Heterogeneous Signaling Framework for End-to-end QoS support in Next Generation networks

Rui Prior[1], Susana Sargento[2], Diogo Gomes[2], Rui L. Aguiar[2]
[1] DCC & LIAAC, Faculdade de Ciências da Universidade do Porto, Portugal
[2] Universidade de Aveiro, Instituto de Telecomunicações, Portugal

Abstract

Next generation wireless communication systems aim to handle diverse types of services across different types of access technologies in a seamless way. This paper proposes a next generation network architecture and evaluates possible associated signaling strategies, focusing in network-level QoS support aspects. The scenarios handled cover terminal-initiated signaling, network controlled signaling, and application-provider controlled signaling. Possible message sequence charts associated with these scenarios are presented and discussed. The paper compares the relative merits of each approach and concludes that the optimum QoS signaling solution depends on the QoS models that will be used, which are directly related to the business models chosen by the operators.

1. Introduction

Next generation wireless communication systems will handle diverse types of services, across different types of access technologies. This trend, already present in 3G networks and in the current explosion of hotspots, is expected to become an universal characteristic in communications by the end of this decade. Providing mobility across domains using different access technologies in a seamless way, with no perceived service degradation for the user, is a major requisite for the next generation networks. Scalability concerns make this requirement still harder. Current wireless operators have dozens of millions of customers, and, as cell sizes decrease, handovers will become more and more frequent, potentially reaching the hundred thousands per second in a large telecom operator.

The scalable support for end-to-end QoS in such a universal mobile and heterogeneous scenario is one of the main topics in networks research nowadays. This technical problem is further compounded by the complex telecom business, with multiple types of operators foreseen in the market, covering a wide range, from basic transport to intelligent service and multimedia provision.

Although the use of the IPv6 protocol as a convergence layer much simplifies the support for seamless mobility and QoS across heterogeneous networks, the provision of multimedia and value-added services in such multi-provider environments requires a common signaling framework for session negotiation, network resource reservation plus session and QoS renegotiation. This framework must integrate application signaling and resource reservation protocols in order to ensure that enough resources are available for a good user-perceived service quality, and that the use of those resources is authorized. Thus, proper interaction between QoS and both mobility and charging mechanisms has to be in place.

The aim of this paper is the presentation of a next-generation 4G architecture (focusing on QoS-related entities) and the evaluation of the different associated signaling strategies that may be used. The paper analyzes their merits and shortcomings with regard to session setup and negotiation, session renegotiation and seamless handovers, as well as the security and flexibility provided by each signaling solution. The different concepts of signaling strategies are analyzed according to four major scenarios, discussed at session setup and renegotiation time: (i), the mobile terminal itself performs the QoS requests to a QoS Broker (responsible for resource management at the access network); (ii) a service proxy is responsible for requesting network resources to the QoS Broker; and (iii) a novel network entity, usually co-located at the access router, capable of QoS and application signaling (and signaling parsing and modification) issues the QoS requests.

The rest of the paper is organized as follows. Section 2 presents the basic components considered in the network architecture. Section 3 presents the different signaling scenarios and illustrates their main characteristics using message signalling charts. Section 4 performs a comparison between these different strategies, and section 5 discusses our key conclusions.

2. Network architecture

Next generation communication systems will aim at providing seamless mobility of users through networks with different access technologies and services. In this sense, the network needs to be capable of supporting heterogeneous access technologies. These communication systems, usually referred to as 4G networks [8], may support network technologies such as Wireless Fidelity (Wi-Fi), Universal Mobile Terrestrial System (UMTS), and new emerging technologies, such as WiMax and Digital Video Broadcast – Terrestrial (DVB-T). These technologies are quite different from each other, ranging

from Local Area Networks (LAN) to Broadcast Diffusion Networks with quite distinct network architectures. The main focus of 4G systems is the support of all these technologies under a unified network architecture capable of supporting the different access technologies, the provision of advanced services to the users, and the provision of the means for the network and service operators to increasingly develop new advanced services.

The IPv6 protocol is an adequate convergence layer to provide such unified platform. IPv6 creates an abstraction layer for services that hides technology specific parameters from advanced services. Moreover, its intrinsic support of mobility is quite important for 4G communication systems, since it provides "almost" seamless mobility between different technologies. In order to provide completely seamless mobility, extensions to IPv6 mobility such as the support for fast mobility [4] must be used. These issues, and their relationship with QoS aspects, have already been addressed in the literature (e.g. [6]).

Figure 1 depicts the proposed next generation network, supporting several access networks, each of them capable of supporting several access (wireless) technologies. This architecture allows for different operators to work in a common environment, with support for access services, other transport services, and advanced services. All operators can have special contracts between each other, federation mechanisms, enabling a better integrated service to the end user.

The Differentiated Services (DiffServ) model [1] is used to support QoS in the core network, achieving scalability and performance. The most important QoS element of the architecture is the QoS Broker, which performs admission control and manages network resources; it controls the network routers according to the active sessions and their requirements. It also performs load balancing of users and sessions among the available networks (possibly with different access technologies) by setting off network-initiated handovers. This is a quite important feature, since it provides the means to optimize the usage of operator resources.

While basic QoS services are provided intrinsically by the Access Network (AN), more advanced services are supported by a Service Provision Platform (SPP), in the core network. In the access network, service proxies are deployed for efficient service provision.

This MultiMedia Service Proxy (MMSP), aware of the requirements of user services, and the QoS Broker in the Access Network (AN QoS Broker) can have a very close relation. Merging this high level knowledge of running services with the available network resources might, for instance, enable the network to move a video stream from a Wi-Fi network to a DVB-T link, and provide the adequate network-level QoS to a multimedia stream.

For the provisioning of multimedia streaming services, MultiMedia Servers (MMServer) may also be present, located in the application server garden. The QoS definitions at the domain level are provided by a Policy Based Network Management System (PBNMS), and then proxied by the AN QoS Brokers to the Access Routers (AR) in the different access networks. For authentication and accounting purposes, an A4C (Authentication, Authorization, Accounting, Auditing and Charging) server is also present in each domain. The Core Network (CN) also has a QoS Broker, to deal with aggregates of flows traversing the core network and communication with other administrative domains.

Figure 1: 4G Network Architecture

4G networks must support all types of services. Although the support of multimedia services can be provided by means of interaction between the MMSP and the QoS Broker, this solution is not suitable for IP legacy applications, which may equally have service quality requirements [6]. In order to provide QoS to those legacy, QoS unaware applications, some advanced functions must be added to the access routers. The required functionality comprises connection tracking, similar to what may be found in a Network Address Translation (NAT) router with port translation, per-application flow DiffServ Code Point (DSCP) marking, and the means to translate other QoS reservation mechanisms, such as Integrated Services (IntServ) [9] resource ReSerVation Protocol (RSVP) reservations, into DiffServ DSCP marking and QoS Broker requests. We refer to the entity supporting all these functions as the Advanced Router Mechanisms (ARM). The ARM provides functionality equivalent to a basic proxy without the need to change any of the legacy applications, and can be considered as a dedicated intelligent transparent proxy. Note that the ARM can also perform application to network level QoS mapping for multimedia services, e.g. for Session Initiation Protocol

(SIP) [7] services, issuing the resource reservation requests to the QoS Broker and filtering the QoS configurations in the application signaling messages [11]. The ARM may, therefore, perform the QoS related functions that are typically required also from a MMSP, if the operator so desires, since the application/service logic allows these operations to be delegated to the ARM.

This 4G network provides overall control and management mechanisms, relying on the overlay setup of a distributed set of QoS Brokers, and achieving QoS at the access link by means of explicit reservation in an IntServ like manner. In order to coordinate all these mechanisms, several signaling strategies arise. Implicit signaling in the DiffServ environment is very simple, but requires applications to produce marked packets with the right DSCP code, and reduces the control flexibility achieved by IntServ-like reservations in the access network, where (radio) resources are scarce. Explicit signaling can be done by application signaling protocols (for example, RSVP messages or using SIP and its companion Session Description Protocol – SDP [3] to describe the required session and QoS parameters) and interaction between the MMSP and the QoS Broker. These solutions are more complex and involve a larger set of signaling messages, but increase the flexibility of the characteristics of the services to be offered to the user. These different solutions are very closely tied to the QoS service model that will be used, which is directly associated to the business model chosen by the operators: (1) application oriented, IntServ like, (2) user oriented, where the user asks for the service characteristics he wants to, and (3) service oriented, where the user has some well known contracted services with the network and/or service operator. Note that these QoS models are not disjoint and independent from each other; they are closely related to the QoS signaling strategies in place, which will be discussed in the next sections.

The proposed 4G network architecture holds several advantages when compared with other solutions. For instance, when considering the UMTS (Universal Mobile Terrestrial System) [1] system, it is apparent that our approach allows the support of heterogeneous access technologies, which permits optimization of the coverage/performance/cost factor under very different utilization scenarios. Contrary to UMTS, where handovers are performed at layer 2 without change in the QoS parameters, in our 4G system handovers are handled at layer 3 and are coupled with triggers for session renegotiation, allowing for features such as the automatic increase in the quality of a videoconference when arriving at a 802.11 HotSpot or the dropping of the video component of a multimedia call without dropping the call when leaving the HotSpot. The centering on layer 3, using IPv6 as a convergence layer, also leads to a simpler stack in 4G than that of UMTS, which translates in less overhead. Finally, the decoupling of PDP and proxy functions, which in UMTS are performed by a single element, the P-CSCF (Proxy Call Session Control Function), frees the 4G networks from being tied to a single application protocol (SIP in the UMTS case). This flexibility is important not only to simultaneously support different applications requiring QoS, not necessarily related to multimedia (QoS for legacy applications is supported by the proposed architecture), but also in order to make the investment in infrastructure future-proof: the network infrastructure should be able to provide support for a new protocol providing functions not possible to implement in SIP with minimal changes.

Other proposals for 4G architectures have been made, e.g. [13] and [13], and in fact our work has been influenced by the previous work performed in [6]. In general, these proposals are also based on IP-core networks, although oriented towards different scenarios. In broad terms, our architecture is more flexible, and presents a more comprehensive set of characteristics, such as: a fully integrated approach to IP-based communication with different types of applications and protocols (e.g. both legacy and SIP-based applications are supported, as described in Section 3), including adaptive applications; the customization/optimization of the architecture according with the expected service mix to support; and the integrated support of multiple QoS service models, according to the overall network configuration (defined by operator policies).

All the signaling strategies that will be presented below are based on this architecture, and therefore, are based on the QoS Broker concept for resource reservation. New signaling methodologies are being thought for the support of QoS in mobility environments, like the ones being defined by the Next Steps In Signaling Working Group (NSIS WG) [12]. However, these signaling approaches can be used both in distributed and centralized admission control approaches, and therefore, could be used in this architecture for the signaling in the access network. This is a topic for further work.

3. QoS Strategies

This section describes the different QoS signaling strategies that may be adopted for the provisioning of legacy and value added services in this 4G environment. Although all the strategies are able to support the three QoS service models referred above, the MMSP strategy is targeted at application oriented models, whereas ARM and terminal strategies are, respectively, more targeted to network service and user oriented models, as we will see.

Two main types of services are analyzed here: a multimedia conference service, initiated using an out-of-band service initiation protocol; and a more “traditional” data transfer service. Both services may require end-to-end QoS support. The study of the strategies adapted to these different services will consider two phases in the service provisioning lifetime: session initiation with QoS

support and session renegotiation due to intra-domain fast handovers as specified in [4].

For multimedia services, the architecture is not tied to any particular session setup protocol. Therefore, our analysis is based on a general application signaling protocol, App_Sig, with 3 messages for a session setup: Initiation, Reply and Ack. These messages are easily mapped into real protocols: in SIP, for example, they roughly correspond to the INVITE, OK and ACK messages.

Some assumptions are made in the signaling scenarios. We assume that, at registration time, the A4C sends subsets of the user profile, the Network View of User Profile (NVUP) [6] and the Service View of User Profile (SVUP), respectively to the AN QoS Broker and to the MMSP, along with information provided by the terminal on its network and service capabilities. Therefore, the QoS Broker can act as a Policy Decision Point (PDP) [10], performing decisions based both on the user profile and the terminal's capabilities, in addition to resource availability, without consulting the A4C every time a service is initiated or handover occurs. Similarly, the MMSP may perform service level authorization without resorting to the A4C. When an update to the user profiles occurs in the A4C it asynchronously pushes the new profile(s) into the QoS Broker and/or MMSP. This way, service initiation and handover are faster, and the load on the A4C is reduced. This assumption is inadequate in the case of pre-paid services; in that case, communication between both the QoS Broker and MMSP with A4C is required during the session lifetime, but this would only complicate our discussion, without affecting the analysis.

Furthermore, in the following diagrams, resource reservation requests are issued to AN QoS Brokers but no reservations in the core are shown. In fact, due to scalability reasons, resource reservation at the core is performed on an aggregate basis, and not per-flow. Communication exists between AN and CN QoS Brokers for providing the former with the necessary information regarding resource availability at the core, but that message exchange is decoupled from per-flow service signaling, and is, therefore, out of the scope of this paper.

3.1. Signaling: Terminal

The first strategy is based on network resource requests issued from the terminal. This strategy is general enough to support all types of applications. Adding support for a new application merely requires a software update in the terminal. In this case, the terminal has a module, referred to as QoS client, which is able to issue QoS requests and map the application parameters and QoS requirements into network parameters.

Figure 2 illustrates a multimedia session initiation in this scenario. The terminal (MT1) begins by mapping the application requirements to network service and QoS requirements. It then sends a request to its local QoS Broker – QoSB1 (indirectly, via a QoS attendant at the access router – AR1) with information on the required network and QoS parameters for the session. The QoS Broker answers with information on the services that may be used according to the user profile and the current network status. This step prevents the terminal from trying to initiate services that cannot be supported by the network, or that the terminal is not allowed to use, in face

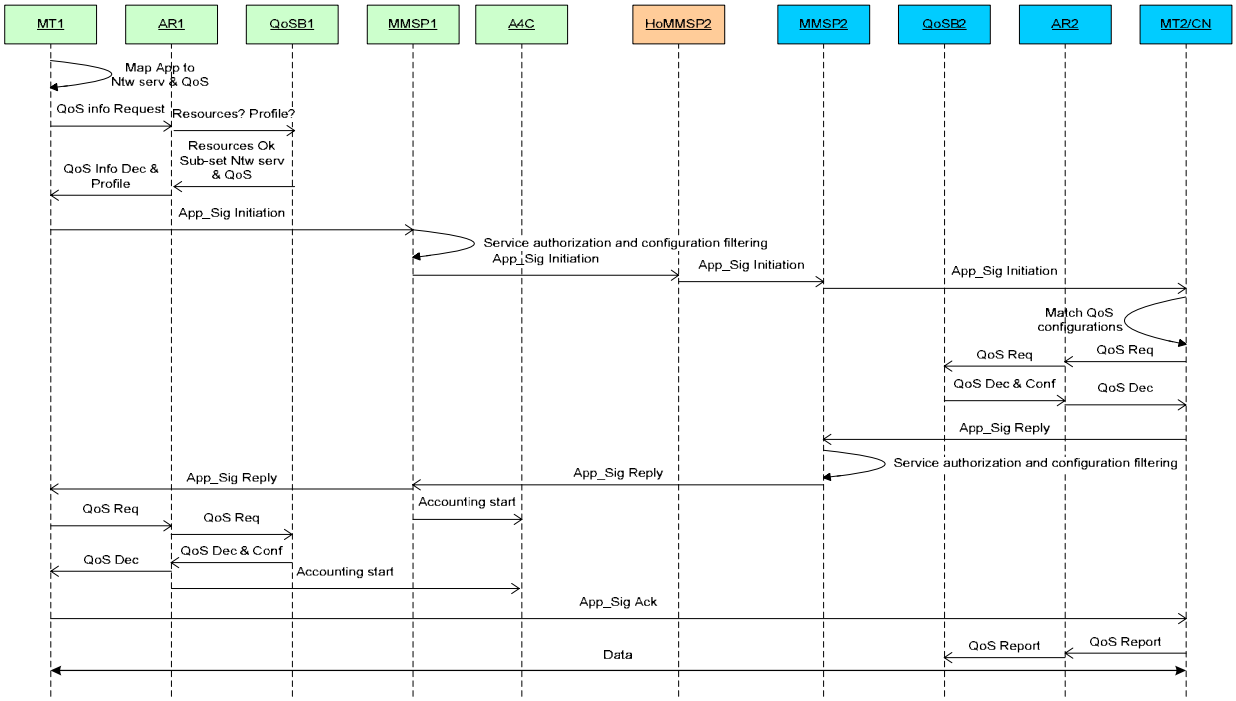

Figure 2: Multimedia conference setup with reservations issued by the terminal

of the user profile (subscribed services).
If allowed by the QoS Broker, the QoS client module in the terminal sends the App_Sig Initiation message, a service initiation message containing a set of QoS configurations corresponding to the composition of the terminal capabilities and the possible service set indicated by the QoS Broker, for negotiation with the other terminal. The MMSP in the caller domain (MMSP1) may then perform service level authorization: in the case of some service in the requested set not being allowed, it will be filtered. Notice that this authorization and filtering is optional, since the SVUP may indicate that all services are allowed as long as adequate transport is supported by the network.
Since the location of MT2 (correspondent node) is not yet known, the App_Sig Initiation message is forwarded to its home proxy (HoMMSP2), which in turn forwards it to the network where it is currently connected. On receiving this message, MT2 matches the set of service configurations to its own capabilities and issues a request to its QoS Broker (QoSB2), asking for resources for the matched services. In the case of a positive answer from the QoS Broker, MT2 issues an App_Sig Reply message containing the set of configurations supported by both terminals, constrained by the QoS Broker's answers. Upon receiving the App_Sig Reply message, the MMSP2 will also perform service authorization for the receiver, and forwards the message to the sender, possibly filtered again due to service authorization considerations.
Finally, MT1 issues a QoS request to the broker considering the set of QoS configurations supported by both sides. Since MT1 is now aware of MT2's Care-of Address, the request is directly sent to its current physical location.
Accounting start messages are sent to the A4C from AR1 and MMSP1 to initiate, respectively, the transport-based and the service-based accounting processes. These messages can be sent independently in both caller and called domains, depending on business policies.
The App_Sig Ack message contains the final QoS configuration to be used. If it is different (lower) than the previous reservation in the access network of the callee (MT2), a QoS report is sent to AR2, which forwards the information to the QoS Broker. At this time, the multimedia data can be transmitted between the two end-points with QoS guarantees.
The initiation of a legacy data service with QoS requirements in this terminal-based scenario is illustrated in Figure 3. The data service can be, for example, a File Transfer Protocol (FTP) session. If the application is QoS unaware, the QoS client module in the mobile maps the application's requirements to network resources and issues the appropriate requests to the QoS Broker on its behalf (it is necessary to support legacy applications without modifications). If the application is QoS aware, this module is bypassed.

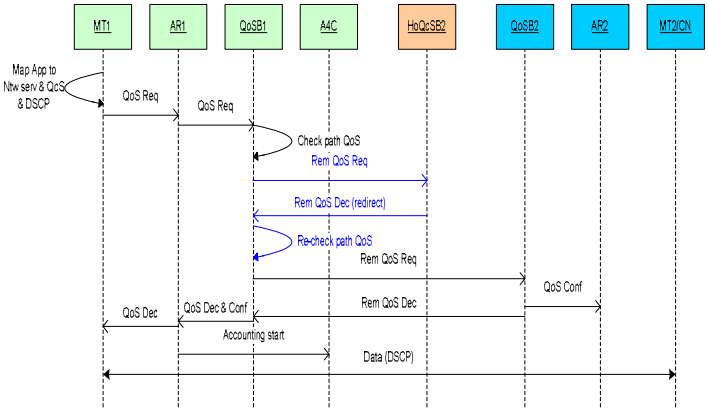

Figure 3: Legacy application setup with reservations issued by the terminal

After mapping the application requirements to the appropriate class of service and amount of necessary resources, a request is sent to the QoS Broker, which checks the availability of resources and then sends the request to its counterpart on the called network for local resource checking. If there is no entry for MT2 in MT1's binding cache, this request is made to its Home Address; therefore, it goes to the QoS Broker at MT2's home network. If MT2 is away, the Home QoS Broker redirects QoSB1 to the QoS Broker of the network where it is currently attached. QoSB1 checks for resources on the new path and sends the request to QoSB2. After admission control for local resources, QoSB2 configures AR2 for the service. When the remote decision arrives, QoSB1 communicates its final decision and configures AR1. An Accounting Start message initiates the accounting process, and data packets may flow with QoS.
Notice that the previous case considered explicit signaling for the QoS request. In the case of implicit signaling, after mapping the application to a network service (QoS client), the terminal sends the first application packet marked with the appropriate DSCP for the desired service [5]. On receiving this packet, both ARs issue requests to the respective QoSBs and resources are reserved in both domains similarly to the previous case.
A different problem to discuss is the handover process, which could be triggered either by network optimization aspects or by user movement. This handover process is shown in Figure 4 for a multimedia session. The proposed handover and renegotiation processes work as follows. The network uses fast handover concepts/messages, similar to those defined in [4], associated with some network information discovery mechanism, for example, the Candidate Access Router Discovery (CARD) mechanism [5], to propagate information to the user on the prospective networks for handover. In the case of a user-initiated handover, upon having this information from CARD, the terminal sends a Router Solicitation for Proxy (RtSolPr) message with information on the new

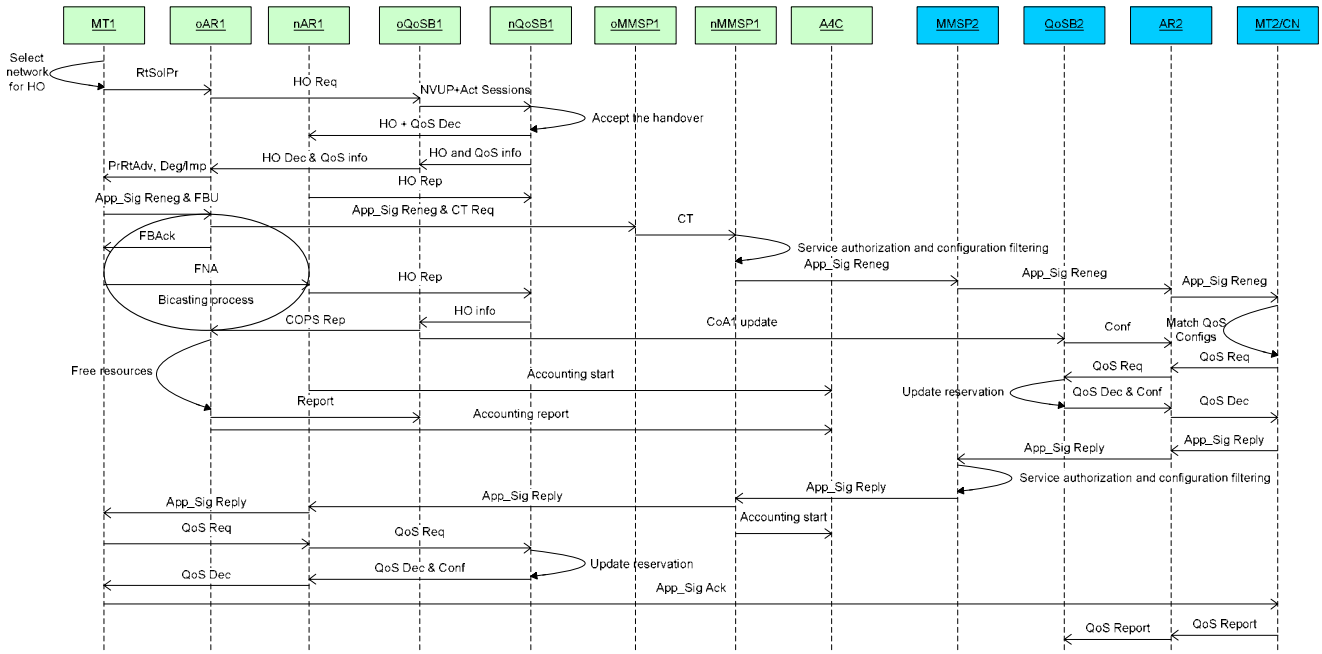

Figure 4: Handover with service renegotiation with reservations issued by the terminal

network to perform handover. The old AR (oAR1) sends a handover request message to the old QoS Broker, which pushes the NVUP, along with information on the set of active sessions, to the QoS Broker of the prospective network (nQoSB1). If the nQoSB1 accepts the handover with the required characteristics, the new AR is configured and the decision is communicated to the oQoSB1 and to the terminal by the Proxy Router Advertisement (PrRtAdv) message.

If service degrading is required or improvement allowed (information that sent in the PrRtAdv message), the terminal renegotiates the session parameters, sending an App_Sig Renegotiation message together with the Fast Binding Update (FBU) confirming the handover. The FBU indicates that the terminal will move and triggers a bicasting process [6], where each packet sent to MT1 via the old network is duplicated at oAR1 and also sent via the new network. Furthermore, it also triggers a Context Transfer (CT) Request, sent to the oMMSP1.

In case of service degrading, if the bicasting process starts before the session renegotiation finishes, more traffic will be sent to the new network than it can handle. There are several possible solutions, trading-off performance for handover precision: (1) the handover is only performed after the renegotiation process, which can have the problem of starting the handover too late, after the user actually moved; (2) provide content adaptation for the new network, which substantially increases complexity, and (3) the nQoSB1 can decide if it is possible to handle that amount of traffic for the time interval required to complete the renegotiation process. In the last case, if there is sufficient available bandwidth to handle the larger amount of traffic in the renegotiation time interval, or if it is possible to degrade some services with worse profile, the handover is performed before the renegotiation process finishes.

The Fast Neighbor Advertisement (FNA) message informs the new AR1 that the handover was completed. Both QoS Brokers are informed of the fact, and the bicasting process stops, since the terminal is no longer receiving information via the old network. Furthermore, oQoSB1 informs QoSB2 of MT1's new CoA; QoSB2 may then update filter configurations at AR2.

Meanwhile, the session renegotiation process continues. If possible, the reservation is updated in the receiving access network. The QoS Request sent by MT1 in the new network is likely to not fail since it is based in recent information from the network. After the App_Sig Ack message, the multimedia session continues with the renegotiated session and QoS parameters.

In this section we considered the case of a user-initiated handover. In the case of a network-initiated handover, the user just receives a notification from the network (oQoSB1) to perform handover, and the rest of the process is similar to the user-initiated handover.

3.2. Signaling: Multimedia Service Proxy

In the second strategy, signaling is performed through an intelligent proxy server, capable of parsing QoS configurations and mapping them to network resource requirements. The proxy issues resource reservation requests to the QoS Broker, freeing the terminal from this burden. Besides issuing QoS requests, the proxy may also apply policies configured by the operator concerning the services allowed by the user contract, based on a subset of the profile (SVUP) pushed to the proxy by the A4C at

registration time. Since it does not apply to legacy applications, this scenario is especially appealing to environments where value added services are a priority. This scenario is quite similar to scenarios where application servers are performing QoS signaling – but in this case the "proxy functionality" would reside in the application server node, and not in a machine in the middle of the communications.

A multimedia conference in this scenario is initiated as shown in Figure 5. The terminal starts the process by sending an App_Sig Initiation message, containing a set of QoS configurations. When receiving this message, the proxy, MMSP1, queries the QoS Broker on the resources allowed to the user, in face of the user profile and the load at the access network. A pre-reservation is performed in QoSB1, which answers with the allowable network services and QoS. The proxy performs service authorization and modifies the App_Sig Initiation message according to this answer (filtering some set of services and QoS configurations). Once again, since the current location of MT2 is not yet known, the Initiation message is forwarded to the proxy at MT2's home (HoMMSP2). HoMMSP2 knows MT2's visited network and forwards the message to the respective proxy (MMSP2).

On receiving the Initiation message, MT2 matches the QoS configurations to those it supports, and sends an App_Sig Reply with the common set. MMSP2 sends a request to the QoS Broker (QoSB2) and, based on the QoSB2's decision, performs service authorization and filters the set of QoS configurations. When the App_Sig Reply arrives at MMSP1, it performs a request to the QoSB1 corresponding to configurations supported by both sides. Since MT2's location is now known, QoSB1 makes a final decision taking into account the availability of resources along the path. If this implies further restriction on the QoS configurations, the MMSP1 filters the App_Sig Reply accordingly.

The App_Sig Ack message contains the final configuration that will be used. QoS Report messages inform both QoS Brokers of the amount of resources that will actually be used, triggering the configuration of the access routers (note that in the case of SIP sending the ACK message via the proxies implies the use of the RecordRoute/Route mechanism); Accounting Start messages are sent by both AR1 and MMSP1 to the A4C. These messages allow for service- and/or transport-based charging. Again, in the case of shared payment between MT1 and MT2, messages may also be sent to the A4C in the receiving domain.

3.4. Signaling: ARM

Similarly to the previous scenarios, in the ARM-oriented scenario the terminal performs only application level signaling, but now the Access Router (through the ARM capabilities) performs the application to network level QoS mapping, issues the resource reservation requests to the QoS Broker and filters the QoS configurations in the application signaling messages. Since the AR is always on the data path, legacy applications with in-band signaling only are equally supported by this scenario: for example, when the ARM sees a TCP SYN packet with destination port 23, it knows a Telnet service is being started. Legacy Applications need extra intelligence in the ARM in order for it to identify and request proper QoS for the application, since no QoS information is present in

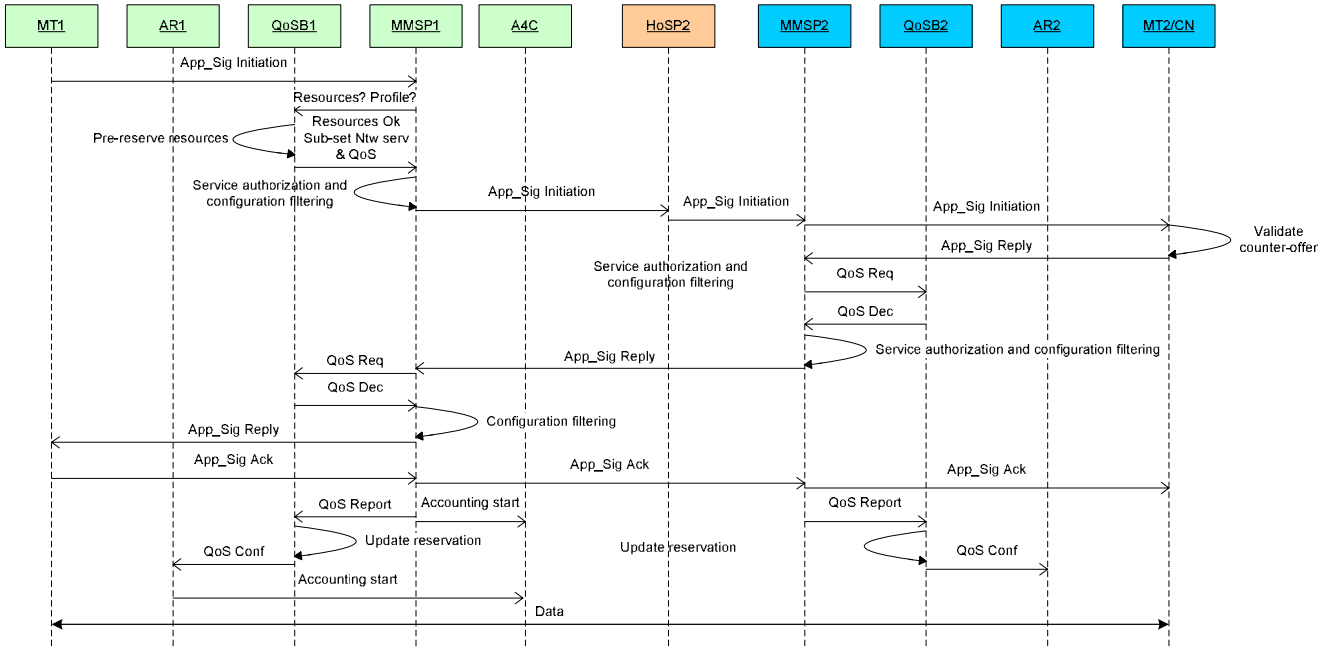

Figure 5: Multimedia conference setup with reservations issued by the MMSP

the initiation message. The information needed to perform this action should be supplied by the QoS Broker on the boot-up of the Access Router. The information on QoS profiles for legacy applications comes directly from the PBNMS, and reflects a mapping of operator business models into network optimization policies. This aspect is quite an important feature for the ARM concept, as it enables it to be in a strict relationship between operator business models and deployed QoS for user applications. Since the ARM has to be efficient, this scenario is preferred for business scenarios where a set of simple well-known services must be universally supported.

A legacy application setup is shown in Figure 6. It is very similar to the terminal scenario (Figure 3): the main difference is that neither the application nor middleware in the terminal request QoS, the application merely starts sending unmarked data packets. When AR1 receives the first packet from the terminal, with in-band signaling, the ARM infers the application being used and requests the appropriate resources to the QoS Broker, appropriate being defined by the operator policy. Notice that the data packet is buffered at the AR until the final response comes from the QoS Broker along with the configuration.

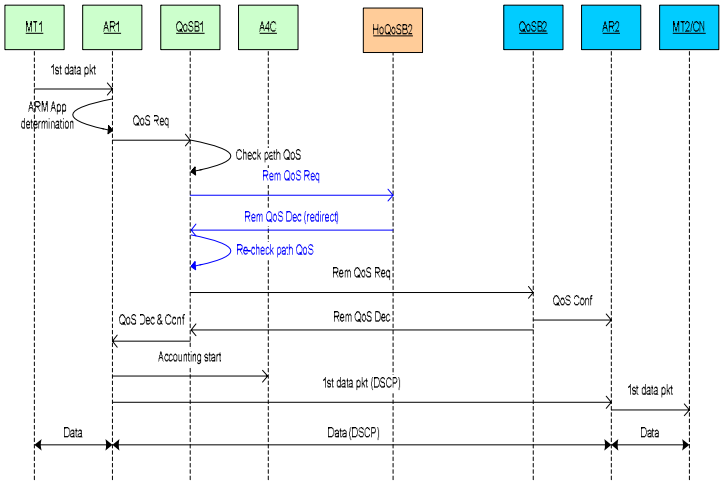

Figure 6: Legacy application setup with reservations issued by the ARM

The need for QoS Broker to QoS Broker requests is related to the charging model: this sequence considers that only the caller pays, therefore an explicit request must be sent from network 1 to network 2. If a split charging model is considered, the initiation may be simpler and more efficient, as the ARM at each AR identifies the application and performs a request to the respective QoS Broker.

A multimedia session setup using an ARM scenario is illustrated in Figure 7. On receiving an App_Sig Initiation message, the ARM verifies application and QoS requirements, maps them to network resource requirements, and queries QoSB1. The ARM can then modify the App_Sig Initiation message depending on the answer received. At the called network, the ARM module sends a request to QoSB2, which responds according to the local network status and cached information on MT2's capabilities. If necessary, the ARM further modifies the App_Sig Initiation message. In the case of a positive answer from QoSB2, MT2 matches the QoS configurations in the App_Sig Initiation message to those it supports and answers with a Reply message containing the common set of QoS configurations. Knowing MT2's CoA, the ARM at the caller network issues a request to QoSB1 taking into account the set of commonly supported configurations and the availability of resources to the network where MT2 is attached. The App_Sig Ack contains the final configuration. If it requires fewer resources than where previously reserved, the ARMs reduce the provisioned rates and send QoS Report messages to inform the Brokers of the fact. Since the ARM performs control at service level, a single accounting start message from the AR contains enough information for service- and/or transport-based charging.

4. Comparison of the QoS Strategies

This section presents a comparative analysis of the different scenarios, accounting for technical, QoS model and business issues.

The scenario where network resource reservations are directly requested by the terminal is general enough to support all types of services and applications without requiring special support from the network. Adding a new service is a simple matter of installing the appropriate software on the terminal. On the other hand, the application (or the "installed" QoS client) must be capable of requesting network resource reservations. The amount of infrastructure required from the operator is minimal, since an operator may provide only a transport service with QoS; operators wishing to upgrade to more advanced services provision may then deploy larger infrastructures. The great flexibility in terms of services provided by this scenario is obtained at the cost of increased terminal complexity since service intelligence is pushed to the terminals. In terms of privacy, this is a favorable scenario; since the operator is not concerned with the applications, all data may be encrypted, only its destination is known. In handovers, the requirements for context transfer are minimized, since the terminal is the common element in a handover. In business terms, this scenario is especially appropriate if the main service sold by the operators is transport with QoS.

The scenario where proxy servers control resource reservations and perform call admission control is obviously only valid for applications that use proxy servers, and is particularly suitable when QoS information is included in application signaling messages. While many other services, including legacy data transfer services such as FTP, may be wrapped in a protocol with

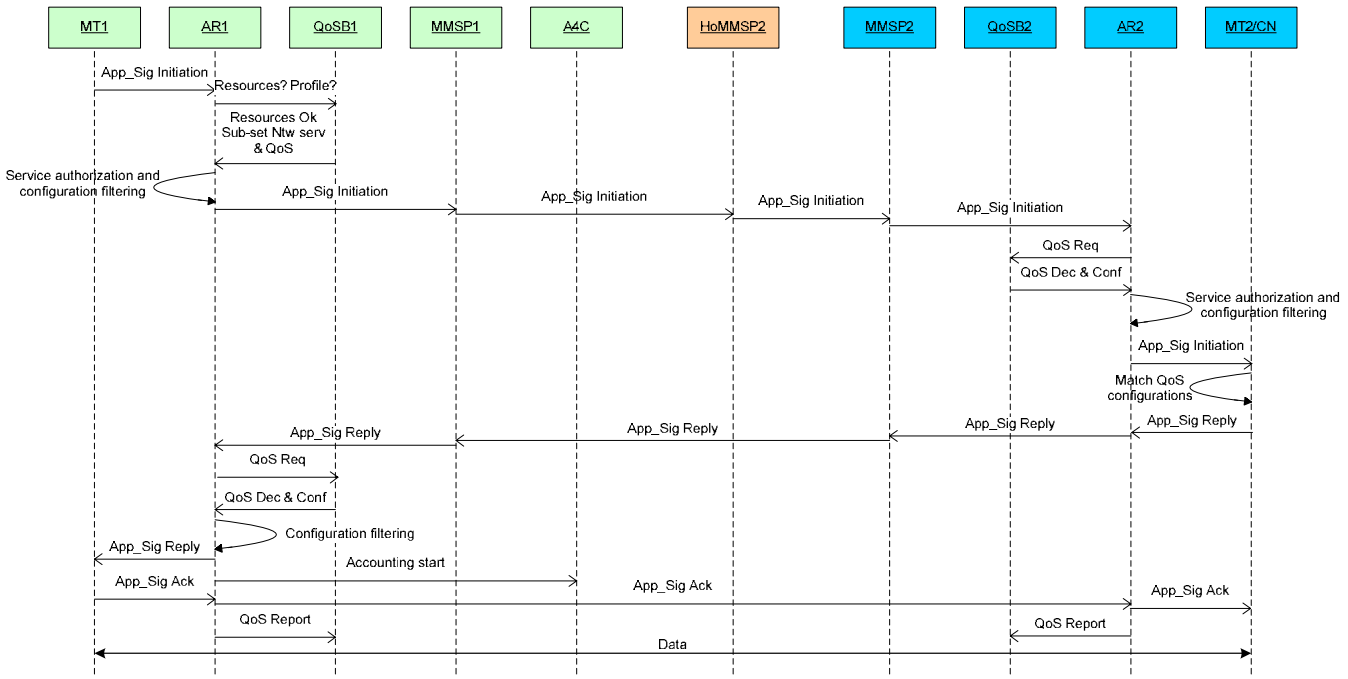

Figure 7: Multimedia conference setup with reservations issued by the ARM

those characteristics (SIP, for example), it is unnatural and cumbersome to do so, adding unnecessary complexity to the terminals and defeating one advantage of this scenario: for proxied applications the terminal is simplified by only performing application level signaling. For multimedia services, the proxies actively limit the set of QoS configurations inside the application signaling to those allowed by the user profile and for which network resources are available, improving the efficiency of session setup and negotiation. Moreover, with this approach current and future applications may be used with minimal or no need for modifications. However, in this case, proxies for all relevant application signaling protocols must be provided by the network, increasing the amount of required infrastructure. Proxies are a fundamental piece in this scenario, and the network operator itself must own them in order to have the complete control over the network and signaling. Note that, in order for the reservations to be optimal, the application signaling protocols must include all relevant QoS parameters.

Proxies that perform and keep track of network resource reservations need to be stateful, implying that context transfer between proxies must be performed during handovers. Additionally, handovers must be coordinated with application signaling. Notice that in order to avoid performance bottleneck in proxies, load-balancing solutions must be implemented, which may further complicate the context transfer process (and increase the amount of equipment required). Finally, regarding privacy, this scenario is less favorable than the previous one, since the operator must be involved in all aspects of the application signaling.

In the ARM scenario, complexity is pushed to the network edge. Although scalability in proxies may be achieved by load balancing among a number of different servers, a solution where the signaling parsing and reservation initiating entity is as close as possible to the terminals is scalable without special needs for load balancing, since a smaller number of terminals will request its services. Although not as scalable as the first solution, it allows the use of simple terminals incapable of performing QoS requests. Regarding signaling complexity, since the AR is naturally in the signaling path, acting as PEP (Policy Enforcement Point) [10] at transport level, less signaling is required in this scenario. Moreover, in case of handover, no additional entities are required to perform context transfer, and the handover is easily coordinated with the application signaling, since the AR is naturally involved in the process. In this case, the ARM controls both the handover process and the session renegotiation. With the ARM, QoS for legacy applications is easily supported without modifying them or requiring middleware in the terminal, and adding support for another application signaling protocol requires only a software update to the AR (push of a new application translation module). In terms of application support this model is as flexible as the first: since the ARM is capable of trans-signaling [11] (which is the ability to translate a signaling protocol into a different one by the network operator infrastructure, on request of the operator), minimum support can be provided even for application signaling protocols with no corresponding

ARM module. Overall, this is the most flexible scenario, since it provides a choice between dumber terminals using only the limited set of well-known services supported by the ARM and more intelligent and costly terminals that support any application. Both service- and transport-based charging are easily supported. The ARM, however, needs to maintain some state machine consistency with the application signaling, and signaling messages cannot be encrypted, as they may be processed by the ARM. This last aspect may raise privacy issues if the network operator is not trusted. The other disadvantage of the ARM is centered on its performance, since it must perform access control and intelligent processing simultaneously; however, given the relationship between wireless link capabilities and computing power, this does not currently seem to be a major concern.

5. Conclusions

This paper presents a 4G communication system, based on IPv6. The paper addresses the problems of QoS signaling in this network, and discusses the different entities involved in this process, both for the case of "traditional" (legacy) applications and in the case of novel multimedia applications (such as SIP telephony). Several scenarios are here analyzed, discussing the relationships between application-level and network level QoS signaling, and how this interrelation can be established: centered on the terminal (PDA, intelligent cellular phone), on service proxies (e.g. SIP proxy or application servers), and in Advanced Router Mechanisms.

Although presented and analyzed separately, these scenarios are not necessarily mutually exclusive. More than one scenario may be supported by the network, and their usage can be determined by the type of application and also by the QoS model being adopted by the operator. In fact it is clear that different models are especially adequate to different types of "services", and thus will depend on the services which the operator wants to support. In our view, each QoS scenario should be associated to a QoS model (user, network service or application service oriented), being directly related to the business model chosen by the network operator.

With such a diversity of services, it is clear that the flexibility to change network behavior through the use of policies is an important issue to consider. In this aspect, both the ARM and the proxy scenarios seem to be the most flexible to deploy, with the former having the advantage of handling both legacy and advanced multimedia applications.

6. Acknowledgements

This work is in part supported by the EU Framework Programme 6 for Research and Development Daidalos project (IST-2202-506997).

7. References

[1] 3GPP TS 23.002 V6.4.0, *Network Architecture*.

[2] S. Blake (ed) et al., *An Architecture for Differentiated Services*, IETF RFC 2475, December 1998.

[3] V. Jacobson, *SDP: Session Description Protocol*, IETF RFC 2327, April 1998.

[4] R. Koodli (ed.), *Fast Handovers for Mobile IPv6*, Internet Draft, October 2003.

[5] M Liebsch et al., *Candidate Access Router Discovery*, Internet Draft, December 2003.

[6] V. Marques, R.L. Aguiar et al., *An IP-Based QoS Architecture for 4G Operator Scenarios*, IEEE Wireless Communications, June 2003.

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

[8] Y. Kim et al., *Beyond 3G: vision, requirements, and enabling technologies*, IEEE Communications Magazine, Volume: 41, Issue: 3, March 2003. Pages: 120–124.

[9] Braden, R., Clark, D. and S. Shenker, *Integrated Services in the Internet Architecture: an Overview*, RFC 1633, June 1994.

[10] R. Yavatkar, "A Framework for Policy-Admission Control", IETF RFC 2753, 2000.

[11] D. Gomes, R.L. Aguiar et al., "A trans-signalling approach to strategy for QoS support in heterogeneous networks", ICT'2004, International Conference in Telecommunications, eds: José Neuman de Souze, Petre Dini, Pascal Lorenz, ISBN: 3-540-22571-4, pp 1114-1121, Springer Verlang.

[12] NSIS Working Group Charter, http://www.ietf.org/html.charters/nsis-charter.html.

[13] D. Wisely, E. Mitjana, "Paving the Road to Systems Beyond 3G - The IST MIND Project", Journal of Communication and Networks, December 2002.

[14] Joachim Hillebrand, et al., "Quality-of-Service Signaling for Next-Generation IP-Based Mobile Networks", IEEE Communications Magazine • June 2004, pp72-79.