

Design of Quality-of-Service Signaling for IP-based Mobile Networks

Joachim Hillebrand and Christian Prehofer
DoCoMo Communications Laboratories Europe
Munich, Germany
Email: {hillebrand,prehofer}@docomolab-euro.com

Roland Bless and Martina Zitterbart
Institute of Telematics
University of Karlsruhe, Germany
Email: {bless,zit}@tm.uka.de

Abstract. Efficient support for multimedia communication is essential for next generation mobile wireless networks. In order to provide the necessary Quality of Service (QoS) in mobile wireless networks, appropriate signaling has to be defined. We discuss several approaches and design issues of mobility-aware QoS signaling including on/off-path signaling and receiver-initiated reservations. Our QoS signaling architecture integrates resource management with mobility management. It is based on a Domain Resource Manager concept and nicely supports various different handover types. In particular, our approach supports anticipated handover with pre-reservation of resources before the mobile node is attached to the new access point.

1 Introduction

In future IP-based mobile networks, users expect not only access to the widely deployed and well-known, best-effort Internet services, but also to enhanced multimedia communications, which often have real-time requirements for data delivery. Therefore, Quality of Service (QoS) must be provided for data flows and maintained even when a mobile node (MN) changes its current point of attachment to the network. On the network level there are several techniques such as Differentiated Services [1] or Integrated Services [2] to enforce QoS during packet forwarding. Additionally, a *resource management* has to take care of admission control, allotment and release of requested resources for QoS provisioning within a network.

The main requirement for mobile networks is that an existing resource reservation has to be taken over to a new access point in case of a handover of the mobile node. In this case, the path of the reservation has to be adapted. The handover for the data flow itself is accomplished by a *mobility management* protocol like Mobile IP [3]. Its main task is to forward packets to their destinations although the access point and the IP address of a node changes. Our approach establishes a coupling between mobility management and resource management in order to provide the required resources along the current data path. As there are different approaches to (micro-)mobility management, we use a minimal, generic interface to mobility management.

In addition, the QoS signaling protocol must support a variety of different handover types, e.g., *hard* or *soft* [4] handovers as well as *anticipated handovers*, which reserve resources along the new path in advance and impose new requirements on signaling. Combinations of these handover types have to be considered as well, because an MN may be forced to continue an already started antic-

ipated handover process with a hard handover due to a sudden loss of connectivity. An enhanced form of anticipated handover uses mobility prediction (see for example [5]) to anticipate the next access point to be used by a mobile node.

This paper discusses the main approaches to QoS signaling and addresses several design issues, e.g., receiver-initiated reservations, with the focus on mobile networks. Then, we present our approach to mobility-aware QoS signaling which integrates resource management with mobility management and location management to provide QoS signaling for a wide variety of handover types. While most other approaches consider the handover cases individually, an integrated solution is presented that uses a single model to describe all possible handover cases. A main advantage is that this model allows one to switch between these handover cases during a handover process.

2 A Mobility-Aware QoS Signaling Architecture

In the following, we give an overview of MARSP (Mobility-Aware Reservation Signaling Protocol) and its corresponding signaling architecture that is shown in Figure 1. They provide QoS in a mobile IP-based network with the following features:

- Independence of a particular QoS technique for provisioning of QoS-based services, at IP layer as well as at link layer. The architecture does not depend on a specific QoS model, but works with various QoS solutions such as Integrated Services and Differentiated Services.
- Independence of specific radio access technologies. It can be expected that different *radio access networks* (RANs) are used at different locations. The presented architecture does not depend on a specific radio access technology. However, it supports the utilization of RAN characteristics.
- Interworking with different micromobility concepts. The solution allows one to integrate different micromobility approaches as a base for seamless handovers. This integration is possible due to a very loose coupling between mobility management and resource management at the MN. As described below, coupling is merely achieved by triggering mobility management signaling depending on successful reservation procedures.
- Support for inter-domain handovers, i.e. if a MN changes its point of attachment to a network that is administered by another organization.

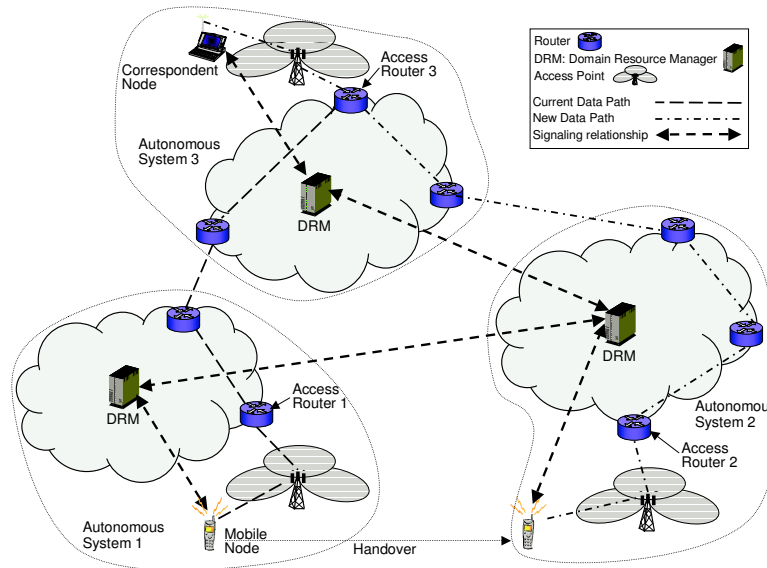


Figure 1: Components of a domain providing wireless access

These features provide QoS support for a wide variety of handover cases that differ with respect to different cellular environments, mobile node capabilities and connectivity, network configurations and operator needs.

In IP-based mobile networks, an *Access Router (AR)* provides IP connectivity to a MN within a domain as illustrated in Fig. 1. An AR can be connected to multiple *Access Points (APs)*. These APs provide link layer connectivity by radio transceivers and are part of the *RAN*. Moreover, each domain has a *Domain Resource Manager (DRM)* that controls all resources at IP level within this domain. It can be viewed as a dedicated logical entity for management purposes. An actual implementation, however, may use several distributed entities.

Due to the fact that the resource management is not necessarily directly located on the data path some basic functions have to be provided. For instance, the DRM must monitor routing protocols (e.g., OSPF and BGP) in order to perform admission control for the current data paths. DRMs must locate the next DRM along the path in order to forward signaling messages. Additionally, a DRM must configure routers by using a management interface, e.g., installation of traffic profiles for differentiated services by using SNMP, COPS or command line interfaces.

The architecture separates resource management signaling from mobility management signaling, because it aims at independence of the different technologies. Resource management can be seen as an additional function to improve traditional best-effort data transport. Generally, more flexibility is achieved if signaling procedures are decoupled from each other. For instance, if a MN requests resources along the new path before it finally registers at the new AR, data packets traversing the new path will directly receive the corresponding QoS during forwarding.

Within the mobility-aware QoS signaling architecture several logical interfaces can be identified. Basically, it can be distinguished between signaling for resource management, signaling for mobility management and signaling at application level. Direct signaling between the ap-

plications may be required to let applications explicitly adapt the content of their data flows to the current resource availability. The latter may change due to a handover, especially if a change of wireless access technology is involved (a so-called *vertical handover*). Thus, applications should be notified via an internal interface of QoS changes and they should signal the sender at application level, which can then adapt its sending rate.

In addition to explicit signaling for resource reservation, location management or some of the proposed protocols for seamless mobility can be integrated into the architecture as well. Location management information may be available at a DRM because of domain-specific conditions such as a network along a highway or railroad track. In this case, movement prediction can be used by a DRM to reserve resources in advance without the need to let the MN request resources explicitly.

3 Design Issues for QoS Control

While there are many papers on QoS architectures, some of the essential design issues have not been discussed, especially for mobile systems. In this section, we discuss on-path vs. off-path signaling and the issues of receiver initiation of reservations, which is difficult for asymmetric paths. An overview over some approaches is given in [6], with a nice classification of the interaction between resource and mobility management.

3.1 On-path vs. off-path Signaling

In the following, we compare the two main QoS approaches for their applicability in IP-based mobile networks. The two approaches are on-path or hop-by-hop signaling and centralized resource management in the form of a domain resource manager, that is based on a bandwidth broker concept.

The on-path reservation approach is characterized as follows:

- QoS resources are managed locally by each router.

- Signaling is triggered by the terminals and follows the data path.
- End-to-end reservations are set up hop-by-hop by a signaling protocol that installs states in routers.

The RSVP protocol [7] is the current the Internet standard for on-path QoS signaling and is used for other signaling purposes as well. RSVP is not mobility-aware and for instance does not support changing of IP addresses. An overview of current research on RSVP extension can be found in [8]. In the same direction, the IETF currently works on a new resource signaling protocol in the NSIS working group [9]. Several European research projects use centralized resource management entities, e.g., [10, 11, 12]. The centralized QoS architecture is characterized as follows:

- A Domain Resource Manager (DRM) handles the resources for one domain.
- The DRM maintains an up-to-date image of resources and reservations in its domain.
- The DRM requests resources from DRMs in adjacent domains along the data path in order to provide end-to-end reservations.

There are several reasons why we decided for the option of a central resource management:

- It is flexible with respect to QoS mechanisms on the data path, e.g., DiffServ or IntServ, or overprovisioning. Furthermore, complex management functions can evolve independently from router implementations.
- It allows one to integrate resource and location management as well as policy and accounting aspects.
- It is more flexible with respect to resource management optimization, especially for the wireless links.
- It is more suitable for anticipated handover, especially with pre-reservation of resources.

The last item requires some explanation. Although pre-reservation is also possible using on-path signaling, it is considerably more difficult to initiate this from the network side. Existing approaches (for instance [8]) for on-path signaling do not use network-assisted pre-reservation. With the DRM approach, a DRM can determine the route and reserve resources for a new access point within its domain or by contacting a neighboring DRM. If the resource management is done locally by each router, it is difficult to determine the routers on the path to the new access router and to reserve resources at each of them. For instance, the extension of RSVP presented in [13] uses so-called proxy agents (e.g., in the access routers) to set up a reservation on behalf of the mobile node. Similar to DRM discovery, the proxy agent for a new, prospective access point has to be discovered by additional protocols.

Another issue is the use of alternative routes in case of router failure. For the hop-by-hop case, a router failure is not directly noticed by the terminals, but re-establishment

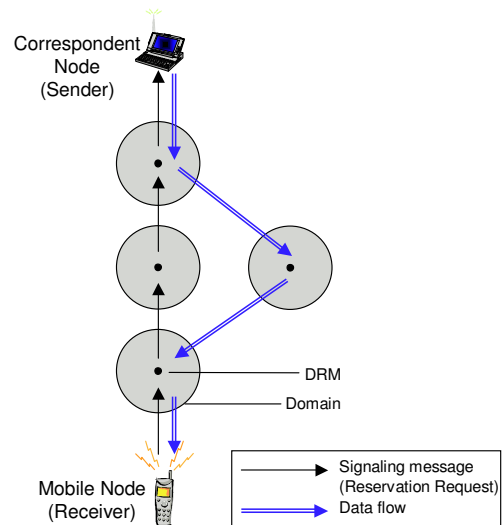


Figure 2: Signaling receiver-initiated reservations

may require new signaling from the terminals. In comparison, a DRM can be notified of local changes and can adapt the reservation [14]. On the other hand, the DRM itself is a single point of failure, though the reliability can be improved with redundant systems.

3.2 Receiver-Initiated Reservations

Support for receiver-initiated reservations is an important signaling protocol feature, especially for mobile networks. Without this feature, the sender must always initiate the signaling. A mobile node, however, acts usually as sender or receiver, often performing both roles simultaneously. If a signaling protocol at application level between the two end-systems is available, it could be used to trigger a sender-initiated reservation from the receiver-side application.

The basic problem with receiver-initiated reservations lies in the fact that there is no easy way to find out the forwarding path of packets from the reverse direction. This is caused by the unidirectional nature of routing, i.e. only the path and next hop towards the destination of a packet is known, but not the previous router that forwarded it. RSVP's solution to this problem is to install state while a PATH control message is forwarded, and, to use this state for finding and traversing the path from the reverse direction. However, this requires that the sender is sending a PATH control message first and that the reverse forwarding path state is updated frequently to detect route changes.

The problem of unidirectional paths exists at the inter-domain level as well as at the intra-domain level. Thus, it is hard to find out the previous autonomous system (AS) as well as the ingress router into the own domain.

The signaling path must follow the data path (at domain level) in order to reserve resources properly. Figure 2, however, shows that the data forwarding path may be different than the signaling path in the reverse (with respect to the data path) direction. There are several options to trigger the sender to initiate a reservation on the data path when considering mobility.

1. The MN, that is the receiver in this case, sends a reservation request to the DRM of its own domain.

The DRM forwards this request DRM-hop by DRM-hop (along the DRM chain of trust) to the sender. In this case, DRMs do not really process the signaling message (particularly no state is installed), but simply forward it to the destination domain. Then a reservation request is sent from the corresponding node DRM. There are two obvious advantages of this approach:

- the MN has to maintain only a signaling transport connection with the corresponding DRM in its own domain. This transport protocol may use an optimization for achieving a better performance over the wireless link.
- the approach fits also well together with an intra-domain handover, i.e., independently of inter- and intra-domain handover the MN has to signal only its own DRM.

A further advantage is that in some cases of a local handover, the DRM can locally adapt to the route.

2. The MN establishes a direct signaling transport connection to the correspondent node (CN) and sends a reservation request directly to the sender. This has the disadvantage of requiring the MN to maintain end-to-end signaling transport connections to all its correspondent nodes in addition to the signaling transport connection to the DRM of its own domain. This means also that the receiver must use different signaling transport connections for issuing inter- and intra-domain handovers, i.e., the signaling protocol must be aware of the current handover type. However, this approach has better scalability properties compared to the first approach, because these signaling messages do not have to be forwarded by DRMs.

A further requirement is that reservation setup shall not interfere with mobility for end-to-end signaling. This requires that signaling messages follow mobile nodes and hence home addresses of the nodes are used for signaling instead of possibly changing care-of-addresses. For the first solution, the signaling transport connection between DRM and MN should be transparent for the node movement. Thus, even if a mobile node left the DRM's domain, the signaling transport connection should track this change. However, even if the signaling connection is already closed, the old DRM may forward requests to the new DRM. For the second solution, both nodes use an end-to-end signaling connection that is transparent for node movement.

Another, related issue occurs in case of bi-directional reservations, which are required for many applications. An interesting optimization is to combine the signaling needed to set up two reservations in both directions. However, due to possible asymmetric data paths it does not make sense to explicitly combine the requests for both directions into a single protocol data unit. Even if the paths at domain level are symmetric, processing both reservations together is still difficult in transit domains, because the ingress router is not known in one direction.

4 QoS Signaling for Handovers

A flexible QoS signaling protocol has to deal with different handover types, e.g., hard, soft, anticipated handover, and variations of these. This section focuses on QoS signaling for different handovers for the previously presented architecture. The particular goal is to analyze different handover types and to find a signaling solution which covers all potential cases. Our protocol MARSP is designed to consider all these issues, it supports both sender- and receiver-initiated reservations. The specification of MARSP is derived from an integrated handover state model (presented in section 5). Particularly, this integrated state model allows one to describe transitions between different handover types. In other words, if the operation of one handover type fails, the handover can still be finished by using another handover type. This represents an increasingly important feature, especially when applied in heterogeneous mobile networks including several wireless access technologies.

4.1 Handover distinctions and scenarios

Several types of handovers may be considered for IP-based mobile networks. Each of the handover variants can be applied to a particular configuration of the network. Our goal is to design signaling protocols which support multiple handover types. It is conceivable that a RAN may not support all handover variants (e.g., soft handovers). Furthermore, MNs may also show different capabilities. Some nodes are only able to maintain a single 'connection' at IP layer without being able to listen or scan for other APs simultaneously. Thus, these nodes have to drop their current connection before selecting a new AR. Other nodes are able to scan for new access points while still staying connected to the current AR. Different handover cases can be distinguished that have impact on mobility management and resource reservations. At first, a MN may perform an *intra-domain handover* or an *inter-domain handover* depending on the location of the new AR. Several optimizations for intra-domain handovers are available, that result in faster signaling procedures because the signaling messages may stay local within the domain, and, there is no need to perform a full re-authentication.

A further important distinction can be made between vertical and horizontal handovers. It is possible that the currently allocated resources must be adapted if a vertical handover will be performed, because different link layer technologies offer very different capabilities (e.g., wireless LANs and GSM with respect to bandwidth). Consequently, this adaptation may require signaling at the application level as well as signaling for resource adaptation along the complete path between CN and MN, even for intra-domain handovers.

When considering QoS, the actual handover decision depends on two main criteria:

- Signal availability—This comprises radio parameters like signal strength, signal-to-noise ratio, etc. The IP layer must be informed by lower layers about conditions and availability of radio connections.
- Resource availability—When carrying out a handover to a new AR, it has to be ensured that the avail-

able resources on and to that AR are sufficient to satisfy the QoS requirements on the MN. A handover can also be rejected or directed to another adjacent AR in order to balance the load in a group of radio cells.

Therefore, handover strategies for IP-based mobile networks should be based on both criteria. However, for some cases of anticipated handover or vertical handover no signal measurements may be available. A consideration of these handover types is especially important, as they are expected to be used more often in future, heterogeneous networks.

4.2 Example for anticipated handover

In this section we show an example of the operation of the protocol to emphasize the requirements and assumptions used in our approach. The example in Figure 3 shows a message sequence diagram for *anticipated handover in an inter-domain case*. The advantage of an anticipated handover is that resource management signaling can be performed before attaching to a new access point. The following example presents the interaction between DRMs of *different domains*.

At first, the MN detects new access points and selects its target AR. Subsequently, a request for changing the MNs point of attachment from AR₁ to AR₂ is issued to the current DRM₁ (RChgReq). This DRM must detect that AR₂ is located in a different domain and has to determine the responsible DRM (a special DNS entry may be used for this purpose).

A handover request (RExtHoReq) is subsequently sent from DRM₁ to DRM₂ in order to request resources from a downstream DRM. Depending on resource availability, DRM₂ sends a corresponding response message (RExtHoRsp) back to DRM₁, which, then, in turn informs the MN about the result (RChgRsp). If resource allocation has succeeded at DRM₂ it waits for a handover confirmation message of the MN (RHoCompl). If this message is not received within a certain time, the pre-reserved resources are automatically released. Thus, the MN will connect to the new AR and is, then, able to confirm the completion of the anticipated handover procedure. DRM₂ informs DRM₁ that the reserved resources in the old domain are not longer used (RExtRelReq).

DRM₁ can then explicitly release unused resources. Otherwise the reservation will time out if no refresh messages are received from the MN within a certain time period.

5 Integrated Handover State Model

A handover can be described with the Integrated State Model presented in Fig. 4. It forms the basis for the further design of MARSP on the interface to the MN. The state model is drawn by using a grid with connection states on the horizontal axis and reservation states on the vertical axis (see below). By using this grid, any possible handover type can be described with a path from the top left (state S) to the bottom right (state F). The intermediate handover states are denoted as H₁ to H₁₁. The depicted transitions in the state model do not comprise reject or failure transitions in order to simplify the model.

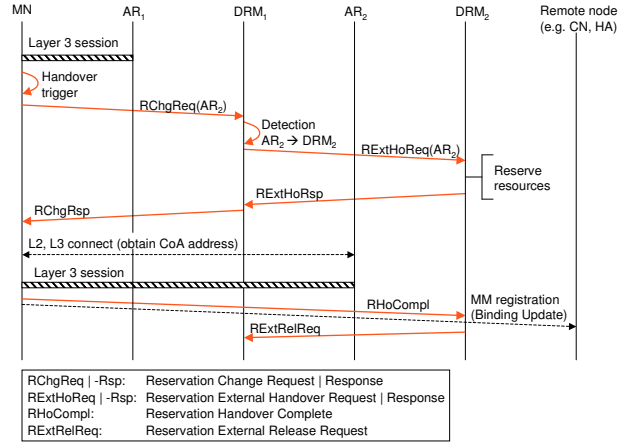


Figure 3: Sequence diagram for inter-domain anticipated handover

In the following a description of the horizontal and vertical axis of the diagram is given.

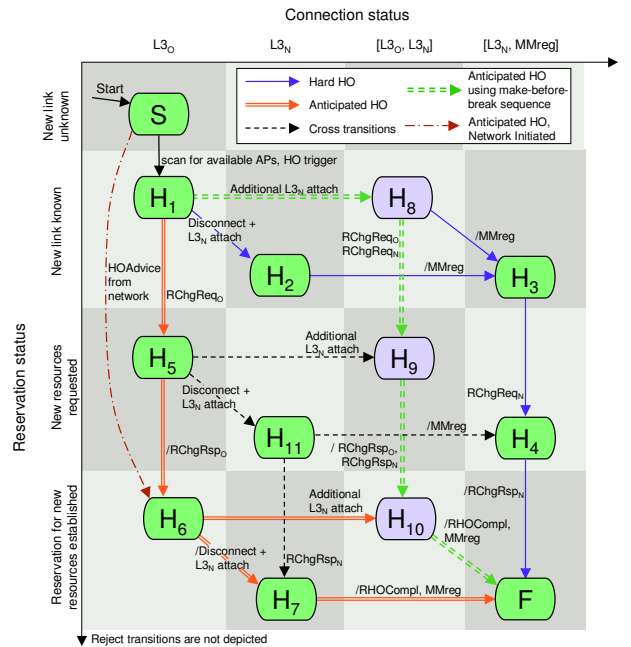


Figure 4: Integrated Handover Model

The integrated state model describes the connection states of handover sequences on the horizontal axis. The states comprise connection with the old AR at layer three (L3_O), connection with the new AR at layer 3 (L3_N), and a connection with both ARs simultaneously (L3_O,L3_N). Moreover, a separate state with conducted mobility management registration (i.e., sending of binding updates) is introduced (L3_N, MMreg) in order to describe the trigger for the handover of data packets between the old and the new data path. In a similar manner the significant states of resource reservation are described on the vertical axis.

The main benefit of this integrated state model is that in case of any network change it can be switched appropriately between handover types to preserve as much QoS as possible. When the operation of one handover type fails, the handover can still be finished by using another handover type, e.g., with a hard handover instead of an anticipated handover.

An isolated view of individual handover cases would not allow transitions between different handover types. The state model only describes transitions which are important for resource management. We only make minimal assumptions about mobility registration in order to achieve solutions for different mobility management mechanisms. Resources may be reserved for the new path before mobility management switches the data flow to it or after the MN registered at the new AR.

In the following we describe the transitions for the anticipated handover type in the state model, since this handover type represents a preferable method for efficient and fast handovers. By using the anticipated handover type the resources are already reserved when the actual handover is performed. Hence, there is no need to send packets without traffic guarantees for a certain time interval after handover. The path for anticipated handover consists of the following transitions:

- S–H₁: The MN scans for available APs or makes use of potential handover triggers.
- H₁–H₅: In the state H₁, the MN may be aware of several alternative ARs for the handover. A significant advantage of anticipated handover is that the MN has several target ARs to choose from as long as the signaling is still carried out over the old AR. Therefore, it is possible to include the whole set of potential ARs in the resource request. This gives the network the possibility to prefer a specific AR for the handover.
- H₅–H₆: Depending on the available resources the DRM answers with a reply containing the identification of the best AR fulfilling the requirements. This reply already denotes the successful resource reservation for the new path. Our approach does not include a separate query for resources without reserving them.
- H₆–H₇: Now the MN disconnects from the old AR, connects to the new AR on layer two and layer three (in case the MN supports simultaneous layer three connections: H₆–H₁₀) and triggers a mobility management registration (MMreg). This procedure also illustrates the coupling of mobility management signaling with resource signaling procedures.
- H₇–F: The MN sends a message about completion of the handover process to the network (*RHoCompl*). This message is needed in case of anticipated handover to show that the MN really connects to the AR where it has reserved resources before. Furthermore, the *RHoCompl* message may trigger the release of unused resources in the network after handover. Additionally to resource management, the MN triggers a mobility management registration in this step.

An advantage of the integrated handover model is, that it also shows combinations of the previously mentioned handover types. For example when layer three is unexpectedly lost to the old AR during an anticipated handover (state H₅), it is still possible to finish the handover as a hard handover without the need for a retransmission of the resource request.

6 Conclusions

We have discussed several approaches and design issues for mobile, IP-based networks. Based on this, we presented our QoS signaling architecture with its associated signaling protocol MARSP. The large variety of different handover types requires an integrated signaling solution. A novel, integrated model was developed to describe the different cases in a unified model. This permits to switch between these handover cases dynamically during a handover. The designed QoS signaling protocol supports all these handover cases, including anticipated handover as well as transitions between different handover types.

References

- [1] S. Blake *et al.*, "An Architecture for Differentiated Services," RFC 2475, IETF, Dec. 1998.
- [2] R. Braden, "Integrated Services in the Internet Architecture: an Overview," RFC 1633, IETF, June 1994.
- [3] C. Perkins, "Mobile IP," *IEEE Communications Magazine*, May 2002.
- [4] C. I. Bauer and S. J. Rees, "Classification of handover schemes within a cellular environment," PIMRC 2002, Sept. 2002.
- [5] W.-S. Soh and H. S. Kim, "QoS Provisioning in Cellular Networks Based on Mobility Prediction Techniques," *IEEE Communications Magazine*, Jan. 2003.
- [6] J. Manner, A. López, A. Mihailovic, H. Velayos, E. Hepworth, and Y. Khouaja, "Evaluation of Mobility and QoS Interaction," *Computer Networks, Elsevier Science Publisher*, vol. 38, Feb. 2002.
- [7] R. Braden, S. Berson, S. Herzog, S. Jamin, and L. Zhang, "Resource ReSerVation Protocol (RSVP) – Version 1," RFC 2205 (Standard), IETF, Sept. 1997.
- [8] B. Moon and H. Aghvami, "Reliable RSVP Path Reservation For Multimedia Communications under an IP Mobility Scenario," *IEEE Wireless Communications*, Oct. 2002.
- [9] "Next Steps in Signaling (nsis) Charter," <http://www.ietf.org/html.charters/nsis-charter.html>, 2003.
- [10] T. Engel, H. Granzer, B. Koch, M. Winter, P. Sampatakos, I. Venieris, H. Hussmann, F. Ricciato, and S. Salsano, "AQUILA: adaptive resource control for QoS using an IP-based layered architecture," *IEEE Communications Magazine*, Jan. 2003.
- [11] V. Marques, R. L. Aguiar, C. Garcia, C. B. Jose Ignacio Moreno, E. Melin, and M. Liebsch, "An IP-Based QoS Architecture for 4G Operator Scenarios," *IEEE Wireless Communications*, June 2003.
- [12] E. Mykoniati, C. Charalampous, P. Georgatsos, T. Damlatis, D. Goderis, P. Trimintzios, G. Pavlou, and D. Griffin, "Admission control for providing QoS in DiffServ IP networks: the TEQUILA approach," *IEEE Communications Magazine*, Jan. 2003.
- [13] A. Talukdar, B. Badrinathm, and A. Acharya, "MRSVP: A Resource Reservation Protocol for an Integrated Services Network with Mobile Hosts," *Wireless Networks*, vol. 7, Jan. 2001.
- [14] O. Schelén and S. Pink, "Resource Sharing in Advance Reservation Agents," *Journal of High Speed Networks, Special issue on Multimedia Networking*, vol. 7, no. 3-4, 1998.