Mobility Management and QoS Support in Wireless Environments: Trends and Open issues

A. Kaloxylos[1], S. Paskalis[2], D. Vali[3], A. Boucouvalas[1]

1. Department of Telecommunications Science and Technology, University of Peloponnese
2. Department of Informatics and Telecommunications, University of Athens
3. OTE Research, Hellenic Telecommunications Organization - OTE S.A,, Greece

Abstract

For quite some time researchers were trying to find viable means to support the quality of services for fixed IP users. This was a difficult task and even up until today there is no universally accepted mechanism to assure the quality of an active service from the one end to the other. As years were passing, the need of IP users to keep their network connectivity while on the move, introduced a family of mobility management protocols. However, it was soon noted that these mobility management protocols were inter-working rather inefficiently with the protocols for the quality of services' support. Thus, new protocols are under design to tackle this issue. However, even with these new protocols there are important issues left unchallenged. This chapter provides all the necessary information for this very interesting research area and its current status.

Keywords

Mobility management, Quality of Service, mobile networks, QoS session, handoff

INTRODUCTION

The wireless and mobile communication devices industry sector is experiencing an enormous growth. People are getting accustomed to be productive while on the move, utilizing the capabilities their wireless and mobile devices offer. The connectivity support, one of the most fundamental requirements, is certain to rely on the ubiquitous Internet Protocol (IP). There are, however, some fundamental challenges that need to be overcome in order to be able to use the same protocol architecture as the fixed users do.

Mobility support, the first of them, stems from the users' need to communicate in every imaginable way, while on the road, on the train, at home or in the office. The IP protocol suite needs to adapt to nowadays' era and start offering uninterruptible connectivity to devices and users, irrespective of their location and movement conditions. Several years of research work have been performed to accommodate mobility management in IP. The primary efforts focused on the ability of the mobile node to communicate with any other Internet connected node while being attached to a different network, which led to a procedure satisfying the target set, but not the actual mobility requirements of moving and maintaining an uninterrupted, good-quality communication. After the baseline for mobility support was set, multiple optimization efforts began to achieve minimization of disruption time, optimization of resources used and generally satisfaction of the mobile users.

Quality support, the second of the essential requirements, has a more complicated history. The IP protocol stack follows the end-to-end principle, which dictates to keep functionality and complexity out of the core of the network and push it to the end-points. In other words, end-devices can and should bear the complexity of evolution and capability additions, whereas internal devices, i.e. the routers, should be kept as simple as possible. The simplicity requirements for the routers extend to simple processing for data packets, and memory-less operation. Following the guidelines strictly, every incoming packet would receive the same priority and would be forwarded to the same output

queue waiting for transmission. In today's Internet, however, not all packets are created equal. Some packet flows can cope with long delays and/or packet loss, whereas other flows can only bear extremely small delays and jitter (real-time voice or video communication). The desire to offer prioritized treatment to certain packets, so as to offer either guarantees or just better service, led to Quality of Service (QoS) schemes, which add processing and possibly state-fullness requirements to the routers.

The aforementioned efforts to provide mobility support and QoS guarantees in the Internet began – and mostly continued independently, leading to inefficiencies and/or incompatibilities. The most obvious and cited example is the usage of the end-points' IP addresses to refer also to a QoS state along the data path. This QoS state identifies the packets that will receive a certain priority treatment and needs to be modified in an end-to-end fashion when mobility causes the modification of an end-point's IP address.. Thus, the net result is the invalidation of the existing QoS setup along the data path, and, thus, the need to re-establish the QoS reservation according to the new IP address and the need to tear down the now invalidated QoS state throughout the data path containing the previous IP address.

The mobility management and QoS incompatibilities have been identified in the research literature practically from the beginning of the individual standardization efforts. However, the relative isolation between the QoS and mobility management specification groups, the experimental nature of the schemes, and the lack of operational use cases, prevented any harmonization attempt at least in the relevant standardization avenues. This finally has been changed with the creation of the IETF "Next Steps in Signaling" working group (Next Steps In Signaling IETF Working Group), which undertook the difficult task to propose a generic signaling architecture for the Internet, capable of dealing with the multiple and contradictory signaling needs for the Internet infrastructure. Among them is also the undisturbed interaction between mobility management and QoS signaling and state maintenance.

Another interesting issue that has to be dealt with stems from the fact that dynamic control of end-to-end QoS schemes requires that signaling has to travel from one end to the other each time a new session is to be established. However, if individual flow states are maintained at each router along the data path, scaling issues are raised, especially for the core network routers. To ameliorate this problem, aggregation of signaling state information is required. The scheme that one should use to minimize the processing load and the signaling information stored in the routers still remains an open issue, but is definitely something that needs to be specified if QoS provisioning is ever to be deployed in a real end-to-end fashion. Unfortunately, regarding their interaction with mobility management, the proposed aggregation schemes exhibit similar incompatibilities as the existing individual QoS proposals. This is a rather important drawback since it is expected that in the future, users will be able to vertically handoff their connections from one radio access technology to another and even from one Internet Service Provider to another, thus moving partly away from an aggregation path.

All the aforementioned issues are described in this chapter that is organized as follows. In the next section we provide a description of the mobility management protocols and what is expected to prevail in the near future. Then, a discussion follows about the advantages and inefficiencies of QoS mechanisms. We present those that have already been standardized or have been proposed to deal with specific issues (e.g., aggregation of signaling states). Next, we discuss on current trends to tackle the interworking issues between QoS protocols and mobility management mechanisms. Finally, we conclude this chapter.

PRESENTATION OF MOBILITY MANAGEMENT PROTOCOLS

As it is well known, the core protocols of the Internet (i.e., IP and TCP) were not designed to handle mobile terminals. In these networks, the destination IP address has topological information and is used to determine the next hop to forward a packet. On the transport layer, the TCP protocol maintains information in the form <source IP address, source port number, destination IP address, destination port number>. Thus, without any additional provision, supporting mobility in TCP/IP networks faces a conflict. From the one hand, a terminal entering a new network has to receive a new IP address in order to be reachable. On the other hand, a TCP connection has to keep the source IP address constant because otherwise the connection will fail since the new information will not correspond to the old one. As we will present, the Internet community came up with a mobility management protocol that makes mobility transparent to the higher level protocols and requires minimum changes to the existing infrastructure.

Mobility management consists of two distinct sets of operations. The first one has to do with the location management of the terminal. This set of operations includes all the procedures for updating the knowledge of the network about the current location of a terminal as well as the procedures for finding the current location of a terminal when it is required to deliver data to it. The second set has to do with the handover of an active flow from the old data path to the new one.

All location management mechanisms, independently if they are deployed in cellular or IP networks, follow some common principles. The overall network is divided into different areas. For each terminal one of these areas is called the “home area”. Inside the home area there is an entity that is aware at any time for the current location of a terminal. In terms of Mobile IPv4 (Perkins, 2002), this entity is called the home agent (HA) and maintains a mobility binding table with the following information: <permanent home address, temporary care-of address, association lifetime>. This information is kept for each terminal currently located outside its home area. To maintain the information of this table up to date, each time a terminal moves into a “foreign area” it will receive a temporary IP address that is called care of address (cCOA). As soon as the terminal receives this address it will inform the home agent. This way a terminal may have two IP addresses. The permanent home address is used to identify the terminal while the temporary care-of address represents the current location of the host and used mainly for routing purposes (i.e., reaching the terminal in its current location). This temporary address is usually assigned to a terminal by another specialized entity called “foreign agent” (FA).

In summary, when a terminal enters into a new area, it discovers the existence of a new FA. Then, it requests a registration with the FA and send its home address, its media address (e.g., Ethernet MAC address) and the address of its HA. The FA then sends a registration to the HA by sending a message that contains the home address of the terminal and its own address. This way, location update is performed in the HA. When a node, called correspondent node (CN) in the MIP terminology, wishes to communicate with the mobile terminal, it will send packets using the permanent home IP address of the terminal. The packets will end up as expected in the home area of the terminal. There, they will be intercepted by the HA. The HA, after consulting the mobility binding table, will construct a new IP packet in which the payload will be the original IP packet sent by the CN. The header of the new IP packet will have the temporary COA as its destination address. This process is called encapsulation (Perkins, 1996) or tunneling. The new packet will reach the FA that is responsible to de-capsulate the original packet and use the media address of the terminal to forward the packet to it. Note that if the mobile node wants to send a packet to the CN, this can be routed directly. This forms an asymmetry in the routing of packets between the two communicating end nodes and is called *triangular routing*. The problem of the triangular routing has been solved in the MIPV6 (Johnson, Perkins & Arkko, 2004)

since the COA can be communicated to the CN directly. Another important difference from MIPv4, is that in MIPv6 there is no need to use any FA, since the required functionality is supported by the mobile terminals themselves.

As we already mentioned, apart from location management, mobility management has to do handing over an active flow from the old data path to the new path. This functionality by definition built in the MIP. If a terminal receives a new COA while having an active communication flow then the new packets will reach the terminal in its new location as soon as the bindings have been updated with the new COA. Of course, if no special provision is taken then, until this binding update takes place, some packets will be forwarded to the old location of the terminal, even though it has already change its location, and will be lost. To tackle this issue several alternatives have been proposed like packet forwarding from the old location to the new one, or prediction schemes that start sending copies of packets to possible future locations of a terminal.

Because of the dominance of the IP protocol and the simplicity of MIP, it is expected that it will be the de-facto standard even for future mobile cellular networks. However, MIP faces some serious problems in terms of the flow disruption time and packet losses. This is why alternative solutions like Fast MIP (Koodli, 2008), Hierarchical MIP (Soliman, Castellucia, Malki & Bellier, 2005) and Proxy MIP (Gundavelli, Devarapalli, Chowdhury & Patil, 2008) have been introduced to deal with the minimization of the disruption time during a HO execution, the localization of signaling exchanges inside a pre-defined area and the transparency of mobility support for the end users.

More specifically, a handover involves not only layer 3 operations (e.g., acquire a new IP address, notify the HA etc), but also some layer 2 actions (e.g., scan for neighboring access points, associate and de-associate in link layer etc). FMIP has been specifically designed to take into consideration these operations and minimize as much as possible the disruption time by having the terminal collecting information and the network components to be configured appropriately before the execution of a handover. In FMIP, the terminal can request the serving access router (called Previous Access Router – PAR) to send it information about a target router for a handover (called New Access Router – NAR). With the received information the terminal can form a perspective new COA. Moreover, it can request the PAR to start transmitting any received packets for the terminal to the NAR, by having the PAR to establish an appropriate tunnel between the two routers. To do this, the PAR will communicate with the NAR to check several issues (e.g., the validity of the perspective COA). When everything is ready the terminal is notified by the PAR to hand-off to the NAR. To do this, the terminal will perform a layer 2 association to the new network (e.g., a WLAN association) and will send to the NAR a notification to inform it that it has been attached under its control area and that the NAR should start forwarding packets for it. Up to this point, the packets are arriving to the terminal following the tunnel between the PAR and the NAR. The tunnel between the two nodes will remain valid until the terminal completes its binding updates with its CNs. Obviously, the aforementioned description is valid only when there is enough time for the terminal to collect all the required information through the old path and the routers to be configured appropriately. If however there is not sufficient time then it is most likely that several operations will be made through the NAR, resulting in an increased disruption time.

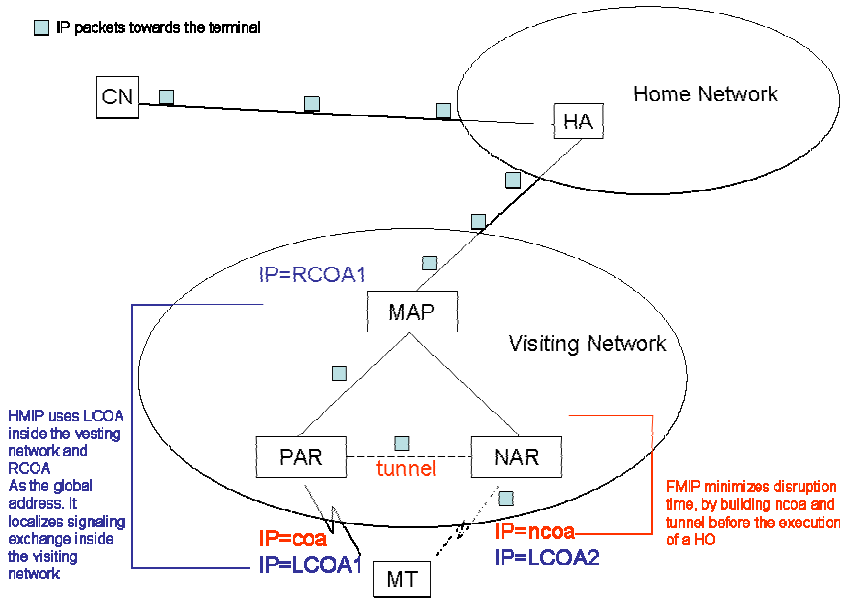

Figure 1: Basic architecture for FMIPv6 and HMIPv6

In the case of Hierarchical MIP, the target is to achieve smaller disruption times by keeping any signaling exchange inside a pre-defined area. This is achieved with the use of two different COAs namely the Regional COA (RCOA) and the Link COA (LCOA). More specifically, inside a pre-defined area a new component is introduced, called Mobility Anchor Point – MAP. This component actually acts as a "local HA" for the terminals that remain inside the domain of the MAP. To reach any terminal, CNs are using a globally visible address (i.e., RCOA). Using this address, the packets will eventually arrive in the MAP entity. Note that as long as a terminal remains inside the same domain, it uses the same RCOA. Any packets targeted to the terminal are intercepted by the MAP entity and encapsulated using the LCOA. When a terminal moves inside a domain it may acquire a new LCOA. This information has to be transferred up to the MAP entity. It does not however needs to be transferred up the HA of the terminal. This is because the HA only knows about the RCOA that is unchanged as long as the terminal keeps moving inside the same domain. The only case the HA needs to be updated is when the terminal moves into the area of a new MAP entity and receives a new RCOA. HMIP presents several advantages. First of all, signaling does not usually require traveling long distances and remains local. This means that the network is burdened with less signaling exchanges and more importantly experiences a smaller handover execution time. Moreover, there is some location privacy for the users since their exact location inside a domain is not known to the outside world. However, this solution requires the introduction of a new component (i.e., MAP). Figure 1 illustrates the operation of FMIP and HMIP.

Finally, we briefly present the case of Proxy MIP since it has been selected for the interworking of future mobile cellular networks (i.e., LTE/SAE). The main goal of this mechanism is to execute mobility management functions in a totally transparent way for the terminals. In other words additional functions are added inside the network in order to leave the terminals totally unaware of any mobility management procedures. For this, two new functions are introduced namely the Local Mobility Anchor (LMA) and the Mobile Access Gateway (MAG). The former acts as the "home agent" for a terminal located inside a specific domain. The latter is a function located on an access router and handles all mobility related signaling for a terminal that is attached to the access router. When a terminal enters a Proxy MIP domain, the terminal communicates with a MAG and after being authenticated, MAG is responsible to contact LMA and create a bi-directional tunnel among them for

the specific terminal. The MAG function is responsible to keep track of any movement from the terminal and update accordingly the LMA.

Although there has been vast research efforts to develop mobility management protocols, until now there has been no real adoption of these protocols in every day life. There are many reasons for this. First of all, standard MIP cannot easily support real time services when the disruption time during a handover is measured to be in the order of some seconds. Moreover, the specifications are considered to be quite "heavy" to be executed on small mobile end devices. Finally, operators are in favor of network based MIP solutions such as PMIPV6. Obviously there are several open issues to be dealt like how to combine the aforementioned techniques to achieve better results and also how these protocols can be combined with the appropriate QoS support mechanisms.

QUALITY OF SERVICE MECHANISMS

QoS mechanisms rely on prioritizing some traffic over other less important/urgent data. The prioritization usually aims at providing guaranteed minimum bandwidth, and occasionally other parameters, such as guaranteed maximum delays and jitter. The specifics of QoS provided depend directly on the underlying link-layer technology and the provisions it is able to make. However, the common attribute in all QoS service provisioning schemes is that in order to provide different services to packets belonging to different service groups, the network, i.e. the routers that comprise it, must have a way to differentiate between the packets. In other words, the existence of some packet classification criteria at each router is the single common attribute in every QoS architecture solution in the Internet.

The existence of the packet classification mechanisms in every router along the data path implies more or less two other distinct functionalities:

- QoS state maintenance at each router, and
- QoS signaling, to establish, maintain and teardown the QoS state.

The Internet community introduced the Integrated Services (IntServ) architecture (Braden, Clark & Shenker, 1994) to implement the vision of end-to-end QoS services into specifications. IntServ supports end-to-end signaling, QoS state establishment and management for per-flow differentiated treatment in intermediate routers along the data path. The signaling protocol that emerged to meet the Integrated Services requirements is RSVP (Resource reSerVation Protocol) (Braden, Zhang, Berson, Herzog & Jamin, 1997). The IntServ architecture was designed to facilitate every QoS element (router functionality, signaling, and accounting) in a fine-grained manner. To achieve this goal, IntServ was founded on the underlying assumption that a homogeneous Internet environment equipped with IntServ enabled routers and end hosts would be the common case.

The IntServ architecture in general and the RSVP protocol in particular received criticism, mainly due to the scalability issues raised by the state maintenance for every data flow in intermediate routers across the end-to-end path. The Internet community considered, therefore, other alternatives to the QoS provision problem. This time, the target was a lightweight QoS architecture putting as little burden in the routers as possible and providing coarse-grained traffic prioritization based on the statically contracted Service Level Agreements (SLAs) between users and the network. SLAs specify the amount and types of traffic each side has agreed to send and receive. The outcome was DiffServ (Differentiated Architecture) (Blake, Black, Carlson, Davies, Wang & Weiss, 1998).

DiffServ networks are statically configured to support a small set of QoS levels (PHBs – Per Hop Behaviors) and do not use any QoS signaling for state establishment and maintenance in routers.

DiffServ routers prioritize the data packets according to a 6-bit field in the IP packet header (DSCP, DiffServ Code Point) that reflects the requested QoS level. This procedure results in aggregating reservations for different users sharing the same QoS level. Appropriate packet marking takes place either at end-hosts or at DiffServ edge routers before the traffic enters the DiffServ network. DiffServ edge routers perform, in addition, traffic classification and traffic conditioning procedures (including metering, marking, shaping and policing) based on the contracted SLAs. In other words DiffServ establishes a minimum, static QoS state at each router, eliminating any QoS signaling and QoS state differentiation, the opposite of the ultra-fine grained IntServ approach.

MPLS (Rosen, Viswanathan & Callon, 2001) employs a somewhat similar treatment to packets as does DiffServ. It assigns each packet an MPLS label (32-bit header) and forwards it in the interior of an MPLS network solely based on the value of that label, without any packet inspection. The difference to the DiffServ approach is that the MPLS network creates MPLS paths for each different label throughout the MPLS network, and forwards the packets through the established paths. Traditional IP routing is not employed inside the MPLS network, and the provider has ample opportunities to configure the service received as well as the path taken for each individual MPLS label path. Note that MPLS was not designed as a means to provide QoS but rather a faster way to route packets. However, it has been recognized that its built-in functionality can be used for QoS purposes if handled appropriately.

The realization of the fact that the Internet is a concatenation of technologically and administratively different domains (Autonomous Systems - ASs) led to the identification of separate QoS techniques for the efficient support of intra- and inter-domain QoS. Thus, a two-tier resource management model was proposed in (Terzis, Wang, Ogawa & Zhang, 1999), with the intra-domain QoS signaling performing resource management inside a domain, and the inter-domain signaling managing resource allocation between domains. The two tiers must be closely coordinated to enable provision of the necessary end-to-end QoS support. The two-tier model increases the degrees of freedom regarding end-to-end QoS support, since each domain is free to choose any QoS support mechanism for allocating resources internally, as long as proper co-operation takes place with the respective inter-domain signaling protocol.

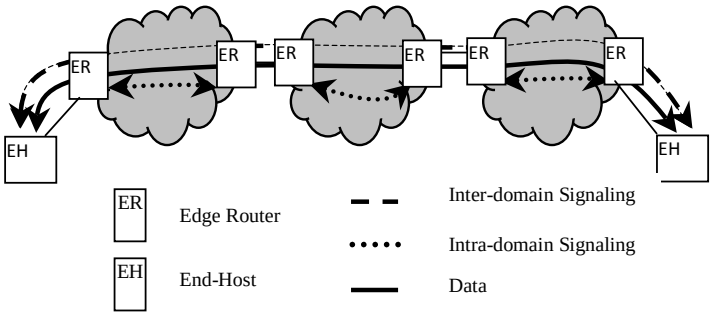

Figure 2. Two-tier QoS signaling

The two-tier signaling architecture, illustrated in Figure 2, implies that each domain is allowed to use its own QoS mechanism or protocol internally, allowing for concatenation of the various heterogeneous domains. The provision, however, of end-to-end QoS requires that appropriate interworking between the intra- and the inter-domain QoS protocols take place at the domain boundaries. Multiple configurations are possible with this approach, including both fine- and coarse-

grained QoS schemes, in the intra- and the inter-domain signaling, depending on the flexibility and simplicity desired to achieve.

One of the important trends in multi-tier QoS schemes is the de-coupling of QoS state (as in packet classification and resource reservation) and QoS signaling state. An example of that scenario is the maintenance of individual flow signaling state at the QoS signaling capable border routers and packet marking to impose intra-domain specific QoS handling (i.e. DiffServ marking or MPLS labeling).

A side-effect of the two-tier architecture is the need to maintain both intra- and inter-domain QoS state at the border routers as well as to perform the necessary mapping and parameter translation between them. As such, the QoS state (and by extension the QoS signaling if needed) needs to be implemented for both intra- and inter-domain signaling at the border routers performing the necessary transitional steps. The gain from imposing such a complexity weight on the border routers is the bare simplicity of the internal routers and the minimal QoS state and signaling needed. An extensive review of the tradeoffs imposed on the specific paradigm can be found in (Vali, Paskalis, Kaloxylos & Merakos, 2004).

There are only a few alternatives for aggregating signaling information. Arguably, the most straightforward way to aggregate signaling is to re-use the existing signaling protocol (Baker, Iturralde, Faucheur & Davie, 2001). Using a single RSVP reservation to aggregate other RSVP reservations across a transit routing region consolidates the signaling state inside that region in just one entry. Other approaches completely dismiss the use of RSVP and propose their own mechanisms for aggregation purposes. The DARIS (Dynamic Aggregation of Reservations for Internet Services) architecture (Bless, 2002) assumes the existence of a central resource management entity inside each DiffServ domain that has a complete knowledge and control of the resources inside the domain. DARIS enables the creation of an aggregate between two arbitrary domains as soon as a threshold of active common reservations between the two domains is exceeded.

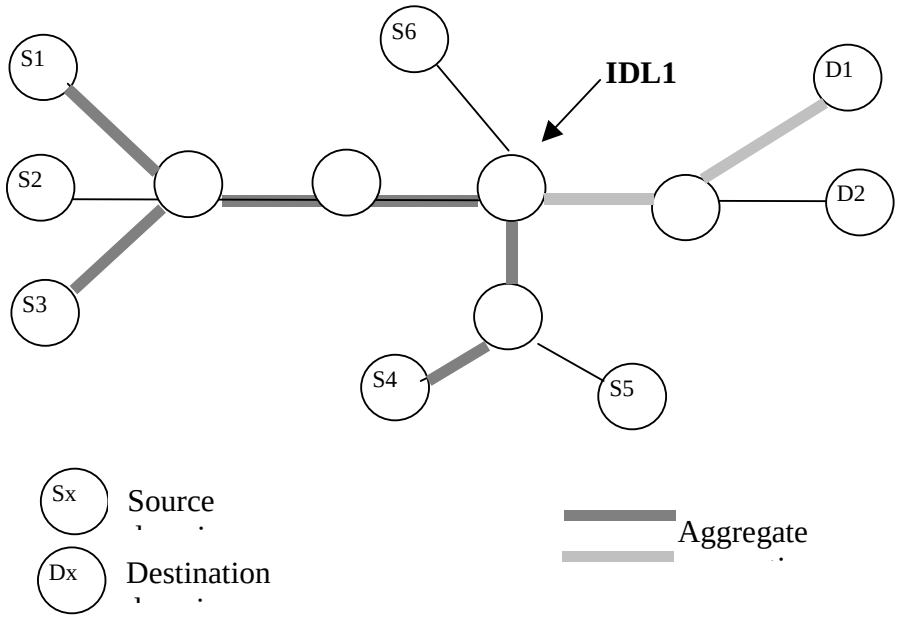

Figure 3. Signaling state aggregation toward domain D1

A couple of other aggregation techniques focus on the reservation at the AS (Autonomous System) level. A scheme designed for aggregate inter-domain usage between heterogeneous domains (Autonomous Systems) is the Border Gateway Reservation Protocol (BGRP) (Pan, Schulzrinne & Hahne, 2000). BGRP operates end-to-end only between domain border routers and aims at aggregating reservations between domains improving scalability. BGRP uses the sink-tree aggregation approach and performs reservation aggregation by building a sink tree for each destination domain (Figure 3). Reservations from different source domains that are destined towards the same destination domain are aggregated along the path, forming a sink-tree rooted at the destination domain edge router. The Shared-segment based Inter-domain Control Aggregation Protocol (SICAP) (Sofia, Guerin & Veiga, 2003) is another approach for supporting aggregate inter-domain reservations between Autonomous Systems (ASs).

The need for a generic Internet signaling framework led to the development of the NSIS framework (Hancock, Karagiannis, Joughney & den Bosch, 2005). NSIS consists of two layers, the underlying General Internet Signaling Transport (GIST) (Schulzrinne & Hancock, 2009 and the application specific NSIS Signaling Layer Protocol Protocol (NSLP), which in the QoS case is the QoS-NSLP (Manner, Karagiannis & McDonald, 2009). The QoS-specific signaling protocol of NSIS is similar in concept to RSVP, dealing with individual flows, and maintaining soft state for each of them. The important difference regarding mobility interaction is the choice of a Session ID irrelevant to the end-point location identifiers (IP addresses).

Each of the proposed QoS schemes provides a more or less different packet classification option for the packet flows. Table 1 lists some of those:

RSVP	5-tuple <Protocol, Source Address, Source Port, Destination Address, Destination Port>
DiffServ	DiffServ code point (DSCP) – a 6-bit field in the IP TOS field
MPLS	MPLS Label stack (shim header)
QoS-NSLP	Packet classifier object (more flexible than RSVP's 5-tuple)

Table 1. Packet classification arguments

Two- or multiple tier models usually employ a simple packet classification approach (DiffServ or MPLS), enjoying the fast and stateless operation for the intra-domain packet classification. The inter-domain packet classification, as well as the necessary signaling state required, is a much bigger issue. If end users are capable of QoS signaling, i.e. requesting and receiving specific time- and service-based QoS, then either RSVP or QoS-NSLP is used, the packet classification criteria are session-based and the signaling state contains information about each session. If only AS border routers handle QoS provisioning, then the packet classification may still be simple (i.e. DiffServ), but the signaling state contains information about session aggregations.

The most relevant to QoS and mobility management interaction attribute is the content of QoS signaling state, and especially the identity of each state. Table 2 presents the signaling state maintained for some of the presented QoS proposals:

RSVP	5-tuple <Protocol, Source Address, Source Port, Destination Address, Destination Port>
DiffServ	None
MPLS	None
BGRP	Autonomous System Number
SICAP	Intermediate border router (Intermediate De-aggregation Location – IDL)
DARIS	Intermediate router or AS
QoS-NSLP	Session ID (random 128 bit number)

Table 2. Packet classification arguments

MOBILITY AND QOS INTERACTION

The most important clash between IP mobility management assumptions and IP QoS assumptions is the consideration of the IP address as an immutable identifier for the end host. The historic dual consideration of the IP address as a geographic/topological qualification and a unique identifier provides many advantages when networked devices do not move. In our post-classic era, though, this is no longer the case, and mobility considerations dictate the use of several IP addresses from a node as it moves and changes points of attachment. The same logic applies to multi-homed devices, which connect to multiple networks simultaneously, seeking optimal connectivity.

The optimum utilization of access networks resources is a critical issue, especially in large domains, containing a significant number of wireless devices. In such networks, the administrator has to deal not only the scarce resources on the wireless links, but also with the efficient utilization of the expensive resources in the link(s) to the upstream ISP.

This task becomes more challenging when the wireless devices are also mobile ones.If no special care is provided mobile terminals face long service disruption times. Moreover, the delay to reorganize reserved resources in the end-to-end path (reserve resources in the new path and release the ones in the old path) results in a waste of network resources for this time period. This percentage is expected to be significant when a large number of mobile terminals resides in the access network or when the terminals demonstrate a high mobility rate.

Interaction with RSVP

The existing QoS protocols, such as RSVP, do not interact well with the IP address change, imposed by Mobile IP. If a mobile host (MH), with established RSVP data flows, performs a network layer handoff, it acquires a new IP address (Care of Address, COA) (Perkins, 2002; Johnson, Perkins & Arkko, 2004), and a new round of RSVP signaling exchanges must be triggered. RSVP creates soft session states in every intermediate router of the traffic flow. Each session is uniquely identified by the "Session" object, which is defined by the triplet <DestAddress, DestPort, [ProtocolId]>. Thus, the downlink reservation (the packet flow toward the mobile) becomes invalid, when the DestAddress parameter is modified due to mobility. The new uplink re-establishment is also affected, since the RSVP "Path" messages sent by the MH contain its new IP address in the "Sender Template" object. These messages sent and received by the MH at its new location are considered to correspond to a new session, and generate a new "Path" state (Braden & Zhang, 1997).

The major problems emerging from this mobility-QoS interaction are: (a) the sessions from and to the MH may not receive expected QoS treatment, and (b) the reserved resources that correspond to the

old COA will not be available to other traffic until the RSVP soft states expire thus, becoming what we will refer to as *stale sessions* and *stale reservations*, respectively. As illustrated in Figure 4, when the MH switches to a different point of attachment and assigns itself a new IP address, end-to-end RSVP signaling (Path/Resv messages) must traverse the network in order to re-establish the QoS reservation pointing to the current COA of the MH. Moreover, explicit teardown of the stale sessions (through PathTear/ResvTear messages) may be initiated, either by the MH or the correspondent node, in order to avoid resource waste until the stale reservations expire.

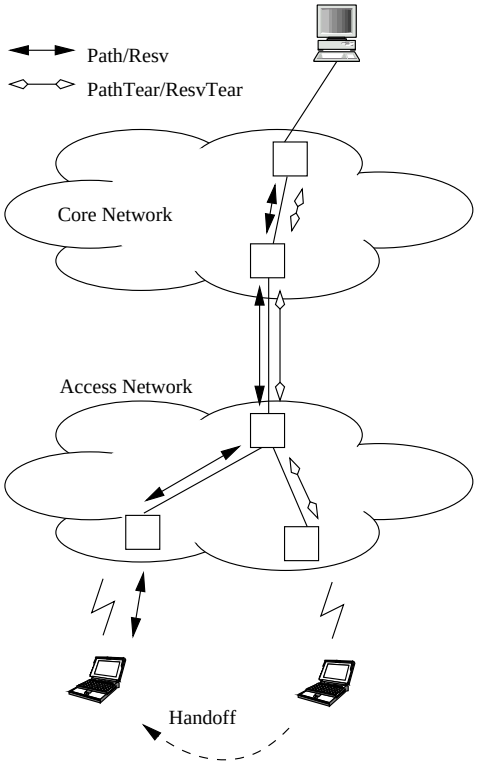

Figure 4. Handoff and RSVP interaction

The research community has identified the need for efficient interaction between mobility and QoS, and has proposed a number of different approaches to the problem. The various proposals try to meet the diverse requirements (Chaskar, 2003) from different angles.

To minimize the time needed for re-organization of the network resources, proposals included RSVP modifications and extensions that deal with context transfers between successive access routers (Kempf, 2002), proactive reservations to neighboring access routers with the use of mobile agents for reserving passive or pseudo-reserved resources, or multicasting in hierarchical or not domains (Talukdar, Badrinath & Acrarya, 2001; Huang & Chen, 2003; Lee, Kim, Chanson, Yu & Lee, 2003; Chang, Lee & Lee, 2005; Tseng, Lee, Liu & Wang, 2003; Chen & Huang, 2000). These solutions address the time minimization issue on the cost of complex procedures.

Other researchers suggested that one needs to find the cross-over router for a moving terminal and simply reconfigure the network in a way that the resources of the old branch are now reserved in the new branch (Moon & Aghvami, 2001, 2004). Some proposals limit their scope to an administrative domain (e.g. an access network) and propose either modifications to RSVP (Manner & Raatikainen, 2003) or the addition of functionality to cross-over routers (Paskalis, Kaloxylos, Zervas & Merakos, 2003).

The obvious solution to the state modification after a handoff is the de-coupling of state session identification from end point IP addresses. One of the proposals suggested modifying RSVP signaling, so that a unique session identity (possibly a random integer) was included in the Session and Sender Template fields (Thomas, 2002). The message processing rules should also be modified to deliver the same QoS to packets originating or destined to different IP addresses, but conforming to the same session ID. Similar approaches were suggested by (Shen, Lo, Seah & Ko, 2000), where the immutable Home Address assumes the role of session ID, and (Kuo & Ko, 2000), where the IPv6 Flow Label is the session reference identifier. The urge to create a signaling protocol with a mobility-immutable Session ID gave birth to the creation of the IETF NSIS Working Group and the proposal of a suite of signaling protocols, which dealt with RSVP shortcomings.

Interaction with QoS-NSLP

The development of QoS-NSLP had one particular feature regarding mobility considerations. QoS-NSLP decouples QoS state and flow identification. Session ID (SID), a cryptographically random number, which is probabilistically globally unique, is the state reference object. The NSIS Transport Layer Protocol (GIST) notices when a routing path associated with a SID changes, and provides a notification to the NSLP. It is then up to the NSLP to update the state information in the network (T. Sanda, Fu, Jeong, Manner & Tshofenig, 2009). Thus, the effect is an update to the states, not a full new request.

The actual signaling and state re-establishment/maintenance burden imposed on the network depends on the location of the cross-over router (the router where the old path and the new path converge), and whether the mobile end point changed IP addresses or not.

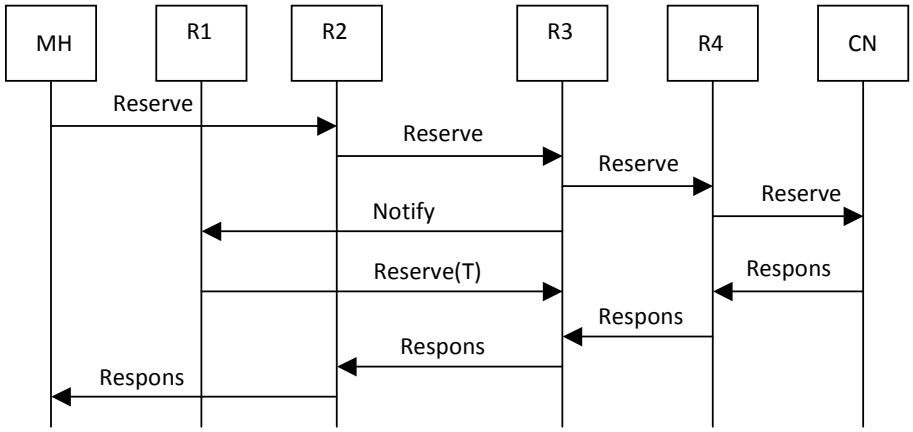

Figure 5. Message exchange after a handoff in QoS-NSLP

Figure 5 illustrates the typical message exchange for QoS-NSLP after a handoff, where the mobile host (MH) switched points of attachment and possibly changed its IP address. To re-establish a QoS reservation with its corresponding node (CN), it needs to re-initiate the reservation procedure sending RESERVE messages along the new data path toward the CN. The first QoS-NSLP aware router that receives a RESERVE message that contains the same Session ID, but different SII (Source Identification Information) or MRI (Message Routing Information), realizes that one of the end points

performed a handoff and that itself is the crossover router (CRN) for the specific session involving a handoff.

For the part of the path that did not contain any session state, (i.e. the MH→CRN part), Connection Admission Control, and state establishment must be processed as with any new QoS request. For the part of the path, though, after the crossover router (i.e. the CRN→CN part), the implications of QoS-NSLP signaling are just state updates to reflect the new conditions regarding the IP address of the mobile host, and, consequently, performed much faster and easier.

The crossover router can also issue a NOTIFY toward the previous location of the MH, which travels hop-by-hop, until it reaches its old access router (R1). R1 figures out that MH has left its network service and sends a teardown request RESERVE(T) toward the other end of the session, i.e. the correspondent node, which destroys QoS states along the way. CRN intercepts this message and discards it, since its purpose was to free the reserved resources in the now stale part of the old data path (i.e. R1→CRN).

The use of QoS-NSLP solves one of the biggest incompatibility problems between mobility management and QoS setup. Despite these efforts, the need for end-to-end signaling has not disappeared, since state updates regarding the packet classification scheme employed, must be re-installed along the data path. Note also that if localized signaling is desired, one must deploy a mobility management protocol such as HMIP, that "hides" the movement of the mobile host inside an HMIP domain, presenting a stable IP address as the MH COA to the outer environment.

Aggregation Resources and mobility

As already mentioned, there are only a few alternatives for aggregating signaling information. Their main operation is to aggregate a large number of signaling states into a single state through a specific path located between one aggregation and one de-aggregation point. As it is obvious, between these two points any signaling information for a specific flow is lost. This functionality is advantageous for fixed terminals since it eliminate a vast number of signaling states, especially from the core network routers.

Future scenarios expect users to be able to vertically handoff their connections from one radio access technology to another or even dynamically select to switch from one Operator, or Internet Service Provider, to another (e.g., handover from/to WLAN to/from WiMAX, UMTS etc). In these cases it is possible that the new end-to-end path, although sharing a large segment with the old one, will not include the previous aggregation or de-aggregation points. Since the intermediate routers do not have any means to recognize that some resources have already been reserved, under an aggregated signaling state, they will attempt to reserve from scratch new resources. Thus, resources need to be re-established in an end-to-end fashion, despite the fact that a portion of the previously established path could be re-used. This is actually a problem similar to the one arisen for the interworking between MIP and RSVP.

The existing standard and the few alternative proposals have not been designed to deal with this problem. Thus, additional functionality is needed. A possible solution to this problem has been proposed in (Kaloxylos, Vali, Paskalis, Panagiotou, Gonianakis & Zervas, 2006) where appropriate extensions to BGRP have been designed. However, since the signaling aggregation problem has not solved yet, we note that any future solution should take into consideration the inefficiencies that may be caused by the mobile terminals.

Conclusions

In this book chapter we present the main mobility management protocols that are present in the Internet and the QoS protocols that attempt to solve the end-to-end QoS support. In the last category we also present protocols that deal with the aggregation of QoS signaling states. Our aim is to identify the problems that arise from the interworking of these mobility management and QoS support protocols. The chapter identifies open issues, even in the latest standardization attempts and provides hints on how these can be tackled.

REFERENCES

Baker, F., Iturralde, C., Faucheur, F. L. & Davie, B. (2001). *RFC 3175: Aggregation of RSVP for IPv4 and IPv6 Reservations.*

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. & Weiss, W. (1998). *RFC 2475: An Architecture for Differentiated Services.*

Bless, R. (2002). *Dynamic Aggregation of Reservations for Internet Services*. In International Conference on Telecommunication Systems – Modeling and Analysis (ICTSM (pp. 26–38).

Braden, R., Clark, D. & Shenker, S. (1994). *RFC 1633: Integrated Services in the Internet Architecture: an Overview.*

Braden, R. & Zhang, L. (1997). *RFC 2209: Resource ReSerVation Protocol (RSVP) — Version 1 Message Processing Rules.*

Braden, R., ed., Zhang, L., Berson, S., Herzog, S. & Jamin, S. (1997). *RFC 2205: Resource ReSerVation Protocol (RSVP) — Version 1 Functional Specification*

Chang, M., Lee, M. & Lee, H. (2005). *An Efficient Resource Reservation Scheme Based on Gray-Cell in Wireless Mobile Networks*. In IEEE Wireless Communications and Networking Conference (WCNC), Volume 3 (pp. 1311–1316).

Chaskar, H., ed., (2003). *RFC 3583: Requirements of a Quality of Service (QoS) Solution for Mobile IP.*

Chen, W.-T. & Huang, L.-C. (2000). *RSVP Mobility Support: A Signalling Protocol for Integrated Services Internet with Mobile Hosts.* In IEEE/ACM INFOCOM (pp. 1283–1292). Tel Aviv, Israel.

Gundavelli, S., ed. Devarapalli, V., Chowdhury, K. & Patil, B. (2008). *RFC 5213: Proxy Mobile IPv6.*

Hancock, R., Karagiannis, G., Joughney, J. & Van den Bosch, S. (2005). *RFC 4080: Next Steps in Signaling (NSIS): Framework.*

Huang, N.-F. & Chen, W.-N. (2003). *RSVP Extensions for Real-Time Services in Hierarchical Mobile IPv6*. Kluwer Mobile Networking and Applications (MONET), 8, 625–634.

Johnson, D., Perkins, C. & Arkko, J. (2004). *RFC 3775: Mobility Support in IPv6.*

Kaloxylos, A., Vali, D., Paskalis, S., Panagiotou, G., Gonianakis, G. & Zervas, E. (2006). *Mobility Support for a QoS Aggregation Protocol*. In International Symposium on Communication Systems, Networks and Digital System Processing (CSNDSP). Patra, Greece.

Kempf, ed., J. (2002). *RFC 3374: Problem Description: Reasons For Performing Context Transfers Between Nodes in an IP Access Network.*

Koodli, ed., R. (2008). *RFC 5268: Mobile IPv6 Fast Handovers.*

Kuo, G.-S. & Ko, P.-C. (2000). *Dynamic RSVP for Mobile IPv6 in Wireless Networks*. In IEEE Vehicular Technology Conference (VTC) Spring, Volume 1 (pp. 455–459). Tokyo, Japan.

Lee, K., Kim, M., Chanson, S., Yu, C. & Lee, J. (2003). *CORP – A Method of Concatenation and Optimization for Resource Reservation in Mobile Internet*. IEICE Transactions on Communications, Special Issue on Internet Technology, E86-B(2), 479–489.

Manner, J. & Raatikainen, K. (2003). *Localized QoS Management for Multimedia Applications in Wireless Access Networks.* In IASTED Internet and Multimedia Systems and Applications (IMSA) (pp. 193–200). Honolulu, HA, USA.

Manner, ed., J., Karagiannis, G. & McDonald, A. (2009). *NSLP for Quality of Service signalling.* Internet Draft, work in progress. draft-ietf-nsis-qos-nslp-10.txt.

Moon, B. & Aghvami, H. (2001*). RSVP Extensions for Real-Time Services in Wireless Mobile Networks*. IEEE Communications Magazine, 39(12), 52–59.

Moon, B. & Aghvami, H. (2004). *Quality of Service Mechanisms in all-IP Wireless Access Networks.* IEEE Journal on Selected Areas in Communications, 22(5), 93–99.

Next Steps In Signaling IETF Working Group. http://www.ietf.org/html.charters/nsis-charter.html.

Pan, P., Schulzrinne, H. & Hahne, E. (2000). *BGRP: A Tree-Based Aggregation Protocol for inter-Domain Reservations*. Journal of Communications and Networks, 2(2), 157–167.

Perkins, C. (1996). *RFC 2003: IP Encapsulation within IP*.

Perkins, C., ed., (2002). *RFC 3344: IP Mobility Support for IPv4*.

Rosen, E., Viswanathan, A. & Callon, R. (2001). *RFC 3031: Multiprotocol Label Switching Architecture.*

Sanda T., E., Fu, X., Jeong, S., Manner, J. & Tshofenig, H. (2009). *Applicability Statement of NSIS Protocols in Mobile Environments*. Internet Draft, work in progress. draft-ietf-nsis-applicability-mobility-signaling-12.txt.

Schulzrinne, H. & Hancock, R. (2009). *GIST: General Internet Signalling Transport*. Internet Draft, work in progress. draft-ietf-nsis-ntlp-19.txt.

Shen, Q., Lo, A., Seah, W. & Ko, C.-C. (2000). *On Providing Flow Transparent Mobility Support for IPv6-based Wireless Real-Time Services.* In IEEE International Workshop on Mobile Multimedia Communications (MoMuC) (pp. 2B41–2b46). Tokyo, Japan.

Sofia, R., Guerin, R. & Veiga, P. (2003). *SICAP, A Shared-segment Inter-domain Control Aggregation Protocol*. InWorkshop on High Performance Switching and Routing (HPSR) (pp. 73–78).

Soliman, H., Castellucia, C., Malki, K. E. & Bellier, L. (2005). *RFC 4140: Hierarchical Mobile IPv6 Mobility Management (HMIPv6).*

Talukdar, A., Badrinath, B. & Acrarya, A. (2001). *MRSVP: A Resource Reservation Protocol for an Integrated Services Network with Mobile Hosts.* Kluwer Wireless Networks (WINET), 7(1), 5–19.

Terzis, A., Wang, L., Ogawa, J. & Zhang, L. (1999). *A Two-Tier Resource Management Model for the Internet.* In Global Internet.

Thomas, M. (2002). *Analysis of Mobile IP and RSVP Interactions*. Internet Draft, work in progress. draft-thomas-nsis-rsvp-analysis-00.txt.

Tseng, C.-C., Lee, G.-C., Liu, R.-S. & Wang, T.-P. (2003). *HMRSVP: A Hierarchical Mobile RSVP* Protocol. Kluwer Wireless Networks (WINET), 9(2), 95–102.

Vali, D., Paskalis, S., Kaloxylos, A. & Merakos, L. (2004). *A Survey on QoS Signaling*. IEEE Communications Surveys and Tutorials, 6(4).