A Modular Approach to Mobile QoS Signaling - Motivation, Design & Implementation.

Nicole Berier, Martin Karsten, Jens Schmitt and Ralf Steinmetz
Multimedia Communications Lab - KOM, Darmstadt University of Technology, Germany[1]
{Nicole.Berier,Martin.Karsten,Jens.Schmitt,Ralf.Steinmetz}@kom.tu-darmstadt.de

***Abstract* -- In order to support mobile multimedia applications in next generation wireless IP-based networks, it is necessary to deliver seamless voice, video and data at high quality. Therefore, session mobility and Quality of Service (QoS) for mobile end systems are required. Within this article, the authors point out a new way to approach the problem. Instead of tightly coupling a modified QoS signaling mechanism with a certain mobility mechanism, a more generic and long-term solution is proposed and exemplified on the basis of existing IETF protocols. The connection-less IP network layer is enhanced by a lightweight and truly optional connection-oriented mobile network service, which offers the possibility to establish soft state unicast connections at the network layer. Hence, a connection-oriented network service is available within a radio access network (RAN) architecture to all end systems - mobile or fixed - independent of the application. Thereby, it is possible to integrate QoS and connectivity signaling for mobile end systems, as well as other connection-oriented services like explicit routing or load balancing.**

1. INTRODUCTION

Future multimedia networks are expected to support ubiquitous computing, which enables users to access a wide spectrum of local or remote applications via all kinds of end systems anytime and anywhere. Applications will vary from simple low bandwidth applications with no real time characteristics to high quality audio or video applications, which might require specific transmission performance assurances. Within this scenario, end systems may vary from large fixed devices like huge information touch screens to tiny mobile devices. Therefore, future multimedia networks need to provide connectivity to wired and wireless end systems as well as mechanisms supporting network layer mobility for seamless integration of fixed and mobile end systems.

Obviously, this vision of future multimedia networking gives rise to several challenging questions and problems, which have to be solved first. Not only feasible applications and end systems have to be developed but also smart network technology has to be devised. To support future multimedia networking in IP-based networks, Internet technology has to change: besides the ability to connect a huge amount of fixed and mobile end systems seamlessly, high bandwidth and quality constraints in wired and wireless networks have to be satisfied as well.

This article focuses on the networking aspects, especially on the integration of mobility into

---

[1] The work is partially funded by Siemens AG, Dept. CT IC 2, Munich, Germany.

IP-based *radio access networks (RAN)*. In order to support mobile multimedia applications in RANs, it is necessary to seamlessly deliver voice, video and data at high quality. Therefore, the support of *session mobility*, which allows seamless network layer connectivity while staying flexible with regard to the underlying transmission medium as well as achieving stable or at least predictable transmission performance - usually termed *Quality of Service (QoS)* - is required. Hence, to support mobile multimedia, both notions have to be combined. The result of this integration is henceforth termed *mobile QoS signaling*.

The article is organized as follows. In Section 2, the definition of the problem and related work are presented. Additionally, a new integration approach to implement mobile QoS signaling is proposed. Based on this approach, the required mobile QoS signaling components are investigated with regard to the distinction of control and data path in Section 3. While the solution for the control path is based on previous work, the solution for the data path is investigated from scratch. The required mobile QoS signaling components and their prototypical implementation are described as a proof-of-concept in Section 4. The article is wrapped up by a summary and an outlook to future work in Section 5.

2. PROBLEM DEFINITION AND RELATED WORK

2.1 Session Mobility

To enable session mobility, an application entity has to be addressable at all times. Therefore, every mobile end system has to receive an identification, which is addressable at any time. To provide connectivity to a mobile end system, two functions are necessary. First, the location information of the mobile end system has to be signaled and stored at a suitable node in the network. Second, the packet forwarding function of this node has be configured to redirect packets addressed to the mobile end system to the actual location. Every protocol, which is designed to support session mobility has to implement both functions. A good overview of the possible design options for these functions can be gained from the discussion of session mobility protocols in [1].

Mobile IP [2] has been standardized by the IETF to support session mobility. Here, the mobile end system's identifier is the *home IP address*. However, IP addresses are not only used to identify end systems but also contain the forwarding information of a packet. Therefore, the mobile end system's home IP address becomes topologically incorrect, when moving. To overcome this problem, Mobile IP applies an additional *readdressing* and *tunneling* mechanism.

The home IP address is bound to a topologically correct *foreign IP address* in order to implement the forwarding to a *foreign network*. The mapping and therefore the location information, is signaled by the end system after changing its point of network access to a suitable node of the *home network* - the *home agent* (version 4) and/or to the *correspondent node* (version 6). The signaling messages are called *registration* (version 4) or *binding update* (version 6). After receiving the new location information of the mobile end system, the home agent and/or the correspondent node establish the connection by configuring their forwarding behavior.

The access of a mobile end system to the Internet can be modeled as a three-tier architecture called *RAN architecture*, as depicted in Figure 1. A *mobile node (MN)* is connected to an *access router (AR)*, potentially through a wireless base station (not shown in the figure). Multiple base stations form the access router's radio region, which are depicted as separate, overlapping ellipses.[2] This tier of the architecture is generally termed *last hop*.

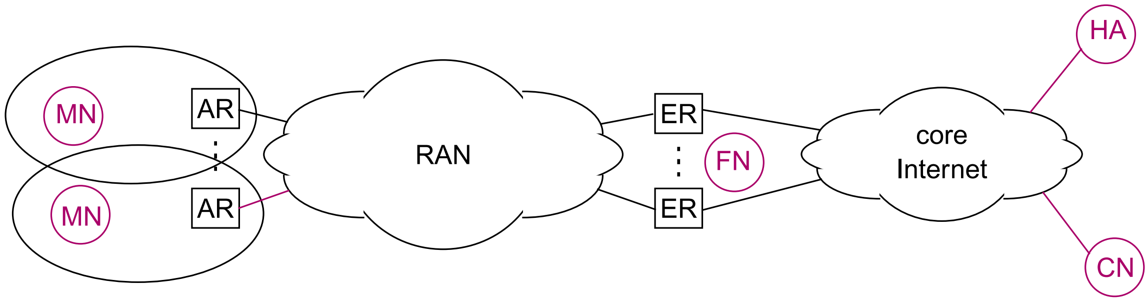


Figure 1: RAN Architecture

The next tier of the RAN architecture is an IP-based RAN, which can also be termed *administrative domain* according to IETF terminology. The access service is implemented by a multitude of access routers (where only two are explicitly shown here). The transition to the third tier of the architecture, the *core Internet*, is performed by *edge routers (ER)*.

Within the RAN architecture, the movement of a mobile node from one access router's radio region to another, is equivalent to an *IP subnet change*. Two terms have emerged: *micro-mobility* for mobility within administrative domains and *macro-mobility* for mobility between administrative domains. Both scopes shall be covered by the IETF mobility architecture of Mobile IP. However, the ability of Mobile IP to support micro-mobility has been studied and doubted in the last years [3]. If, for example, the mobile node is moving from one access router's radio region to another, a new connection has to be established to the home agent (HA) and/or to the correspondent node (CN). This might be a relatively large signaling overhead for

---

[2] Mobility and session mobility within the radio regions are supported by appropriate link layer mechanisms, like handover, which are out of scope for this work.

registration respectively binding updates compared to the actual movement of a mobile node. The farther the distance of a home agent respectively a correspondent node the worse this situations becomes. Both the load for the core Internet and the delay of subnet changes are increased by this mobile signaling.

To support these *intra-administrative* subnet changes, a mobility architecture is proposed, which utilizes a micro-mobility approach. Here, Mobile IP remains as macro-mobility solution for *inter-administrative* subnet changes. Micro-mobility approaches can be divided into two groups. On the one hand, modifications of Mobile IP are proposed, like Hierarchical Mobile IP [4]. On the other hand, proposals of independent solutions exist, where Mobile IP is terminated at the edge router of an access network, for example HAWAII [5] and Cellular IP [6]. Similar, to the latter ones, our work is focused on the micro-mobility part of the overall problem. Thus, as depicted in Figure 1, the whole RAN acts as a *foreign network (FN)* and Mobile IP is terminated at the edge router.

2.2 Mobile QoS Signaling

Any kind of system performance guarantees can only be given, if demand does not exceed capacity. In terms of network QoS, it is necessary to ensure that incoming traffic does not exceed the forwarding capacity of the network resources that is utilized to transmit this traffic. All approaches to network QoS take care of this basic requirement, albeit with differing combinations of mechanisms. In the IntServ architecture [7], for example, demand is controlled by explicit admission control of signaled service requests and subsequent traffic regulation. The DiffServ architecture [8] does not rule out signaling and admission control and further, also encompasses the need for traffic regulation at network edges. Even a simple overprovisioned best-effort network relies on the fact that demand does not regularly exceed capacity, which is often ensured through appropriate safety margins in the capacity dimensioning of a network.

Static provisioning of resources, as envisioned for best-effort or DiffServ-based network design approaches becomes much harder in the presence of end system mobility, because in addition to estimating the amount of traffic, the movement of end systems has to be estimated, as well. Additionally, as discussed in Section 2.1, seamless connectivity requires some signaling interaction between an end system and some other system in the network, anyway. Our conclusion, therefore, is to integrate connectivity and QoS signaling, because this integration allows to offer QoS guarantees more efficiently by the use of admission control. At the same time, this approach avoids an increase of system complexity and potentially signaling delay by two independent signaling channels, which have to be operated in a cooperative fashion anyway, to offer seamless connectivity *and* QoS to mobile end systems.

2.3 Existing Integration Approaches

By the time the IETF decided to specify a session mobility solution, routing and forwarding mechanisms had been based on the implicit assumption of fixed end systems only. Instead of changing all these existing mobility-unaware routing and forwarding mechanisms, new mobility-aware routing and forwarding mechanisms, like redirection through the home agent, tunneling or binding update, have been added to the architecture. With regard to this design decision, the consequences of additional integration work have been implicitly accepted, since all the remaining mechanisms and in particular QoS mechanisms, like queueing, scheduling, classification and signaling, remain mobility-unaware. Whenever a mobile end system requests network service other than pure session mobility, the service is based on mobility-unaware mechanisms. Thus, for the implementation of mobile QoS signaling, two different paradigms - *mobility-unaware* and *mobility-aware* - have to be combined.

So far, there is no fundamental solution in sight. As it would take a strong effort, the Internet community is not willing to change all the mobility-unaware routing and forwarding mechanisms. Thus, required mobility-aware QoS mechanisms have to be added to the Internet as well. Many mobile QoS signaling approaches have been proposed to solve the problem (see for example [9], [10], [11], [12], [13], or [14]). In [15], we show that all of these apply a *monolithic integration method*. Thus, a certain QoS-enabling mechanism of fixed IP-based networks, like modified RSVP, is intertwined with a certain session mobility mechanism, like Mobile IP or micro-mobility approaches, respectively. However, this monolithic integration method results in solutions, where the mechanisms are tightly coupled with the strategies to use them. The integration has to be reiterated every time a new use of the mechanisms should be applied, which of course leads to a large amount of specialized solutions. Additionally, the realized extensions to support mobile QoS signaling cannot easily be reused.

For example, in [10] mobility independent service guarantees are obtained in an Integrated Services Network by establishing advance resource reservations towards all locations a mobile end system may visit during the lifetime of a connection. Since RSVP is not directly adequate to make such reservations, a new reservation protocol MRSVP is proposed. This protocol offers the functionality to establish advance reservations for a given *specification* of a set of locations. A mobile node is granted an *active* reservation for its current location, whereas *pas-*

sive reservation are established by so called *proxy agents* in the locations given by the specification. Although both basic mechanisms, advanced reservation and third party signaling, are included in this protocol, they are tightly integrated and focused on mobile communications. Such an approach does not allow the re-use of these mechanisms for different application scenarios. Compared to our work, this proposal implicitly identifies and approaches similar problems, but fails to clearly structure the problem space. Instead, the proposed solutions are tightly coupled and targeted to a specific scenario.

2.4 A new Integration Approach

For these structural problems, the authors are convinced that utilizing a monolithic integration method is not appropriate to solve the problem. Therefore, a new *modular integration approach* is proposed: the required *components of mobile QoS signaling* are identified and modular solutions are presented. Hence, the implementation of mobile QoS signaling does not result in a monolithic system but in a collection of *required mobile QoS signaling procedures* and strategies for their use and combination. Thus, the mechanisms and their possible utilization are strictly separated by our approach and existing mechanisms can be used and adapted, whenever possible. Additionally, we show, that the new implemented mechanisms are reusable for other application scenarios. We are not aware of any other work applying such a rigorously modular approach.

To identify the required mobile QoS signaling components, a distinction between control and data path is made. While *control path components* already have been investigated in [15], the concept of *data path components* is completely new. Thus, this article focuses on the investigation of the required mobility-aware data path components. As a result of this investigation the need for connection-oriented services is identified and a general, modular, and reusable solution is proposed. The connection-less IP network layer is enhanced by a lightweight and truly optional connection-oriented mobile network service, which offers the possibility to establish soft state unicast connections at the network layer. It is shown, that this service closes the gap of missing mobility-aware data path mechanisms and completes the design of mobile QoS signaling. With regard to the RAN architecture (see Section 2.1), locations of functionalists are determined. As a proof-of-concept, the identified mobile QoS signaling components are implemented as modular procedures and integrated in an existing RSVP implementation.

3. MOBILE QOS SIGNALING COMPONENTS

3.1 Mobility-aware Control Path

The identification of the required mobility-aware control path components and procedures is based on a simple mobility description model, which has been presented in [15]. Within this work, the following mobility-aware control path components and procedures respectively, have been derived.

QoS handling on the control path can be offered to every mobile end system in our context by supporting temporal and spatial indirection. Thus, mobile QoS signaling can be utilized on the control path, by applying *reservation in advance* for temporal indirection and *third-party service* for spatial indirection. In the following, both mechanisms are summarized by the term *mobile signaling* for simplification.

3.1.1 Reservation in Advance

A procedure for temporal indirection of signaling is necessary to achieve predictable and seamless QoS support. Here, the basic idea is to minimize the delay of the intended effect of a re-registration after changing a subnet, by offering the possibility to mobile end systems to carry out the connectivity- and QoS signaling before entering the subnet. Reservations in advance are established according to a given movement prediction. Obviously, the more precise the information of the movement prediction, the more efficient resources can be managed.[3] The details of the reservation in advance service are defined in [16]. As discussed in [16], reservation in advance is a universal QoS-enabling mechanism, which is useful for a multitude of scenarios. Thus, the use within a mobile scenario is just one possible application of the service. The implementation of the reservation in advance mechanism is presented in Section 4.2.1.

3.1.2 Third-party Service

In order to provide reservation in advance to mobile end systems it is necessary to implement spatial indirection as well. In general, a mobile end system, which is not connected to a certain subnet, is not able to use any network service, due to a missing topologically correct IP address. For this reason, a mobile end system would not be able to establish a reservation in advance. As a solution to this problem, spatial indirection by providing third-party services is proposed. In our context, all third-party services, which are required to signal connectivity or

---

[3] In the scope of this work, we assume the movement prediction information to be available.

QoS in advance, like the possibility to authenticate or register in advance by a mobile end system are supposed to be offered. Ideally, all necessary resource reservations are established in advance, so the changing of the subnet at a later point of time can be carried out with minimal delay and no uncertainty about resource availability. As a proof-of-concept, the function of IP address allocation in advance is implemented within the present work. The implementation of the third-party service is presented in Section 4.2.2.

3.2 Mobility-aware Data Path

In an analogous manner to the mobility-aware control path components, mobility-aware data path components have to be identified. Therefore, in Section 3.2.1, the problem is investigated with regard to the distinction of mobility-aware versus mobility-unaware mechanisms. As a result, it is shown, that no proper solution to support mobile QoS handling can be obtained. Hence in Section 3.2.2, a new point of view to the problem is proposed: instead of integrating session mobility in IP-based networks by adding new mobility-aware routing and forwarding mechanisms, a connection-oriented integration approach is investigated. Henceforth, the required mobility-aware data path components are derived in Section 3.2.3.

3.2.1 Integration of Mobility-aware Mechanisms in IP-based Networks

The IETF integration approach for session mobility in IP-based networks is based on the design principle of implementing *transparent mobility*: all nodes of the Internet, except for the mobility-aware agents and mobile end systems, remain mobility-unaware.

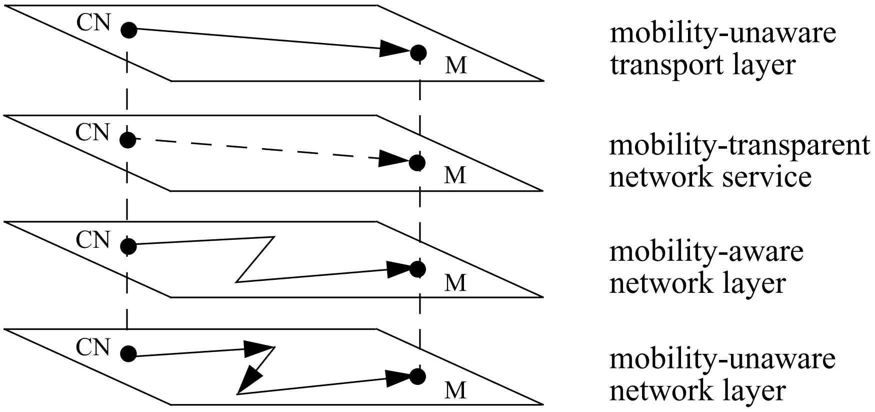


Figure 2: Mobility-transparent Network Service

As shown in Figure 2, the *mobility-aware network layer* offers transparent transmission of IP packets (direct dashed arrow) by implementing the according mobility-aware mechanisms (zigzag arrow), for which it uses *mobility-unaware network layer* mechanism (legs of zigzag arrow). Therefore, as the mobility-aware routing and forwarding information are transparent to the mobile end system (M) and the correspondent node (CN) on the data path, mobility is also transparent for every application above the *transport layer* (direct solid arrow). Hence conceptually, a *mobility-transparent network service* has been created, which hides the mobility-aware network layer from the transport layer.

Now, whenever considering a dedicated application on the transport layer the following problem arises: to identify IP packets, the necessary network header fields have to be examined. However, the required “real” respectively mobility-aware information is hidden by the mobility-transparent network service. As an example the classification as a mobility-unaware basic forwarding mechanism is considered. A class is assigned to an IP packet according to the values of dedicated IP header fields. In our example, all packets with the same source and destination address as well as the same protocol type are mapped to the same class. For this class an according forwarding behavior is defined. However, when supporting mobility in the network layer, a redirection on the forwarding path has to be performed. A corresponding node sends an IP packet to a mobile end system, which is addressed to the mobile end system’s home IP address. Whenever, as in our example, the mobile end system is connected to a foreign network with a new foreign IP address, a mobility-aware router in the home network takes care of this IP packet. The mobility-aware router performs a redirection and the IP packet is forwarded using an IP tunnel to the according foreign IP address. Thus, a packet, which is transmitted from a correspondent node to a mobile node, would change its class when entering the tunnel, due to the different IP header used for the redirection.Therefore, the IP packet would no longer be treated according to the original forwarding behavior.

When considering the combination of both mechanisms, two potential solutions exist: either the classification mechanism becomes mobility-aware or the mobility-aware mechanisms forwarding and tunneling perform transparent operations. In the first case, the classification mechanisms would need to look into the IP tunnel performing the operation on the mobile end system’s home address. In the second case, whenever a mobile end system moves to a new subnet, it has to renegotiate all the services based on the existing classification. However, both solutions would lead to poor QoS handling: while the first solution introduces high overhead to all routers in the Internet implementing mobility-aware classification mechanisms, the second solution does not offer seamless and thus no ubiquitous services to mobile end systems. Hence, both combinations do not offer a satisfying solution.

> The problem of integrating the mobility-aware with the mobility-unaware paradigm is inherent to the integration method of session mobility into IP-based networks. Therefore, to solve the problem it seems inevitable to investigate a new integration approach for session mobility.

3.2.2 Integration of Connection-oriented Mechanisms in IP-based Networks

Instead of considering the mobility-aware aspects of new routing and forwarding mechanisms, the connection-oriented aspects of all the session mobility supporting approaches are considered. In general, packet transmission at the network layer can be classified as *connection-oriented* versus *connection-less*. Connection-oriented transmission is divided into three phases - establishment of a connection, data transmission phase and tear down of a connection - while connection-less transmission only consists of the data transmission phase. That means, for connection-oriented transmission, forwarding information is stored within the forwarding tables of the network nodes at the connection establishment phase and is deleted at the connection tear down phase. In contrast, packets transmitted by the connection-less approach have to carry all the required forwarding information.

The network layer of the Internet has been designed and implemented as a connection-less network layer. Hence, the model is particularly based on the separation of control and data path, and on connection-less forwarding. According to the model, every connection-oriented semantic on the data path is placed within the transport layer. The *IP forwarding* can be classified as connection-less. The determination of paths for a certain point of time or in dependency of a certain metric respectively, is performed separately by *IP routing*. Due to this separation, IP packets belonging to the same sender and receiver class may possibly traverse totally disjoint paths. IP's particular robustness results from that.

With regard to the new requirements for the Internet, like session mobility [2], traffic engineering [17], or constraint-based routing [18], connection-oriented semantics on the data path and the necessary signaling technology for the control path have been added to the IP layer, recently. Here, the connection-less "character" of the IP layer has not been questioned and the connection-oriented service has been bound to particular applications. Therefore, the integration of session mobility into IP-based networks has not been supported by applying connection-oriented services, but by identifying and adding the required mobility-aware mechanisms.

> IP's principle of connection-less forwarding has to be violated to achieve session mobility. Obviously, Mobile IP - as well as every other session mobility protocol - adds connection-oriented semantics to the network layer.

A connection is established by signaling the location information and configuring the forwarding behavior. After establishing a connection, data packets are transmitted. As Mobile IP is a *soft state* protocol, the connection's duration is limited by a lifetime parameter and does not need to be torn down explicitly after moving to a new access point. The signaling of a network access change respectively the connection establishment takes place between the mobile end system and the home agent or end-to-end with the correspondent node. As a matter of fact, this is the only way to provide network connectivity to mobile end systems. Therefore, the following form of signaling is termed *connectivity signaling* of a mobile end system. Further on, the forwarding resulting from the signaling process is called a *connection*.

3.2.3 Identification of the Required Components

To summarize, IP-based networks always have to be enhanced by a connection-oriented paradigm, whenever session mobility shall be supported. Additionally, assuming an increasing number of mobile end systems connected to the Internet, it seems reasonable not only to add mobility-aware mechanisms, like [19] or [20], but to extend the connection-less IP by an optional, but generic *connection-oriented mobile network service*. This service offers connectivity signaling to mobile end systems to establish a *soft state unicast connection*. By applying the service within the RAN architecture (see Section 2.1), the existing connection-oriented paradigm and the connectivity signaling is revised in an explicit manner. With regard to this design decision, the mobile signaling procedures can be reused to provide connectivity signaling for mobile end systems.

3.3 Connection-oriented Mobile Network Service

To provide the connection-oriented mobile network service, the IP layer has to be extended to offer the possibility for mobile end systems to establish and maintain a soft state unicast connection. As already mentioned, instead of implementing completely new mechanisms, existing mechanisms for the service design are investigated and whenever possible, reused and adapted. An important design assumption is that IP remains the basic network service. Particularly, the solution is supposed to be modular, lightweight, and should not introduce disadvantages to the existing architecture.

3.3.1 Service Design

Multiprotocol Label Switching (MPLS) [21] is a connection-oriented network protocol and therefore - in terms of the *OSI layer model* [22] - equivalent to the connection-less IP. Additionally, MPLS follows the separation of data path and control path. Thus, it is possible to isolate the *connection-oriented forwarding* mechanism of the MPLS data path and to influence the connection-oriented network service characteristic by the selection of an according label distribution protocol for the control path. The MPLS architecture provides a choice of possible protocols but no assumption on a dedicated protocol is made. The *Resource ReSerVation Protocol (RSVP)* [23] can be utilized as an MPLS label distribution protocol (RSVP-TE [24]). Our solution combines the connection-oriented forwarding mechanisms of MPLS with the *soft state paradigm* of RSVP to extend the IP layer by the semantic of a soft state unicast connection. While this insight is rather simple in nature, it nevertheless denotes a truly new combination of paradigms between the usual alternatives of hard-state, connection-oriented and stateless, connection-less unicast communication architectures at the IP layer. A similar combination of paradigms can be found in IP multicast. Additionally, it is important to note that the service design is only exemplified on the basis of MPLS for connection-oriented forwarding and RSVP for soft state paradigm.

3.3.2 Mobile QoS Signaling Procedures

The mobile signaling procedures, which are based on RSVP, can be reused to provide connectivity signaling for mobile end systems by integrating RSVP-TE mechanisms for *label switching*. The implementation of mobile signaling procedures are based on experimental *lightweight signaling* extensions of RSVP. Therefore, the required procedures to implement mobile QoS signaling on the control path *and* on the data path are:

- lightweight signaling: *remote client* and *one-pass* (see Section 4.1)
- mobile signaling: reservation in advance and third-party service (see Section 4.2)
- label switching: *label distribution* and *explicit routing* (see Section 4.3)

The mechanisms are implemented as modular extension of our existing RSVP implementation [25] and it is shown that these procedures offer generic functionality, which can be used within a wide variety of scenarios. The integration of the mechanisms results in the possibility for mobile end systems to request QoS signaling and connectivity signaling (respectively session mobility) as an integrated service or separately. Thus, it offers soft state, mobile, lightweight and aggregated[4] signaling, which is summarized as *generic signaling* for simplification.

3.3.3 Routing

Besides these major results, two additional effects can be derived from this service design. As RSVP implements a strict separation of signaling and routing mechanisms, it is designed to work with different routing algorithms. Thus, the characteristic of the connection-oriented network service can also be influenced by the choice of a routing protocol. As an example, basic RSVP is not able to influence a path selection according to the requested QoS. However, the combination of RSVP with QoS routing algorithms offers this possibility. The combination of RSVP and multipath routing addresses an additional aspect, which affects resource management as well as traffic engineering aspects. Hence, multipath routing is able to spread the traffic over several paths. Both combinations might be interesting for future research work. However, within this work a traditional intra-domain routing protocol, like RIP or OSPF, is assumed.

3.3.4 Modular Extension of the IP Layer

In conclusion, the IP layer can be modularly extended by the connection-oriented forwarding, generic signaling and the appropriate routing functionality to offer connection-oriented mobile network service. In Figure 3, the modular design of the extended IP layer is depicted.

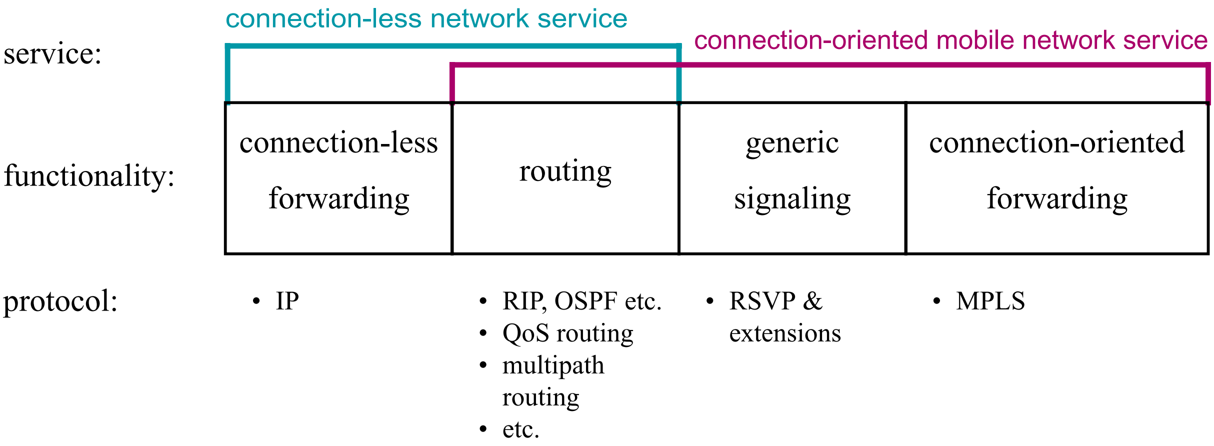


Figure 3: Service Design

Of course, the combination of connection-less and connection-oriented network service is not a fundamentally new idea. Conceptually, the design is based on the OSI layer model, where

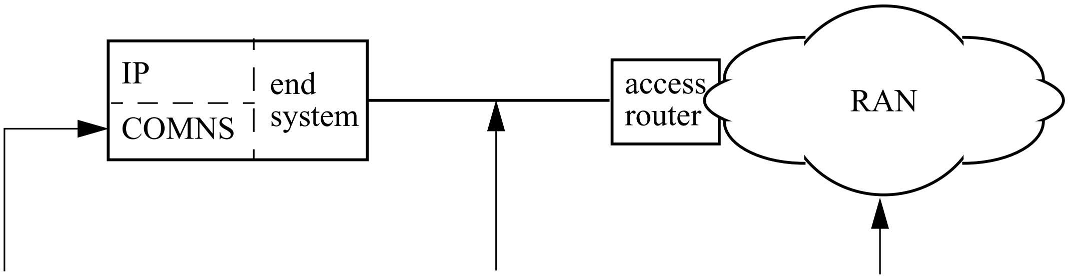
---

[4] Aggregation can be realized by MPLS label stacking, which is out of scope for this work.

it is possible to distinguish between the connection-oriented network service and the connection-less network service on the network layer. These services are offered to a higher layer by a network service access point. In order to extend the IP layer of the connection-oriented mobile network service as lightweight as possible, it is necessary to derive a minimal set of these network service access points within the architecture.

3.3.5 Service Architecture

With regard to the three-tier radio access network architecture (see Section 2.1) as well as the definition of mobile QoS signaling (see Section 2.2), the connection-oriented mobile network service (COMNS) is implemented as an add-on to IP on the end system (see Figure 4). It is important to note, that the connection-oriented mobile network service does not necessarily have to be provided on any other node within the radio access network or within the core Internet. The realization of the service request means that MPLS respectively RSVP have to be implemented on the end system. Here, we assume, a lightweight version of RSVP is not prohibitively expensive for end systems.



Service request:
- connectivity signaling
- QoS signaling
- mobile QoS signaling

Service interface:
- lightweight signaling
- mobile signaling
- label switching

Service provisioning:
- MPLS-based?
- IP-based?
- QoS provisioning?

Figure 4: Service Architecture

For service provisioning, the access routes have to implement a hybrid forwarding architecture, since both packet forwarding and label switching have to be offered to the end systems. The detailed service architecture depends on the architecture of the RAN.

In the preferential case of an MPLS-based RAN, an IP overlay structure is used to transmit regular IP traffic between edge nodes of the RAN (access and edge routers). In fact, this currently seems to be the preferred configuration of many provider subnets. Then, the functionality of access routers to offer the connection-oriented mobile network service basically consists of exposing the MPLS service to end systems. Additionally, it must be possible (and sustainable) to dynamically establish QoS-enabled label switched paths for application flows initiated by end systems. Whether QoS assurances are guaranteed within the RAN based on per-node mechanisms or network-wide resource provisioning is beyond the scope of our work. Any standard IP traffic can be transmitted between the end system and the access router using a *default label* (in contrast to *application flow label*) and is then handled by regular IP processing.

In case of an IP-based RAN, we assume the existence of another micro-mobility solution and the functionality of the access router is to map the traffic from end systems appropriately to the connection semantics of the micro-mobility solution.

In either case, the architecture and design of the RAN can profit from the availability of integrated connectivity and QoS signaling, including the extensions for spatial and temporal indirection, at the edge of the RAN. As discussed earlier, this allows to improve both handover characteristics and QoS provisioning.

4. MOBILE QOS SIGNALING PROCEDURES

In our prototype implementation, all previously described functional components are implemented as extensions to the KOM RSVP engine [25].

4.1 Signaling Extensions

Two experimental extensions to RSVP are used to realize the mobile signaling approach. These extensions are briefly presented here. For a detailed description and experimental investigation of the extensions, the reader is referred to [26].

4.1.1 Remote Client

RSVP defines two alternative methods to transmit messages between RSVP-capable nodes. RSVP messages are either transmitted as raw IP packets or using UDP encapsulation. When using UDP encapsulation, packets are addressed to well-defined ports. If multiple clients run on a single end system, this addressing scheme requires a central manager entity (usually the RSVP daemon) to receive and dispatch incoming messages. For mobile end systems, the effort of running a dedicated RSVP daemon might be prohibitively expensive, even if this daemon does not need the full functionality. An elegant solution is to define additional protocol mechanisms which allow an RSVP daemon running on the first RSVP-capable hop to administer and communicate remotely with a number of clients. These clients in turn only need to implement RSVP stubs and except for the special addressing scheme, participate in the full RSVP signaling procedure.

The only protocol extension needed to realize this scheme is a new message type, termed `InitAPI`, which is used both to register a client and its addressing information at the first hop RSVP daemon, as well as for communicating registration information from the remote RSVP daemon to the client. The RSVP daemon uses a dedicated object, termed `API_Server`, to process and respond to these messages and to administer the clients.

4.1.2 One-pass Signaling

In its basic form, RSVP uses a bidirectional message exchange to set up an end-to-end simplex reservation. This procedure is called *one-pass with advertising* (OPWA) and is used to support heterogeneous requests from multiple receivers within a multicast group. Reservations are eventually requested and established from the receiver to the sender. The advertising phase is needed to route reservation requests along the reverse data path to the sender. Furthermore, service-related data are collected during the advertisement phase and delivered to the receiver. As discussed in [26], there are a number of scenarios in which both features are not needed. In such cases, the original OPWA procedure represents an unnecessary signaling overhead for both end systems and intermediate nodes. Additionally, there might be situations where an initial (potentially duplex) reservation establishment by the initiator is desirable as fast as possible, which can later optionally be overridden by appropriate signaling requests from the responder and in turn from the initiator. The *one-pass* service establishment mechanism is used in this fashion to optimize the signaling interaction for mobile end systems. It fully interacts with traditional RSVP signaling, such that it is possible to optionally override an initial one-pass reservation with later requests.

A new message type, `PathResv`, is defined to indicate that reservations based on the transmitted `TSpec` shall be established through the transmission of this message. Other than the message type, the syntax is exactly the same as for a `Path` message. In order to request a duplex reservation, an optional `DUPLEX` object can be added to a `PathResv` message. The `DUPLEX` object carries the reverse port information, assuming that the same transport protocol is used in both directions. Again, this specification can easily be changed or extended, if necessary for any purpose. The duplex extension is only sensible, when symmetric paths can be assumed between two end systems and furthermore, only for unicast communication. These assumptions are likely to hold true for mobile services in the controlled environment of a RAN.

4.2 Mobile Signaling

The implementation of the temporal and spatial indirection mechanisms - reservation in advance (see Section 3.1.1) and third-party service (see Section 3.1.2) - is based on the signaling extensions presented above, remote client and one-pass signaling.

4.2.1 Reservation in Advance

We have extended RSVP to handle advance reservation requests by specifying a new object and appropriately extending the traffic control module co-located with an RSVP daemon. The principles and details of the advance reservation service are described in [16] and [15]. The basic mechanism is realized by representing the timing information (start and duration) of advance requests as a `RERA` object and inserting it into `RESV` messages. Path-related state has to be maintained during the hold-back time of an advance request, anyway, and it does not govern the assignment of transmission resources to traffic flows. Therefore, it is sufficient to associate the specification and handling of advance requests with the reservation part of the traditional RSVP message exchange. Compared to other advance reservation proposals, the service specification presented in [16] does not restrict the potential parameter space for advance reservations by separating the service period into preemptable and non-preemptable service. However, as discussed in [16], all internal state information at the traffic control module can be represented with essentially the same data structures as proposed by earlier work. In particular, our implementation uses the data structure proposed in [27] to maintain state information for the non-preemtable parts of all service requests.

In the context of mobile end systems requesting transmission service from a RAN, the reservation in advance mechanism is used in combination with one-pass signaling to efficiently establish a (potentially duplex) QoS transmission path in advance. This keeps the necessary signaling overhead for all participating entities at a minimum.

4.2.2 Third-party Service

The implementation of the third-party service is based on the remote client extension described in Section 4.1.1. This mechanism is appropriately extended to also carry out the information exchange for third-party service. The basic service that is implemented in the prototype is that of remotely pre-allocating IP addresses and redirecting service requests. When registering with a remote RSVP daemon, a mobile end system inserts a `THIRD_PARTY` object into the `InitAPI` message to describe the third-party request. In our prototype, the `API_Server` object then requests the allocation of an IP address from a DHCP server and reports it back to the client. This address can then be used to relay further RSVP messages that are emitted by the mobile end system through the remote client association with the daemon. Thereby, the end system can interact with the rest of the network and particularly with a home agent or correspondent node. For example, it can establish advance service requests by using a topologically correct IP address, while not yet being connected to that particular subnet.

4.3 Label Switching

As discussed previously, the connection-oriented mobile network service is particularly useful in combination with RAN architectures that are based on IP over MPLS. Therefore, the prototype implementation also encompasses the MPLS-related functional components, which are implemented according to the relevant standardization document [24]. From the set of functions specified in [24], we have implemented RSVP-based label distribution and explicit routing. The signaling procedures are integrated with appropriate Linux-based MPLS forwarding implementations (see [28] and [29]). Thereby, it is possible to set up explicitly routed paths through a RAN for standard IP routing of MPLS, and further, it is possible to emulate the setup of QoS-enabled dedicated paths for specific service requests. We have not yet integrated the configuration of MPLS forwarding and packet scheduling on Linux-based routers, because the available scheduling disciplines are rather limited, compared to, e.g. ALTQ for BSD-based systems [30]. Unfortunately, there is no open and up-to-date MPLS forwarding implementation available for BSD-based systems at this time.

4.4 Integration

In Figure 5, the integration of the described signaling components is depicted according to the RAN architecture (see Section 2.1). As described in Section 3.3.5, the integration is very simple because only the generic signaling interface between end system and access router has to be implemented. The `API_Client` is implemented on the end system, which is enabled to send an `InitAPI` message to a remote access router, hosting the `API_Server` and the RSVP daemon. As described, the `InitAPI` is extended for generic signaling (see Figure 5). The third-party request, potentially with regard to the reservation in advance request, is responded by the API Server by calling a third-party service. All the other primitives are handed over to the RSVP daemon, which stores it and triggers the according actions, like advance admission control (see Section 4.2.1). The mapping of the mobile QoS signaling mechanisms to the RAN mechanisms has to be done by the access router (see Section 3.3.5). As an example, the mapping of the defined labels, default label (connection-less IP forwarding) and application flow label (connection-oriented MPLS forwarding) are depicted.

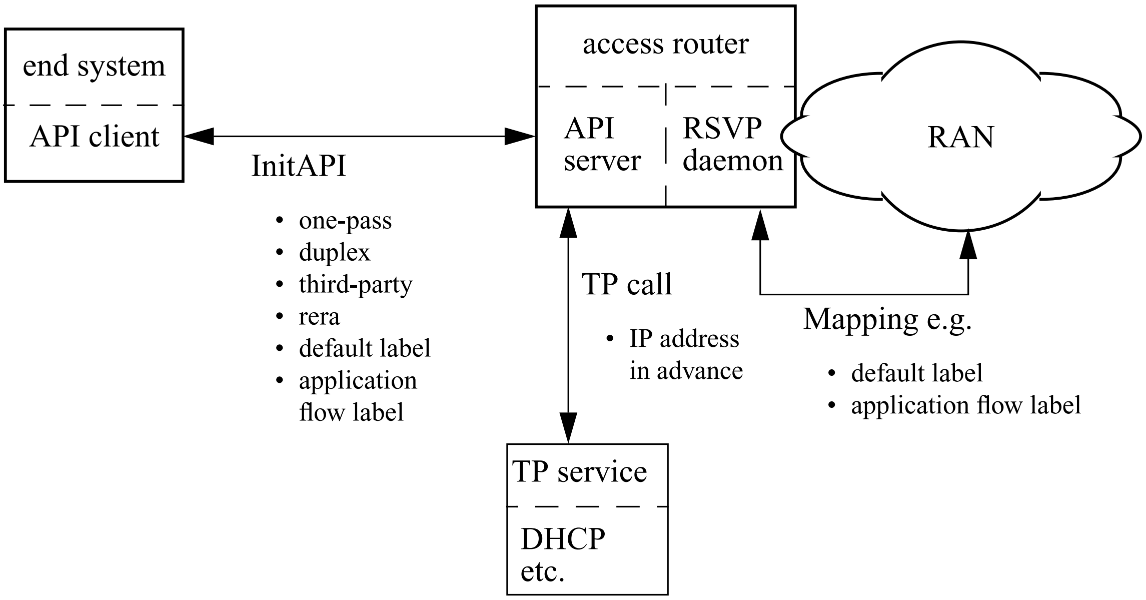


Figure 5: Service Implementation

5. SUMMARY AND FUTURE WORK

In this paper, a modular approach to integrate session mobility and QoS for mobile end systems is investigated. The approach is based on a connection-oriented mobile network service, which offers the possibility for mobile end systems to establish soft state unicast connections within a RAN architecture. Thus, the required connection-oriented paradigm to support session mobility is applied in an explicit manner. It is shown that the service can be used to support different applications like connectivity signaling of mobile end systems as well as mobile QoS signaling. Basic mechanisms are strictly decoupled from the rest of the overall system architecture. Thereby, existing mechanisms can be reused. The presented implementation results in a collection of mobile QoS signaling procedures, which are shown to be applicable for a wide variety of scenarios.

Future work must be directed to proof the general capability of the connection-oriented mobile network service within the RAN. Thus, corresponding applications have to be investigated. We have decided to implement three possibilities: first, as a proof-of-concept, a micro-mobility solution is designed and implemented, where mobile end systems use the connection-oriented mobile network service within the RAN to request network connectivity. Second, the connectivity signaling is combined with QoS requests to measure the handover delay in a test environment. Finally, a mobility-aware load balancing solution is designed and implemented. With regard to the temporal and spatial indirection functionality of mobile QoS signaling, mobility-aware load information is available at the access router of the RAN. This information can be used for load balancing at the edge of the RAN.

6. REFERENCES

[1] C. E. Perkins. *Mobile IP Design Principles and Practices*. Prentice Hall, January 1998.

[2] C. E. Perkins. RFC 2002 - IP Mobility Support. Standards Track RFC, October 1996.

[3] J. Solomon. RFC2005 - Applicability Statement for IP Mobility Support. Standards Track RFC, October 1996.

[4] E. Gustafsson, A. Jonsson, and C. E. Perkins. Mobile IP Regional Registration. Internet Draft (draft-ietf-mobileip-reg-tunnel-06.tx), March 2002. Work in Progress.

[5] R. Ramjee, T. L. Porta, S. Thuel, K. Varadhan, and S. Y. Wang. HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks. In *Proceedings of the Seventh Annual International Conference on Network Protocols (ICNP'98).* October 1999.

[6] A. T. Campbell, J. Gomez, and A. G. Valko. An Overview of Cellular IP. In *IEEE Wireless Communications and Networking Conference (WCNC'99).* September 1999.

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

[8] T. Li and Y. Rekhter. RFC 2430 - A Provider Architecture for Differentiated Services and Traffic Engineering (PASTE). Informational RFC, October 1998.

[9] S. Singh. Quality of Service Guarantees in Mobile Computing. *JSAC - Computer Communications*, 19(4), 1996.

[10] A. K. Talukdar, B. R. Badrinath, and A. Achary. MRSVP: A Resource Reservation Protocol for an Integrated Services Network with Mobile Hosts. In *Proceedings of ACTS Mobile Summit'98*. June 1998.

[11] I. Mahadevan and K. M. Sivalingam. An Architecture for QoS Guarantees and Routing in Wireless/Mobile Networks. In *First ACM International Workshop on Wireless Mobile Multimedia (WOWMOM'98).* October 1998.

[12] A. Mahmoodian and G. Haring. Mobile RSVP: Towards Quality of Service (QoS) Guarantees in a Nomadic Internet-based Environment. In *IFIP TC 6 Fifth International Conference on BROADBAND COMMUNICATIONS '99, Convergence of Network Technologies*. October 1999.

[13] J. W. M. Yu Zeng and Y. Shen. Indirect RSVP for Virtual Cluster Cellular Mobile IP Networks. In *Proceedings of the Networking 2000 Conference.* May 2000.

[14] A. Terzis, M. Srivastava, and L. Zhang. A Simple QoS Signaling Protocol for Mobile Hosts in the Integrated Service Internet. In *Proceedings IEEE INFOCOM'99.* March 1999.

[15] N. Berier, M. Karsten, J. Schmitt, and R. Steinmetz. Mechanisms for QoS Signaling in a Mobile Internet Environment. In *Proceedings of Mobile Multimedia Communications (MoMuC'2000).* October 2000.

[16] M. Karsten, N. Berier, L. Wolf, and R. Steinmetz. A Policy-Based Service Specification

for Resource Reservation in Advance. In *Proceedings of the International Conference on Computer Communications (ICCC'99).* September 1999.

[17] D. O. Awduche, A. Chiu, A. Elwalid, I. Widjaja, and X. Xiao. Overview and Principles of Internet Traffic Engineering. Internet Draft (draft-ietf-tewg-principles-02.txt), May 2002. Work in Progress.

[18] B. Jamoussi, L. Andersson, R. Callon, R. Dantu, L. Wu, P. Doolan, T. Worster, N. Feldman, A. Fredette, M. K. Girish, E. Gray, J. Heinanen, T. E. Kilty, and A. G. Malis. RFC 3212 - Constraint-Based LSP Setup using LDP. Standards Track RFC, January 2002.

[19] A. Terzis, J. Krawczyk, J. Wroclawski, and L. Zhang. RFC 2746 - RSVP Operation Over IP Tunnels. Standards Track RFC, January 2000.

[20] D. L. Black. RFC 2983 - Differentiated Services and Tunnels. Standards Track RFC, October 2000.

[21] E. C. Rosen, A. Viswanathan, and R. Callon. RFC 3031 - Multiprotocol Label Switching Architecture. Standards Track RFC, January 2001.

[22] M. T. Rose. *The Open Book: A Practical Perspective on OSI.* Prentice Hall, 1989.

[23] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin. RFC 2205 - Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification. Standards Track RFC, September 1997.

[24] D. Awduche, L. Berger, D.-H. Gan, T. Li, V. Srinivasan, and G. Swallow. RFC 3209 - RSVP-TE: Extensions to RSVP for LSP Tunnels. Standards Track RFC, December 2001.

[25] M. Karsten, J. Schmitt, and R. Steinmetz. Implementation and Evaluation of the KOM RSVP Engine. In *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2001).* April 2001.

[26] M. Karsten. Experimental Extensions to RSVP - Remote Client and One-Pass Signalling. In *Proceedings of the Ninth IEEE/IFIP International Workshop on Quality of Service (IWQoS'01),* June 2001.

[27] O. Schelén and S. Pink. An Agent-based Architecture for Advance Reservations. In *Proceedings of 22nd IEEE Conference on Computer Networks (LCN'97).* November 1997.

[28] K. Fraser and P. Quiney. MPLS on Linux, September 2000.
Software at http://www.cl.cam.ac.uk/Research/SRG/netos/netx/index.html.

[29] J. R. Leu. MPLS for Linux, November 2001.
Software at http://sourceforge.net/projects/mpls-linux.

[30] K. Cho. *The Design and Implementation of the ALTQ Traffic Management System.* PhD thesis, Keio University, Japan, January 2001.
Software at http://www.csl.sony.co.jp/person/kjc/programs.html.