- A user - a person or company with a service level agreement (SLA) contracted with an operator for a specific set of services. Our architecture is concerned with user mobility, meaning that access is granted to users, not to specific terminals.
- A MT (Mobile Terminal) - a terminal from where the user accesses services. Our network concept supports terminal portability, which means that a terminal may be shared among several users, although not at the same time.
- AR (Access Router) - the point of attachment to the network, which takes the name of RG (Radio Gateway) - for wireless access (WCDMA or 802.11).

- PA (Paging Agent) - entity responsible for locating the MT when it is in "idle mode" while there are packets to be delivered to it [4].
- QoS Broker - entity responsible of managing one or more ARs/AGs, controlling user access and access rights according to the information provided by the AAAC System.
- AAAC System - the Authentication, Authorization, Accounting and Charging System, responsible for service level management (including accounting and charging). In this paper, for simplicity, metering entities are considered an integral part of this AAAC system.
- NMS (Network Management System) - the entity responsible for managing and guaranteeing availability of resources in the Core Network, and overall network management and control.

This network is capable of supporting multiple functions:

- inter-operator information interchange for multiple-operator scenarios;
- confidentiality both of user traffic and of the network control information;
- mobility of users across multiple terminals;
- mobility of terminals across multiple technologies;
- QoS levels guaranties to traffic flows (aggregates), using, e.g. the EF Per Hop Behaviour (PHB);
- monitoring and measurement functions, to collect information about network and service usage;
- paging across multiple networks to ensure continuous accessibility of users.

Simple implementations of the above functions, including management aspects, have been done with the IPv6 protocol stack in Linux. The implementation relies on MIPL (Mobile IP for Linux). Other network and stack entities required for seamless operation of terminals in this heterogeneous environment have also been developed. QoS and AAAC sub-systems are responsible of serving a user according to his service contract. They operate at the network level and at the service level respectively, and employ a differentiated services approach for QoS. Fast MIP extension [3] and security (IPSec) have also been developed and integrated in the network.

3 PROVIDING QUALITY OF SERVICE

The design principle for QoS architecture was to have a structure which allows for a potentially scalable system that can maintain contracted levels of QoS. Eventually, especially if able to provide an equivalent to the Universal Telephone Service, it could possibly replace today's telecommunications networks. Therefore, no specific network services should be presumed nor precluded, though the architecture should be optimised for a representative set of network services. Also, no special charging models should be imposed by the AAAC system, and the overall architecture must be able to support very restrictive network resource usage.

In terms of services, applications that use VoIP, video streaming, web, e-mail access and file transfer have completely different prerequisites, and the network should be able to differentiate their service. The scalability concerns favour a differentiated services (DiffServ) approach [5]. This approach is laid on the assumption to control the requests at the borders of the network, and that end-to-end QoS assurance is achieved by a concatenation of multiple managed entities. With such requirements, network resource control must be under the control of the network service provider. It has to be able to control every resource, and to grant or deny user and service access. This requirement calls for flexible and robust explicit connections admission control (CAC) mechanisms at the network edge, able to take fast decisions on user requests.

3.1 Service and Network Management in Mobile Networks

Our approach for 4G networks and to service provisioning is based on the separation of service and network management entities. In our proposal we define a service layer, which has its own interoperation mechanisms across different administrative domains (and can be mapped to the service provider concept), and a network layer, which has its own interoperation mechanism between network domains. An administrative domain may be composed of one or more technology domains. Service definitions are handled inside administrative domains and service translation is done between administrative domains [6].

Each domain has an entity responsible for handling user service aspects (the AAAC system), and at least one entity handling the network resource management aspects at the access level (the QoS Broker). The AAAC system is the central point for Authentication, Authorization and Accounting. When a mobile user enters the network, the AAAC is supposed to authenticate him. Upon successful authentication, the AAAC sends to the QoS Broker the relevant QoS policy information based on the SLA of the user, derived from his profile. From then, it is assumed that the AAAC has delegated resource-related management tied to a particular user to the QoS Broker.

However, two different network types have to be considered in terms of QoS: the core and the access. In the differentiated services approach, the core is basically managed per aggregate based on the network services, and not by user services. In that sense, core management is decoupled from the access. We assume that the Core Network is managed as the ISPs manage it nowadays or with some new management techniques that might emerge in the future (e.g. aggregation techniques). As a result, the core will have installed the capabilities required to support a voice-call, e.g..

On the other hand, on the access network, the complexity of CAC can be very large, due to the potentially complex criteria and different policies. The QoS broker issues the commands to control both ARs and RGs, configuring e.g. an EF service. The QoS Broker is thus the entity that interfaces between the user-service level and the network-service level.

3.2 Implicit "Session" Signalling

In this architecture, each network service being offered in the network is associated to a different DSCP code. This way, every packet has the information needed to the network entities to correctly forward, account, and differentiate service delivered to different packets. After registering (with the AAAC system) a user application can "signal" the intention of using a service by sending packets marked with appropriate DSCP. These packets are sent in a regular way in wired access networks, or over a shared uplink channel used for signalling in W-CDMA. This way of requesting services corresponds to implicit signalling, user-dependent, as the QoS Broker will be aware of the semantics of each DSCP code per each user (although typically there will be no variation on the meaning of DSCP codes between users). Thus QoS Broker has the relevant information for mapping user-service requests into network resources requirements and based on this information configures an access router.

A novel concept of "session" is implemented: the concept of a "session" is here associated with the usage of specific network resources, and not explicitly with specific traffic micro-flows. This process is further detailed in section 4.

Table 1: Example: Network Services

Service		**Relative Priority**	**Service parameters**	**Typical Usage Description**
Name	**Class**			
SIG	AF41	2a	Unspecified Signalling	(network usage)
S1	EF	1	Peak BW: 32 kbit/s	Real time services
S2	AF21	2b	CIR: 256 kbit/s	Priority (urgent) data transfer
S3	AF1*	2c	Three drop precedences (kbps): AF11 – 64 AF12 – 128 AF13 – 256	Olympic service (better then BE:streaming, ftp, etc)
S4	BE	3	Peak bit rate: 32 kbit/s	Best Effort (BE)
S5	BE	3	Peak bit rate: 64 kbit/s	Best effort
S6	BE	3	Peak bit rate: 256 kbit/s	Best effort
S7	Special Service Requesting AAAC Contact for specific network characteristics (DSCP, bw, etc)			

3.3 Network services offer

Services will be ofered a the network operator independently on the user applications, but will be flexible enough to support user applications

Each offered network service will be implemented with one of the three basic DiffServ per-hop behaviours (EF, AF, or BE), with associated bandwidth characteristics. Table 1 lists the network services used in the tests. The network services include support for voice communications (e.g. via S1) and data transfer services. Delay, delay jitter and packet loss rate are among the possible parameters to include in the future, but no specific control mechanisms for these parameters are currently used. The services may also be unidirectional or bi-directional. In fact, the QoS architecture can support any type of network service, where the only limit is the level of management complexity expressed in terms of complexity of interaction between the QoS Brokers, the AAAC systems and the AR that the network provider is willing to support.

Users will then subscribe to service level agreements consisting of different offerings. The operator may have a portfolio of packages composed by different criteria and targeting different groups of customers. An "Inexpensive service" can be supported at the network layer through S1, and S4 services (Table 1); and "Exclusive Pack", can be composed of S1, S2, S3, and S6. The technical translation of this "service pack" into network level services is the "network view of the user profile". (NVUP). The NVUP structure is not visible to the user, but impacts the way the services will be provided to the user, by the network.

4 END-TO-END QOS SUPPORT

Given the concepts described in section 3, the entities developed in the project can support end-to-end QoS, without explicit reservations at the setup time. Three distinct situations arise in the QoS architecture: i) registration, when a user may only use network resources after authentication and authorization, ii) service authorisation, when the user has to be authorised to use specific services; and iii) handover - when there is a need to re-allocate resources from one AR to another.

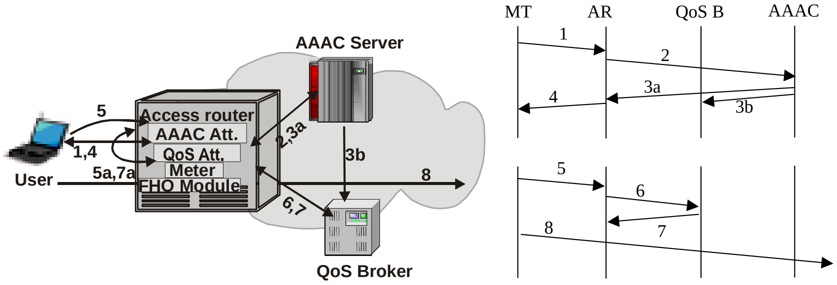

Figure 2: Support for QoS - registration and service authorisation

4.1 Registration and Authorisation

The Registration process (see Figure 2) is initiated after a Care of Address (CoA) is acquired by the MT via stateless auto-configuration, avoiding Duplicate Address Detection (DAD) by using unique layer-2 identifiers [7] to create the Interface Identifier part of the IPv6 address. However, getting a CoA does not entitle the user to use resources, besides registration messages and emergency calls. The MT has to start the authentication process by exchanging the authentication information with the AAAC through the AR. Upon a successful authentication, the AAAC System will push the NVUP (network view of the User Profile) to both the QoS Broker and the MT, via the AR. Messages 1 to 4 on Figure 2 detail this process.

The same picture shows how each network service is authorized (messages 5 to 8). The packets sent from the MT with a specific DSCP implicit signal the request of a particular service, such as a voice call (supported by network service S1, as in Table 1). If the requested service does not match any policy already set in the AR (that is, the user has not established a voice call before, e.g.), the QoS attendant/manager at the AR interacts with the QoS Broker that analyses the request and authorises the service or not, based on the User NVUP (Network View of the User Profile) and on the availability of resources. This authorisation corresponds to a configuration of the AR (via COPS [10]) with the appropriate policy for that user and that service (e.g. allowing the packets marked as "belonging" to voice call to go through, and configuring the proper scheduler parameters, as we will see in section 4.3). After that, packets with authorised profile will be let into the network and non-conformant packets will restart the authorization process once more, or will be discarded.

4.2 Handover with QoS guarantees

One of the difficult problems of IP mobility is assuring a constant level of QoS. User mobility is assured in our network by means of fast handover techniques in conjunction with context transfer between network elements (ARs - old and new - and QoS Brokers).

When the quality of the radio signal in the MT to the current AR (called "old AR", AR1) drops, the terminal will start a handover procedure to a neighbouring AR (called "new AR", AR2) with better signal and from which it has received a beacon signal with the network prefix advertisement. This handover has to be completed without user perception, when making a voice call, e.g.. For achieving this, the MT will build its new care-of-address and will start the handover negotiation through the current AR, while still maintaining its current traffic. This AR will forward the handover request to both the new AR and to the QoS Broker. The two QoS Brokers (old and new) exchange context transfer information relative to the user's NVUP and the set of services currently in use by the MT. The new QoS Broker will use this information to verify the resources availability at the new AR and, in a positive case, configures the new AR to accept the handover. The MT is then informed that the necessary resources are available at the new AR and may then perform the Layer 2 handover. During this last phase, both ARs are bicasting, to minimize packet loss. The detailed messaging is presented in the next figure.

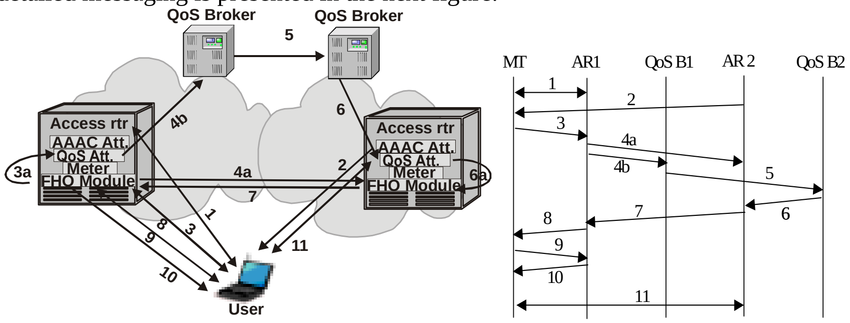

Figure 3: End-to-End QoS Support - Handover with QoS

4.4 EF PHB resource provisioning

Building an all-IP architecture based on a Differentiated Services introduces a problem of how to create per-domain services for transport of traffic aggregates with a given QoS. Per-domain services support data exchange by mixing traffic of different applications, therefore different aggregates are required to support delay-sensitive traffic, delay tolerant traffic, inelastic, elastic, as well as network maintenance traffic (e.g. SNMP, DNS, COPS, AAAC etc.).

As applications generate traffic of different characteristics in terms of data rates, level of burstiness, packet size distribution and because the operator needs to protect the infrastructure against congestion, it is very important that aggregate scheduling will be accompanied by:

- per-user rate limitation performed in the ingress routers (ARs) based on user profile,
- dimensioning and configuration of network resources to allow for a wide range of user needs and services,
- resource management for edge-to-edge QoS.

Deterministic control of the edge-to-edge delay for delay-sensitive applications can be based on mathematical formulation of a delay bound for aggregate scheduling proposed in [8] and [9] – and this is the case of voice calls, usually considered the most delay-sensitive user application. The deterministic upper-bounding of edge-to-edge delay is possible, provided that the resource utilization factor for all links in a DiffServ domain is controlled and does not exceed pre-calculated values. Based on [9] one can calculate utilization factors that will not be overdraft, and thus guarantee that the target delay will not be exceeded.

To maintain resource utilization in the entire domain, the QoS Broker is expected to know the demand, current utilisation factors of all links based on incoming calls parameters or on measurements, and on additional information such as traffic load matrix. The real data traffic is provided by monitoring functions in the network, while traffic matrixes are induced on historical profiling (and with varying degrees of complexity). The QoS Broker will then use this knowledge for admission control and resource provisioning. The mathematical formulations have the disadvantage of relying on the worst-case scenario, which leads to substantial over dimensioning.

For defining simpler heuristics for application in the QoS Broker we conducted simulations in order to get insight into the delay issue, and how it varied in function of different schedulers, and respective parameters. We evaluated a set of per-hop and per-domain behaviours supporting the typical services defined in Moby Dick [11]. The following figures show the comparison of Strict Priority (PRI), Strict Priority with rate limitation (PRIs), Weighted Fair Queuing (WFQ) and Stochastic Fair Queuing (SFQ) scheduling algorithms that were considered to serve typical traffic classes. Figures 6 and 7 present edge-to-edge average and maximum queuing delays as a function of number of hops.

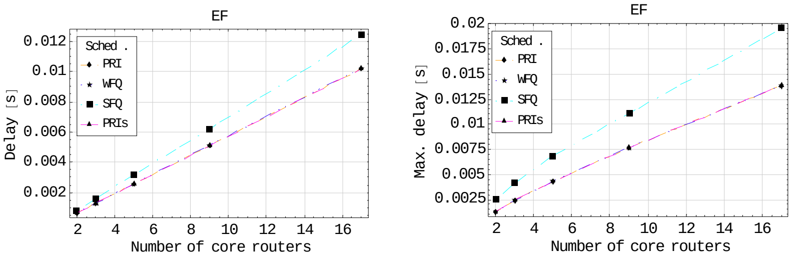

Figure 6. Edge-to-edge average queuing delay and 99,5 percentile of maximum queuing delays of VoIP for EF aggregate (EF load=12% of link capacity, link load: 100%, link rate: 10 Mbit/s).

The basic evaluation criteria was the queuing delay and the delay jitter of EF PDB for flow S1. The SFQ algorithm exhibits the worst performance of all schedulers, especially for medium and high traffic loads on a link. A better performance exhibits the SFQ algorithm at a very low load, but it applies to average delays only. PRI, PRIs and WFQ algorithms produce comparable results. For the Moby Dick architecture we are now considering to recommend PRIs, due to its simplicity when compared to WFQ.

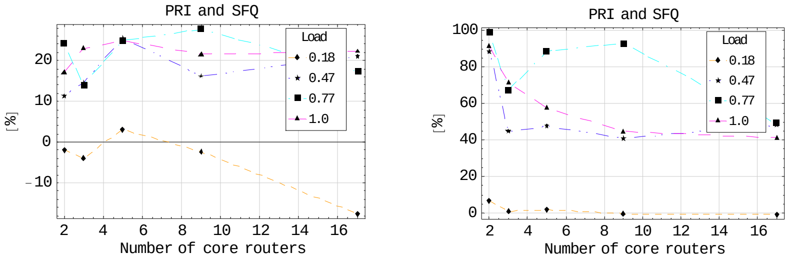

Figure 7. Left: comparison of average delay for PRI and SFQ schedulers Right: comparison of 99.5% quantile of delay for PRI and SFQ schedulers

Notice that in the simulations, a Time Sliding Window (TSW) algorithm was used for rate limitation. The main role of rate limitation was to prevent lower priority classes from being affected by higher priority classes, and is applied to all PHBs in each node. Here, we draw attention to the fact that it is very important to protect a class carrying network maintenance traffic (SIG traffic, Table 1), because this traffic plays an important role in maintaining the network infrastructure, but does not have the highest priority. When the traffic does not exceed the configured rate, the performance of PRI and PRIs is the same since the TSW does not affect traffic characteristics. The SP algorithm also fits into the Moby Dick concept of service classes.

The PRIs limitation has yet another advantage – rate limitation does not have influence on traffic characteristics when traffic level remains within limits, and the limits can be dynamically changed without inducing abrupt delay shift. For WFQ and SFQ algorithms dynamic change of bandwidth assigned for service class changes the service rate for this class, and can cause a transient increase of delay jitter. Thus this seems to be the preferable approach to support real-time services, such as voice calls.

5 CONCLUSION

We presented an architecture for supporting end-to-end QoS. This QoS architecture is able to support multi-service, multi-operator environments, handling complex multimedia services, with per user and per service differentiation, and integrating mobility and AAAC aspects. The main elements in our architecture are the MT, the AR and the QoS Brokers. We discussed the simple interoperation between these elements and depicted the overall QoS concept. With our approach, very little restrictions are imposed on the service offering. This architecture is currently being evolved for large testing in field trials across Madrid and Stuttgart.

Being an architecture specially targeted to support real time communications over packet networks, the network elements configuration must be well dissected. The simulation study summarized in the paper was a valuable input to the QoS Broker implementation and policies design, providing simple heuristics to properly configure the access routers to achieve the best possible performance. The schedulers configuration on the core routers was also determined through the results of this simulation study.

This architecture still has some shortcomings, though, mostly due to its diffserv-orientation. Each domain has to implement its own plan for mapping between network service and a DSCP, and thus, for inter domain service provision, it is essential a service/DSCP mapping between neighbouring domains. Furthermore, an adequate middleware function is required in the MT, to optimally mark the packets generated by the applications and issue the proper service requests, which requires extensions in current protocol stacks.

Nevertheless, our proposal facilitates the deployment of multiple service provision models, as it decouples the notion of service (associated with the user contract) from the network management tasks. It seems to provide a simple, flexible, QoS architecture able to support multimedia service provision for future 4G networks.

ACKNOWLEDGEMENTS

We would like to acknowledge the European Commission support and all partners in the Moby Dick consortium [12]. The support of everyone involved in the Madrid trials is also acknowledged.

REFERENCES

1. J.M. Pereira *et all*: "Fourth Generation: Now it is Personal!" Proceedings of PIMRC 2000, London, Sep 2000.

2 Hans Einsiedler et al., "The Moby Dick Project: A Mobile Heterogeneous All-IP Architecture", ATAMS 2001, Krakow, Poland, (http://www.ist-mobydick.org).

3. G. Dommety, ed. "Fast Handovers in Mobile IPv6", Internet Draft, work in progress, <draft-ietf-mobileip-fast-mipv6-3.txt>, July 2001.

4. Marco Liebsch, et all, "Paging Concept for IP based Networks", Internet Draft, draft-renker-paging-ipv6-01.txt, September 2001.

5. D. Black, S. Blake, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", IETF RFC 2475, December 1998.

6. Thi Mai Trang Nguyen et al, "COPS-SLS: a Service Level Negotiation Protocol for the Internet", IEEE Communications Magazine , Vol. 40 No. 5 , May 2002 , pp. 158-165

7. Marcelo Bagnulo, et all, "Random generation of interface identifiers", Internet Draft, draft-soto-mobileip-random-iids-00.txt, January 2002.

8. Anna Charny, J.-Y. Le Boudec, " Delay Bounds in a Network with Aggregate Scheduling ", *Proceedings of QOFIS*, Berlin, October 2000

9. Yuming Jiang, “Delay Bounds for a Network of Guaranteed Rate Servers with FIFO Aggregation”, Proceedings of ICCC 2002, New York, May 2002

10. D. Durham et al, Internet Engineering Task Force, RFC 2748, The COPS (Common Open Policy Protocol), Jan 2000.

11. Marques V., Aguiar R.,Pacyna P., Gozdecki J., Beaujean Ch., Chaher N., García C., Moreno J.I., Einsiedler H., "An architecture supporting end-to-end QoS with user mobility for systems beyond 3rd generation", IST Mobile & Wireless Telecommunications Summit 2002, Thessaloniki, Greece, June 17-19, 2002, pp. 858-862.

12. http://www.ist-mobydick.org