ELSEVIER

Computer Networks 38 (2002) 137–163

COMPUTER NETWORKS

www.elsevier.com/locate/comnet

Evaluation of mobility and quality of service interaction

Jukka Manner [a,*], Alberto Lopéz Toledo [b], Andrej Mihailovic [c], Héctor L. Velayos Muñoz [d], Eleanor Hepworth [e], Youssef Khouaja [f]

[a] *Department of Computer Science, University of Helsinki, Teollisuuskatu 23, 00510 Helsinki, Finland*
[b] *Department of Communications and Information Engineering, Facultad de Informática, University of Murcia, CP 30071, Murcia, Spain*
[c] *Centre for Telecommunications Research, King's College London, Strand, London, UK WC2R 2LS*
[d] *Department of Microelectronics and Information Technology, KTH, Royal Institute of Technology, Electrum 229, 164 40 Kista, Sweden*
[e] *Internet Systems, Roke Manor Research Limited, Roke Manor, Romsey, Hampshire, UK SO51 0ZN*
[f] *FTR&D/DMR/DDH, France Télécom R&D, 4, rue du Clos Courtel, BP 59, 35512 Cesson Sevigné, France*

Received 6 February 2001; received in revised form 21 June 2001; accepted 1 August 2001

Responsible Editor: G. Morabito

Abstract

With the fast adoption of IP-based communications for mobile computing, users are expecting a similar service in wireless and wired networks. This raises the need for setting guarantees to the quality of the offered service (QoS), despite the technology of the access network (AN) or the mobility of the terminal. As mobile computing is getting more popular on a daily basis, new broadband cellular wireless ANs will appear with overlapping coverage in hot spots. This generates a new challenge for QoS provision, as it will have to deal with fast mobility of terminals. Various QoS architectures have been defined, but none provides full support for guaranteed service levels for mobile hosts (MHs). This paper discusses the problems related to providing QoS to MHs and identifies the existing solutions and future work needed.

Keywords: IP access networks; IP mobility; Micro-mobility; IP QoS; IntServ; DiffServ

1. Introduction

The telecom world is moving towards an "all-IP" network. The fast adoption of IP-based communications for hand-held devices equipped with wireless interfaces is creating new challenges for the Internet evolution. Users expect flexible access to Internet based services, including not only traditional data services but also multimedia applications. Multimedia applications and the generated audio–video streams need a constant circuit-switched-like guaranteed connection. Enabling circuit-switched-like service in a packet-switched network requires some sort of support for the service quality. The quality of service (QoS) is usually understood to mean fast, predictable and loss-free forwarding of data packets.

The use of IP facilitates the design of applications and services independent of the environment

* Corresponding author. Tel.: +358-9-19144210; fax: +358-9-19144441.

E-mail addresses: jmanner@cs.helsinki.fi (J. Manner), alberto@dif.um.es (A. Lopéz Toledo), andrej.mihailovic@kcl.ac.uk (A. Mihailovic), hvelayos@it.kth.se (H.L. Velayos Muñoz), eleanor.hepworth@roke.co.uk (E. Hepworth), youssef.khouaja@francetelecom.com (Y. Khouaja).

1389-1286/02/$ - see front matter
PII: S 1 3 8 9 - 1 2 8 6 (0 1) 0 0 2 5 5 - 9

in which they will operate, be it a fixed or wireless network. The emerging wireless access networks (ANs) and third generation cellular systems constitute the enabling technology for "always-on" personal devices. IP, traditionally developed by the Internet Engineering Task Force (IETF), have mainly been designed for fixed networks. Their behaviour and performance are often affected when deployed over wireless networks. For example, the protocols for supporting mobility and QoS have until recently been worked on separately from each other. Thus, when support for service differentiation is sought in a mobile environment, various enhancements are needed. Therefore, the interaction of these protocols is currently being reviewed.

The telecom world has created various systems for enabling wireless access to the Internet. Systems such as the General Packet Radio Service (GPRS), Enhanced Data Rate for GSM Evolution (EDGE), Universal Mobile Telecommunications System (UMTS) and International Mobile Telecommunications (IMT-2000) are able to carry IP packets using a packet switching network parallel to the voice network. These architectures use proprietary protocols for traffic management, routing, authorisation or accounting, to enumerate some, and are governed by licenses and expensive system costs. Wireless LANs offer higher speeds and more flexible and cost-effective deployment, not to mention easier integration with the IP world, which has driven many to consider that these networks will provide the real mobile Internet.

From the QoS point of view, the problems with mobility in a wireless AN and mobility-related routing schemes are related to providing the requested service even if the mobile node (MN) changes its point of attachment to the network. Handovers between access points, change of IP addresses, and mechanisms for the intra-domain micro-mobility mechanisms may create situations where the service assured to the MN cannot be provided, and a violation of the assured QoS may occur. A QoS violation may result from excess delays during handovers, packet losses, or even total denial of service. In the case where the user only requested differentiation according to a relative priority to flows, a short QoS violation may fit within acceptable limits. If the flows were allocated explicit resources, the new network access point and route from the domain edge should provide the same resources.

Several research projects within the academic community, e.g. INSIGNIA [28], and in the industrial community, e.g. ITSUMO [13], have sought to combine mobility with guaranteed QoS. In the BRAIN project [11], we are envisioning an all IP network, where seamless access to Internet based services is provided to users. By using IETF protocols, we are designing a system that would be able to deliver high-bandwidth real-time multimedia independent of the wireless AN or the wireless technology used to connect the user to Internet. This implies the need for IP mobility support and also end-to-end QoS enabled transport. The provision of QoS guarantees over heterogeneous wireless networks is a challenging issue; especially because overprovisioning is not always possible and the performance of the wireless link is highly variable. We focus our architecture on wireless LAN networks, since these provide high bandwidths but may also create frequent handoffs due to fast moving users—this type of architecture is most demanding in view of mobility management and QoS.

This paper presents the evaluation of mobility and QoS interactions in the given framework. The discussion is structured into four main sections. In Section 2 we study the existing QoS architectures, followed by a presentation of mobility management protocols in Section 3. Section 4 discusses the problems that emerge when provision of QoS is deployed for mobile hosts (MHs). Section 5 presents some existing solutions and identifies the future work needed in this area. Finally, in Section 6 we present concluding remarks.

2. Quality of service management

In this section we present short overviews of existing IETF-presented architectures for providing different levels of services to IP flows. Also the ITSUMO architecture is presented as it is based on the differentiated services (DiffServ) framework and seems a mature proposal.

The IETF architectures can be classified into three types according to their fundamental operation; the integrated services (IntServ) framework provides explicit reservations end to end; the DiffServ architecture offers hop-by-hop differentiated treatment of packets; and the real-time transport protocol (RTP) provides mechanisms for flow adaptation and control above the transport layer.

2.1. Integrated services and RSVP

The IntServ model [5] merges the advantages of two different paradigms: datagram networks and circuit switched networks. It can provide a kind of circuit-switched service in packet-switched networks. IntServ is used to specify characteristics of the transmitted flow and to specify the requested service from the connecting network.

The resource reservation protocol (RSVP) [12] was designed as the primary signalling protocol for the provision of QoS in an IP network. The IntServ model uses RSVP to propagate the attributes of the data flow and to request specific resources along the data path [56]. IntServ and RSVP provide unidirectional resource reservations on a per-flow basis. The sender of a flow first sends a PATH message to the receiver. The message is updated at every router on the path. The receiver responds with RESV message and indicates the resources needed at every hop to support the forthcoming flow. Any router on the end-to-end path may deny the flow if resources are scarce. If the sender receives the RESV message the resources for supporting the flow requirements have been granted.

IntServ identifies three main categories of services that can be provided to users. Guaranteed services [50] (i) provide users with an assured amount of bandwidth, firm end-to-end delay bounds, and no queuing loss for flows. Controlled load [57] (ii) services assure that the users will get service that is as close as possible to the one received by a best-effort service in a lightly loaded network. Best effort services (iii) are characterised by absence of a QoS specification and the network delivers the best possible quality.

2.2. Differentiated services

While IntServ provides per-flow guarantees, DiffServ [2,3] follows the philosophy of mapping multiple flows into a few service levels—an approach sometimes referred to as class of service (CoS). DiffServ are constructed by a combination of: (i) marking packets with a DiffServ code point (DSCP) at boundary nodes, (ii) using the DSCP to determine how packets are forwarded by the nodes inside the domain, and (iii) conditioning the marked packets at boundary nodes.

In IPv4, the DSCP is marked in the 8-bit type of service field in the IP header. DiffServ is realised by mapping the DSCP contained in the IP packet header to a particular treatment or per-hop behaviour (PHB), at each network node along the path of the packet. There are various PHBs being defined in the IETF such as expedited forwarding [25] or premium service and assured forwarding [8] or quality-level based forwarding. Service level agreements (SLA) specify bilateral service levels between domain boundaries.

Lately there has been initiative in IETF to specify formal per-domain behaviours (PDBs) [36], to clear this apparent chaos. In contrast to the more abstract SLA concept, PDB would be a technical building block coupling rules, specific PHBs, and configurations with a resulting set of observable characteristics.

DiffServ performs aggregate classification of packets in contrast to IntServ, which provides a per-flow classification. The aggregation results in more scalable but also more approximate service to user flows and the lack of control signalling can be seen as a weakness in view of the total operation.

2.3. Integrated service over differentiated services

IntServ, RSVP and DiffServ can be seen as complementary technologies in the pursuit of end-to-end QoS. There are a number of 'work in progress' efforts, which are directed towards these aggregated control models. These include aggregation of RSVP [9], the RSVP DCLASS Object [6] to allow DSCPs to be carried in RSVP message objects, and the operation of IntServ over DiffServ networks [7,55] proposed by the Integrated Services over Specific Link Layer (ISSLL) Working group.

The architecture proposed by the ISSLL provides a reservation-based QoS architecture with feedback signalling about the state of the network. The architecture uses RSVP to signal resource needs but uses DiffServ as the technology to do the actual resource sharing among flows. The reference architecture includes a DiffServ region in the middle of two IntServ regions. This model does not fix the sizes of the different regions and their structure. At the other extreme, the IntServ regions could be only the sending and receiving nodes themselves (and possibly the closest router), while all routers between these two are DiffServ enabled. Basically, the more DiffServ routers we have, the more scalable the service is.

The basic requirements and assumptions are that the resource signalling is done with RSVP and that we have a mapping at the border nodes for RSVP-based reservations to DSCP values. Depending on the scenarios, routers within the DiffServ region may be able to produce RSVP messages, even though the forwarding operation is purely based on the DSCPs. This would allow for more accurate resource co-ordination within the DiffServ domain. Also a SLA between the non-DiffServ regions and the DiffServ region is needed. The SLA defines the capacities of the DiffServ region, the resource types and capacities for each type of RSVP-based reservation.

The primary benefit of combining IntServ and DiffServ is the increased scalability, provided through the aggregate traffic control of DiffServ.

2.4. Real-time transport protocol

The RTP [48] provides end-to-end delivery services, such as payload type identification, timestamping and sequence numbering, for data with real-time characteristics, e.g. interactive audio and video. It can be used over unicast or multicast networks. RTP is run "on top" of a transport protocol such as UDP. RTP does not reserve resource on the end-to-end path, but rather tries to adapt to prevailing network conditions.

RTP usually works in conjunction with a control protocol, the real-time control protocol (RTCP), which provides minimal control over the delivery and quality of the data. RTCP provides support for real-time conferencing of groups of any size within an Internet. This support includes source identification and support for gateways like audio and video bridges as well as multicast-to-unicast translators. It offers quality-of-service feedback from receivers to the multicast group as well as support for the synchronisation of different media streams. The feedback mechanism allows RTP to adapt to current network conditions.

2.5. ITSUMO

The ITSUMO approach [13] presents a QoS architecture framework following a bandwidth broker like scheme. The architecture is based on DiffServ in that traffic is aggregated and forwarded in backbone network based on PHBs. In the proposed architecture, there is at least one global server and several local nodes in each administration domain. The server is referred to as the QoS global server (or QGS), and local nodes are referred to as QoS local nodes (or QLN). QLNs are ingress nodes of the DiffServ domain, and they reside generally in the edge of wired backbone networks. The QGS retains the global information of the domain, and informs QLNs what to do when traffic comes in. The MN communicates its QoS requirements directly to the QGS, for example through the use of session initiation protocol (SIP) messages [23]. Once the MN has had such a request accepted it is guaranteed (within SLAs), it can move within the domain and receive the required QoS. The QGS server has a near-complete picture of the state of the network at any time—it achieves this by regular polling of all the QLNs. It uses this knowledge to determine if a particular request can be supported. Once it has determined this, it broadcasts this decision to all nodes likely to be affected by the MN. Mobility guarantees are made by notifying QLNs of MNs likely to handover into their cells.

The service level specification (SLS) is usually agreed by both the user and the service provider when the user signs up with a service provider. To change the SLS in wired network, a user has to contact the service provider. Once the negotiation is done, the user can utilise the new SLS. Once the negotiation between the MS and the QGS is done, the QGS multicasts the decision to all QLNs in the same administration domain. The MS therefore is capable of utilising the new SLS anywhere while it is moving within the same administration domain. Thus, dynamic SLS for mobile environment is achieved with only one negotiation in the same administration domain.

The ITSUMO approach offers classes of services based mainly in the combination of two parameters: *latency* and *loss*. For each parameter possible values are described: High, moderate, low for latency, and high, moderate, low and none for loss. The combination of the two parameters forms a spectrum with 12 classes of services as showed in Fig. 1.

One disadvantage of this approach is that it assumes global domain knowledge, which is difficult to maintain and manage. One article [21] suggests that trying to do admission control based on anything other than just bandwidth requirements for real-time service is intractable in such a situation, yet the ITSUMO approach seems to consider a range of delay and loss parameters.

2.6. Discussion

None of these architectures is able to provide both scalability to a large number of flows and

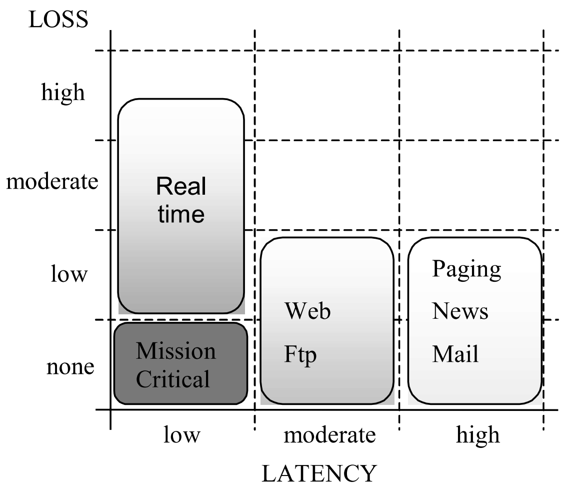

Fig. 1. ITSUMO classes of service.

strict service guarantees. The most interesting architecture to study in the future is the framework for IntServ operation over DiffServ networks. This scheme provides both signalled and assured service while being able to scale to large networks. A bandwidth broker in this architecture would enhance the overall service given to flows. The Internet Architecture Board has noted the issues related to the present architectures and has given some instructions for future work [24] including the combining of the mentioned architectures.

3. Mobility management

In this section we identify some of the fundamental issues related to mobility management in order to define a perspective on the best solutions, which can be then applied in our further study. It is considered that this topic will be closely related to QoS development, since both mobility and QoS protocols are expected to have awareness of certain, if not all, of their functionality. The following sections address the issues for facilitating the selection of the most promising candidates for mobility management and introduce a categorisation for distinguishing protocols and their associated purposes. The focus of this section is placed on an analytical method we call the evaluation framework, which has been adopted for facilitating detailed analysis and comparative evaluation of mobility protocols. The framework assists us in the identification of key protocol features for integration with QoS, such as handover management dealt separately in Section 3.4. It also helps us to select our preliminary choices of mobility protocols based on the most efficient protocol classes.

In the following discussion, the term MN is used to refer to a MH or mobile router. If MH is used, the term mobile router does not apply, and vice versa.

3.1. Evaluation framework

The evaluation framework is presented in more details in Ref. [19] and contains an extensive model for analysing and evaluating mobility protocols. In this section we give a short description of the framework, which consists of three fundamental parts:

(a) *Identification of the essential components of mobility protocols*: This is the breakdown of all functions the protocols are designed to tackle and are called protocol-design issues (PDIs).
(b) *Classification of mobility protocols*: This part contains an extensive survey of all current mobility protocols and attempts to classify them based on their purposes and protocol mechanisms. The protocol mechanisms are, in accordance with part (a), named PDI-solution.
(c) *Evaluation of mobility protocols*: This is the final part of the framework and it uses results of the previous two parts, (a) and (b), to create propositions for the development of optimum mobility protocol(s). The essential element of this task is the application of the evaluation criteria consisting of a set of requirements against which effectiveness of a particular PDI-solution can be assessed.

Here we list the PDIs, along with a short explanation of each:

Packet forwarding: Packet forwarding refers to the delivery of packets to and from the MN modified in order to cope with host mobility. Typically, the solution is based on host routes, with or without tunnelling.

Path updates: This refers to the mechanism for installing information in the fixed network so that packets can be successfully forwarded to the MN at its new point of attachment. It consists of the intelligent transmission of specific *update* messages dependent on the particular protocol.

Handover management: This PDI looks at the impact of handovers on the MN (whereas the previous issue took a network-centric view). Handover management is discussed into more details in Section 3.4.

Support for idle mobile hosts: Paging reduces the frequency of refreshments/updates for an idle MH in order to achieve two goals: reduce the protocol overhead (signalling, route lookups and memory requirements) in the network and minimise a MH's power consumption.

Address management: The way in which a MN is assigned an IP address in a foreign network can have an important impact on various parameters, which affect the overall efficiency of the protocols.

Routing topology: This refers to a general static view of the AN nodes, whilst the other issues above more or less cover dynamic protocol operation. The routing topology has implications on the scalability and robustness of the system and also relates to the reaction upon any failure of links or routers.

Security issues: The user's access to a visited network need to be authorised and the requests for path changes have to be authenticated; the user's privacy should be preserved; the AN's topology should be hidden from MNs; and inter-working of IPv6 Security protocol (IPSec) is required. The majority of IP-mobility schemes include security features or a framework for their realisation.

Requirements for mobile nodes: An important decision is to what extent MNs are required to participate in the establishment and updating of the routing structure that enables mobility.

Requirements for core network interface: This issue defines the functionality in the gateway router of the AN. This is the transition point between the micro- and macro-mobility (see the remaining parts of this section for the explanation of macro- and micro-mobility) and can include functions such as inter-working between micro- and macro-mobility, mapping of addresses, tunnel management, central control of mobility protocol mechanisms.

Classification of mobility protocol can be achieved regarding many of their characteristics. It can be assumed that they share a common goal of overcoming the location dependent nature of IP addresses of Internet hosts by developing mechanisms for translation of addresses and efficient distribution of packets to and from any location both for static and highly MHs. However, the evolution of the protocols has resulted in a first-glance distinction, concerning their *scope* as the dominant parameter. Therefore the current mobility protocols can be classified into two main categories: global or macro mobility protocols and micro- or regional-mobility protocols explained in more details in Sections 3.2 and 3.3, respectively.

Evaluation of mobility protocols should conclude with a recommendation for the optimum PDI-solutions either by promoting the existent models or by proposing guidelines for the design of new ones. The evaluation criteria is used for achieving this task and consists of subgroups of requirements clustered in three main categories: *efficiency*, *scalability/robustness and applicability/ ease of deployment*. Applying the requirements identified by the evaluation criteria performs the comparative analysis of mobility protocols.

3.2. Macro-mobility

The mobile IP [41] protocol is the current standard for supporting macroscopic mobility in IP networks i.e. host mobility across IP domains while maintaining transport level connections. It is transparent for applications and transport protocols, which work equal with fixed or MHs. It can be scaled to provide mobility across the Internet. And it allows nodes using mobile IP interoperate with nodes using the standard IP. There are two versions of mobile IP: mobile IPv4 and mobile IPv6. Each one addresses a particular version of IP.

3.2.1. Introduction to mobile IPv4

The mobile IPv4 protocol was designed to provide a near-term solution for MNs without requiring protocol upgrades in stationary correspondent nodes or routers. In this protocol, there are three functional entities: the MN, the home agent and the foreign agent (FA). The MN is configured with a permanent IP address belonging to its home network. It is called the home address. All packets sent to the MN are addressed to its home address. The home agent is a router in the home network of the MN. It is continuously aware of the MN current location. The FA is a router in the visited network. It is generally used by the MN to obtain a new temporary address and to register with the home agent. The home and the FA are called mobility agents.

Mobile IP performs as follows: when the MN is connected with a network, it listens to mobility agent advertisements broadcast by mobility agents. If the network prefix changes, the MN detects a movement. It is now located in a visited network and tries to acquire a new temporary address. This new address can either be obtained by an auto-configuration mechanism like DHCP or be the actual FA address. The former is called co-located care-of address (CCOA) and the latter is called foreign agent care-of address (FA-COA). If CCOA is acquired, the MN registers this new temporary address with the home agent by exchanging registration requests and responses using CCOA as source address. If FA-COA is acquired, the MN cannot register itself using its home address. In this case, the FA will relay the registration to the home agent. Once registration finishes, home agent intercepts packets sent to the MN and uses IP-in-IP encapsulation to tunnel them to the new temporary address. The home agent must also answer to address resolution protocol [43] requests for MN hardware address with its own hardware address to intercept packets destined to the MN. If the MN uses a FA-COA address, the corresponding FA decapsulates the packets and delivers them to the MN. Otherwise, the MN decapsulates its packets itself, as it is directly reachable using the CCOA address.

To maintain the registration, the MN has to periodically renew its registration. When the MN returns to its home network, it has to remove its current registrations. This will result in home agent stopping to intercept the MN traffic.

As the home agent intercepts all packets addressed to the MN and tunnels them to the visited network, a "triangle routing" effect is produced (Fig. 2, left). All packets must first pass through the home agent even if the current access router (AR) is in the same network as the correspondent node. An extension to mobile IP, known as route optimization [42], has been proposed to overcome this problem (Fig. 2, right). It allows data packets to be routed directly from the correspondent node to the MN using a binding cache in the correspondent node that keeps track of the current temporary address. These binding caches are created and updated by binding update (BU)

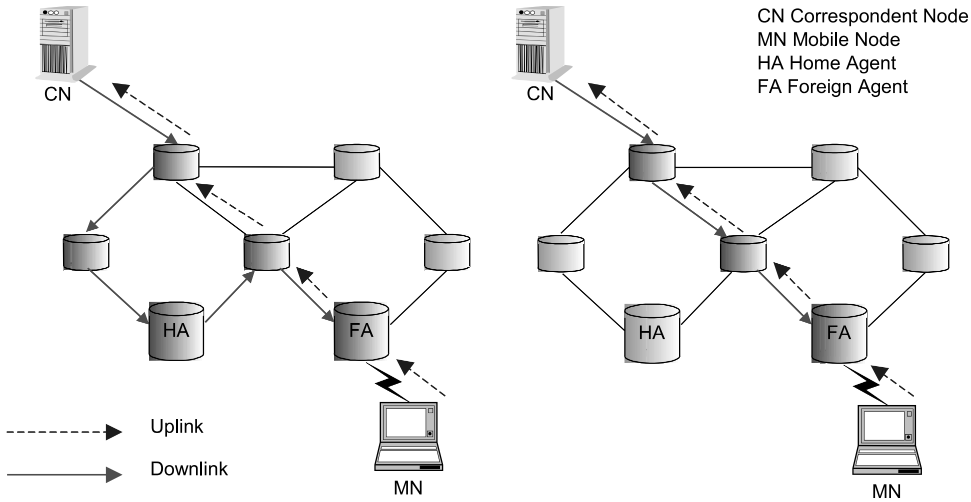

Fig. 2. Illustration of triangle routing and route optimisation.

messages sent by the home agent in response to MN warnings or correspondent node requests.

In mobile IP, the MN has to use its home address as a source IP address so that the current connections are not interrupted. On the other hand, it sends its packets through a router on the visited network, and assumes that routing is independent from source address. Nevertheless, due to security concerns, the use of routers that perform ingress filtering break this assumption and impose on the MN to use a topologically correct source IP address in its emitted packets. Consequently, an extension to mobile IP, known as reverse tunnelling [32], has been proposed to establish a topologically correct reverse tunnel from the FA to the home agent. Sent packets are then decapsulated at the home agent and delivered to correspondent nodes with the home address as IP source address.

3.2.2. Comparison of mobile IPv4 with mobile IPv6

The design of mobile IP support in IPv6 [26] is based on the experiences gained from the development of mobile IP support in IPv4, and the opportunities provided by the new features of the IP such as an increased number of available IP addresses and additional automatic IP-configuration features.

Firstly, in mobile IPv6, the route optimisation process is integrated in the protocol. In fact, the route optimisation and the registration procedure with home agent are both done by new defined BUs. These new BUs use the integrated IPSec for sender authentication, data integrity protection, and replay protection.

Furthermore, mobile IPv6, and IPv6 itself, allows MNs and mobile IP to coexist efficiently with routers performing ingress filtering, as the MN uses its temporary address as the source address. The home address of the MN is indicated in a home address destination option of the IP packet.

Also the use of IPv6 destination options, that carry optional information only addressed to the destination, allows all mobile IPv6 control traffic to be piggybacked on any existing IPv6 packet, whereas in mobile IPv4 and its route optimisation extensions, separate UDP packets were required for each control message.

Finally, in mobile IPv6, there is no longer any need to deploy special FAs. MNs make use of the enhanced features of IPv6 [35] to operate in any location away from the home network without any special support required from its local router.

3.3. Micro-mobility

For the support of regional-mobility within one domain or one site, the mobile IP solution was found non-optimal. Firstly, it generates significant signalling traffic in the core network even for local movement. Secondly, it creates a considerable delay in the diffusion of MNs localisation updates. And finally, it causes long interruptions and packet losses during handovers. Therefore a new protocol providing the management of micro-mobility seems to be necessary.

Due to the large number of micro-mobility protocols it is not possible to present a detailed description of all candidate schemes. In order to more easily observe the key features of different micro-mobility protocols we have introduced a classification, which categorises the protocols regarding their protocol mechanism (Fig. 3). The already applied classification into macro- and micro-mobility protocols justifies the assumption that all micro-mobility protocols are designed to address the same mobility scenarios therefore the categorisation based on their mechanisms is regarded suitable. Emphasising the protocol mechanisms is expected to result in obtaining a useful conclusion about the most optimal solution, which we plan to propose for our further work. The two major categories of *regional-mobility* protocols are:

- Proxy-agent architectures (PAA),
- Localised enhanced-routing schemes (LERS).

Proxy agents architecture schemes: These schemes extend the idea of mobile IP into a hierarchy of MA (which are extensions of mobility independent predictive's (MIP)'s FAs and/or HAs). A MN registers with its local agent ('a') at the bottom level of the hierarchy ("MN is at care-of-address (CoA)"), which in turn registers with its nearest agent at the next hierarchy-level ("MN is at agent a"), and so on up the hierarchy towards the HA. This way, when the MN changes its CoA, the registration request does not have to travel up to the HA but remains 'regionalised'. Packets from a CN travel down the hierarchy, being tunnelled from one level to the next.

Examples include the initial hierarchical mobile IP [40] and its alternatives, which place and

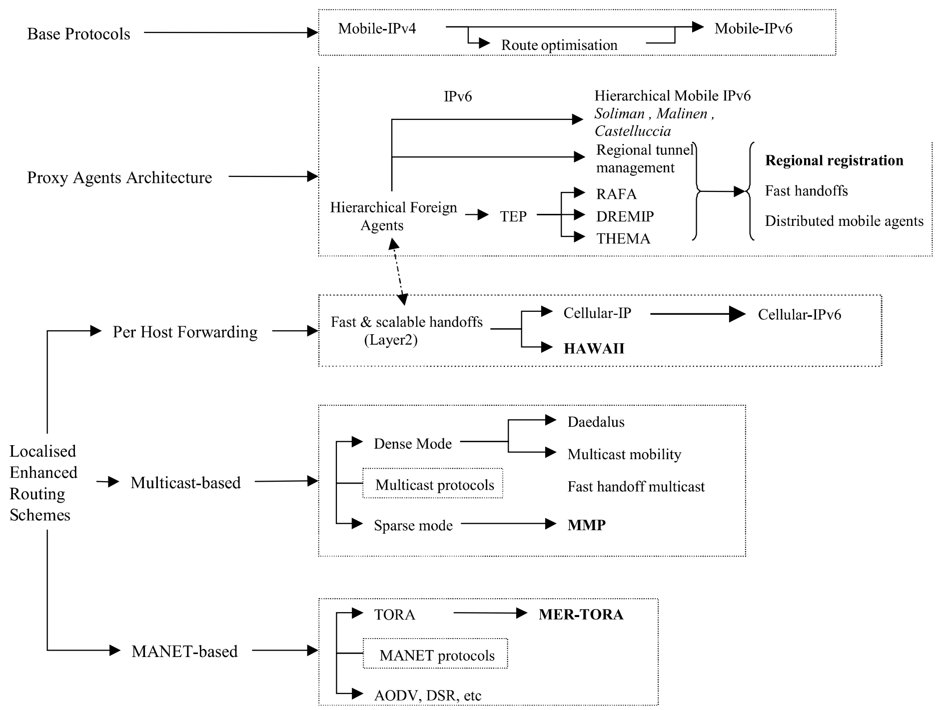

Fig. 3. Classification of mobility protocols.

interconnect mobility agents more efficiently: mobile IP regional registration [20], transparent hierarchical mobility agents (THEMA) [31], fast handoff in mobile IPv4 [17].

The new mobile IP version 6 [26] has had some optional extensions by applying a hierarchical model where a border router acts as a proxy home agent for the mobile nodes. The common goal in the proposed scheme is to reduce the latency and load of BU signalling for a MN moving within a visited domain. Thus these schemes can be classified as new PAA for IP version 6. They include "Hierarchical MIPv6 mobility management" [47] and "Mobile IPv6 regional registrations" [33].

Localised enhanced-routing schemes: These schemes introduce a new, dynamic layer 3 routing protocol in a 'localised' area. There are several distinctive approaches:

- *Per host forwarding schemes*: Inside a domain, a specialised path set-up protocol is used to install soft-state host-specific forwarding entries for each MN. The domain, which appears as a subnet to routers outside the domain, is connected to the Internet via a special gateway, which must be pointed to by the default gateway of the routers (or packet forwarding nodes) inside the domain. Examples include handoff-aware wireless access internet infrastructure (HAWAII) [44] and recently cellular IPv6 [49], which includes IPv6 enhancements to the original cellular IP [14] protocol and a technique for indirect handoffs.
- *Multicast-based schemes*: Multicast protocols are designed to support point-to-multipoint connections. So they share with IP mobility the same design goals of location independent addressing and routing and thus multicast-based mobility solutions have been proposed. A multicast-CoA is assigned to a single MN, which can then be used to instruct neighbouring multicast-enabled routers to join the MN's virtual multicast group, either prior to or during handovers. This can be visualised as a multicast cloud centred on the MN's current location but also covering future locations. Examples include dense mode multicast-based [29,46,53] and the recent sparse-mode multicast-based [34].
- *MANET-based schemes*: MANET protocols were originally designed for mobile adhoc networks, where both hosts and routers are mobile, i.e. there is no fixed infrastructure. The routing is multihop and adapts as the MNs move and connectivity in the network changes. MANET protocols can be modified for our scenario, where there is a fixed infrastructure and only hosts can be mobile. Currently there is only one proposal in this category: MER-TORA [38,39].

Fig. 3 shows some of the many IP mobility protocols, which category they fall into and very roughly how they relate to each other.

3.4. Handover management

This section expands on the handover management PDI identified in Section 3.1. It is considered one of the most important features of the mobility protocols when considering the interaction with QoS protocol because of the likely renegotiation of QoS parameters. Handover refers in general to support for terminal mobility wherever the MN changes its point of attachment to the network. More specifically, the AN may provide particular capabilities to minimise the interruption to sessions in progress.

In wireless mobile networks different handover scenarios might occur. A Layer 2 handover happens if the network layer is not involved in the handover, intra-access network [1] handover when the new point of attachment is in the same AN, inter-access network handover when the new AR is in a different AN. Horizontal or vertical handover are said to happen if the old and the new AR [2] use the same or different wireless interface (technology) respectively. Others handover types

[1] AN: an IP network, which includes one or more ARs and gateways.

[2] AR: an IP router residing on the edge of an access network and offering IP connectivity to MHs connected to access links, acting as a default router to the MHs it is currently serving. The AR may include intelligence beyond a simple forwarding service offered by ordinary IP routers.

can be defined according to different phases of the handover.

Three phases are distinguished in the handover.

Initiation phase

The objective of this phase is to recognise the need for a handover and subsequently initiate it. The handover can be required by the MN or by the network. Generally, it is initiated when the radio link quality between a MN and its AR is degraded. However, it can also be initiated for network management and maintenance reasons. For example, in case of overload some MNs may be moved from an AR to another one.

Decision phase

In this phase, measurements on neighbouring radio transmitters and eventual network policy information are first collected. Then the best target AR is identified taken into account the measurement and information report. The execution phase is finally triggered to perform the corresponding handover.

According to whether the MN or the network handles these operations, four handover types are differentiated: *mobile* or *network controlled* handover if the MN or the network initiates and decides a handover. *Network-assisted* handover when the network collects information that can be used in handover decisions and *mobile-assisted* handover when information and measurements from the MN are used to decide the execution of a handover.

Execution phase

In the execution phase, the MN has been detached from the old AR and attached to the new one. The order of attach and detach events is not fixed. During a *soft* handover the MN communicates simultaneously with the old and the new AR whereas in a *hard* handover it is not able to do it.

Handover may imply re-routing of connections through the fixed network and an address negotiation for the MN like the acquisition of a new care-of-address and the registration procedure in mobile IP. In *planned* handover, contrary to *unplanned* handover, some signalling messages can be sent before the MN is connected to the new AR, e.g. building a temporary tunnel from the old AR to the new AR. If the handover is initiated via the currently serving AR, it is a *backward* handover, else it is a *forward* handover.

Specific actions may be performed depending on the handover phase. For example, the events may initiate upstream buffering or advance registration procedures at the MN. These mechanisms characterise furthermore the handover type: *smooth* handover is a handover with minimum packet loss, *fast* handover allows minimum packet delays and *seamless* handover that is a smooth and fast handover.

3.5. Discussion

Previous sections explain our approach for analysing and selecting mobility mechanisms suitable for possible integration with other Internet protocol, in particular QoS and related extensions. Our approach of separating the functions performed by mobility protocols into independent PDI creates a platform where we can more efficiently define critical features for integration with QoS and other protocols. We also acknowledge that is not entirely possible to separate the mobility protocols into independent PDI and that some interdependence is inevitable but regardless of it the conceptual differences are obvious. We have also highlighted the differences between macro- and micro-mobility protocols since they address separate aspects of mobility and should therefore be treated differently especially when considering integration of mobility and QoS or other protocols. We also propose a method for examining micro-mobility protocols based on the initial classification, which we consider essential for extracting logical patterns in the protocol mechanisms of micro-mobility protocols. Finally we recognise that some PDIs bear more importance in our further work of integrating mobility with other protocols, in particular handover management PDI, which was detailed in the previous section. Some of the topics mentioned in the remainder of this paper expand on the conclusion from our investigation in particular Section 5 where we make some propositions for the creation of optimal mobility protocols based on the results of our evaluation framework. Our intention is to further validate our choices through simulations.

4. Interaction of mobility and quality of service

This section discusses the problems related to guaranteeing service levels to MNs. We classify the problem areas into three groups, namely topology related problems, and macro- and micro-mobility related issues. Solutions to these problems are proposed in Section 5.

4.1. Depth of handovers

We can identify several types of handover situations between different entities, namely intra-AR, inter-AR and inter-ANG. [3] The same physical handover can create different logical handover situations to different MN flows if the flows travel through different AN gateways.

Fig. 4 illustrates possible handovers while a MN moves within and between two administrative domains. The different levels of handovers create variable load of signalling in the AN. In Fig. 4, the higher the handover propagates in the AN topology, the more time it will take to set routing and QoS allocations in place.

A handover between access points 1 and 2 would be considered intra-AR, a handover between access points 2 and 3 would be inter-AR, and a handover between access points 3 and 4 would create an inter-ANG handover. In addition, a handover between access points 4 and 5 would change the serving AN.

In handover situations RSVP has problems guaranteeing the reservations, because the reservation, routing and data transmission are independent phases. RSVP-aware nodes need to send periodic PATH and RESV messages for each flow to refresh the end-to-end reservations. When a route changes, packets will receive only best-effort service until the reservation state has been updated on the new path; the further away from the MN the handover is noticed, the more time it will take to rearrange the reservations. In order to shorten the period when there is no reservation, the refresh messages should be sent immediately after a handover to trigger the intermediate RSVP routers to update the location of the MN and thus do a "local repair".

DiffServ also assumes a fairly static routing and SLAs. Without a bandwidth broker to co-ordinate resource sharing in the DiffServ architecture, new MNs may disrupt the overall resource sharing between existing mobiles in their new cell.

In an intra-AR handover, the handover control only needs to handle radio resources since the flows will still use the same routing paths between the AR and ANG. Even the admission control part may be left out if the admission control has already been done with this AR when the mobile node initiated the transfers. This handover is often also called a Layer 2 handover. It may be transparent to the IP layer if the interface to the mobile does not change. Otherwise the handover triggers some changes internal to the AR.

If, in an inter-AR handover, the AR changes but the ANG remains the same due to similar routing, the handover affects the radio resource availability and the AN resources. In addition, the new AR may need to check for admission control at the same time.

The resource co-ordination due to mobility is much more affected by the change of ANGs. The ANG can change when the MN moves within the same larger AN or when the MN does a handover to another operator's AN. Note that this would have to involve the assignment of a new IP address to the MN. When the ANG changes, RSVP-based flows will experience a longer degradation in their QoS until a scheduled refresh message reaches a router which has the state of the reservation, and can thus initiate an update in the reservation states on the path.

The most complex handover happens when the administrative domain changes. Besides re-routing and QoS allocation handling, the mobile may need to be re-authenticated, authorisation to AN resources needs to be checked, and accounting records initialised. This results in heavy signalling, which is seen as a longer time interval before the MN's QoS allocations are back in place.

[3] Access network gateway (ANG): gateway offering the access network connectivity to the core IP network.

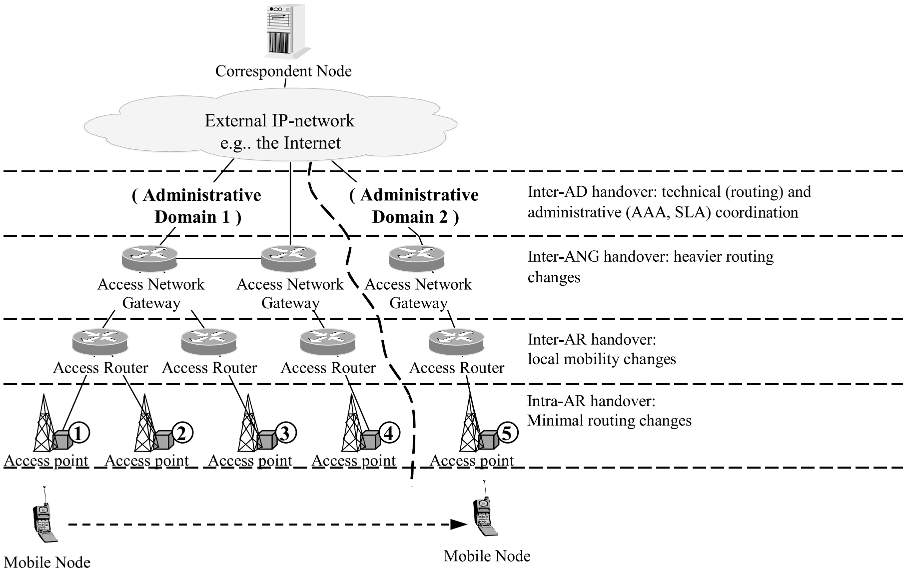

Fig. 4. Example network topology to illustrate different handover scenarios.

4.2. Macro-mobility issues

One of the main differences between fixed and mobile networks is that the MN will have to adapt to changes on the QoS perceived due to its mobility.

This section will focus on the effects that macro-mobility (mobile IP) has on QoS provision. The first problem arises directly from the already discussed *triangular routing* phenomenon. Packets from the MN usually follow a direct path to the CNs; packets from the CNs are re-routed via the MN's home network to its point of attachment in a foreign network, from where they are forwarded to the MN's current location. Several QoS architectures operate best when packets follow the same route in the forward and reverse direction. Triangular routing can affect the service level guarantees of these schemes.

In order to avoid this problem, it is possible to tunnel the upstream flow to follow the downstream using *reverse tunnelling* [32]. In this case, triangle routing is avoided but another problem arises. Routers in the tunnel may not be able to recognise some encapsulated parameters of the QoS protocols apart from IP addresses. For example, if RSVP packets use the *router alert option* to indicate to routers on the path that they require special handling, when RSVP messages are encapsulated with an outer IP header, the router alert becomes invisible. Although solutions to this have been proposed e.g. RSVP extensions to MHs [1], they still add complexity to the operation of QoS protocols on mobile environments.

Other main concern for QoS when the host is moving is the time needed to re-establish the routes, and hence, the time needed to re-configure resource management required to provided QoS in the new location.

Even in the case of using route optimisation, transmission of BUs directly to CNs result in a large update latency and disruption during handover. This effect is greatly increased if MN and HA or CN are separated by many hops in a wide area network. Location updates need to travel over the entire path from the MN to the HA/CN before the change in mobile location is effectively communicated and ongoing connections are restored. Data in transit will be lost until the handover completes and a new route to the MN is established. Route optimisation (as a protocol specification) however includes smooth handoff support using previous foreign agent notification extension, which can be used to avoid the described disruption.

Moreover, location updates are always generated whenever the MN changes a subnet in the foreign network. High mobility users may have frequent notifications to the home agent, resulting in a high control overhead. In situations with an extremely large population of MNs, the signalling load can become a significant portion of the traffic.

There are other problems related to address management. Since the current mobile IP standard requires the mobile to change the care-of address (either FA or co-located) at every subnet transition, it is harder to reserve network resources on an end-to-end path between the CN and the mobile. For example, if RSVP is used to make reservations for QoS for sensitive traffic, new reservations over the entire data path must be set up whenever the care-of address changes. The impact on the latency for re-establishment of the new routes as we have seen before is critical on QoS assurance.

Mobile IPv6

Mobility support in IP version 6 has already been discussed in Section 3.2. Some of the benefits included on this protocol directly affects the behaviour of the QoS protocols running over it.

Mobile IPv6 makes use of the new features provided by IPv6 protocol. They help to solve most of the problems discussed above which arise with the use of mobile IP in IPv4 networks. For example *route optimisation* is included in the protocol, and there are mechanisms for movement detection that allow a better performance during handover. The *routing header* avoids the use of encapsulation, reducing overhead and facilitating, for example, QoS provision.

Although the mobile IPv6 solution meets the goals of operational transparency and handover support, it is not optimised for managing seamless mobility in large cellular networks. Large numbers of location update messages are very likely to occur, and the latency involved in communicating these update messages to remote nodes make it unsuitable for supporting real-time applications on the Internet. These problems indicate the need for a new, more scalable architecture with support for uninterrupted operation of real-time applications.

4.3. Micro-mobility issues

The domain internal micro-mobility schemes may use different tunnelling mechanisms, multicast or adaptable routing algorithms. The domain internal movement of MNs affects different QoS architectures in different way. IntServ stores a state in each router; thus a moving mobile triggers local repair of routing and resource reservation within the network. DiffServ on the other hand has no signalling mechanism, which means that no state needs to be updated within the network, but the offered service level may vary. At least the following design decisions of a micro-mobility protocol need to be considered when combining mobility and QoS architectures within a network:

- the use of tunnelling hides the original packet information and hinders multi-field classification,
- changing the MN care-of-address during the lifetime of a connection,
- multicasting packets to several ARs consumes resources,
- having a fixed route to the outer network (always through the same gateway) is less scalable,
- adaptability and techniques (speed and reliability) to changing routing paths,
- having an optimal routing path from the gateway to the AR, and
- support for QoS routing.

Multicast approaches can have ill effects on the resource availability, for example, because the multicast group can vary very dynamically. The required resources for assured packet forwarding might change rapidly inside the domain, triggering different QoS-related control signalling and resource reservations.

The use of tunnelling can affect the forwarding of QoS-sensitive flows since the original IP packet is encapsulated within another IP packet. However, as long as the tunnel end points are capable of provisioning resources for the tunnelled traffic flows, the agreed QoS level need not be violated. Tunnelling has the advantage that multiple traffic flows can be aggregated onto a single reservation, and there is inherent support for QoS routing.

Micro-mobility schemes that rely on explicit per-host forwarding information do not have such simple support for QoS routing, because there is only one possible route per host. Both IntServ and DiffServ have been extended to cope with tunnelling [10,52] and the changes to the IP address [30]. Some coupling of the macro- and micro-mobility protocols and the QoS architecture may still be needed to ensure an effective total architecture.

4.4. Discussion

In this section we presented the problems faced when trying to provide QoS in mobile environments. The depth of handover can have a large impact on the disruption caused to the QoS for an application. This is because each of the different depths provides various levels of localisation of the signalling required to establish new routes to the MN and install the QoS information in the relevant network nodes in the new path. If the signalling must travel end to end, this introduces more latency than if the changes can be localised. The use of micro-mobility protocols helps to minimise the latency between handover and the installation of the routing and QoS information, but only when the MN remains attached to an AR within the same administrative domain.

On the application layer, the effects of handovers are noticed as sudden disruption in the data flows. Adaptive transport protocols react to these changes by backing off, for example, if enough delay is noticed, TCP reverts to slow start and lowers its expectation of the capacity of the link. Once the new route is stabilised, TCP can regain the original transmission speed. When using RTP, the disruption in the transfer is reported to the application by RTCP, the control protocol of RTP. This allows the application to revert to lower transmission speed, for example, by changing the codec and bit rate used to compress the audio or video stream. The impact and the way RTP behaves due to handovers greatly depend on the stream and the codec used.

Possible solutions to the problems of providing QoS in the mobile environment are introduced in the following section.

5. Solutions

This section identifies various schemes for providing parts of an all-inclusive support of QoS-aware mobility. A full support of mobile terminals with QoS requirements can be accomplished by a combination of these schemes.

5.1. Packet handling at network edges

Network operators already intercept each packet arriving from an external network and decide whether the packet can be allowed into the core network. This admission control is performed by a node called the firewall and is based on IP addresses and port numbers e.g. identifying applications. Firewalls are typically deployed for security reasons and usually scan both incoming and outgoing packets.

The firewall operation can be extended using different rules for performing the admission control. Instead of just preventing known security problems, the edge nodes can use defined bandwidth and QoS policies on a per-flow basis for controlling the traffic admitted into the network. Both the ARs and the gateways perform the admission control, the former for flows originating from MNs and the latter for flows emerging from external networks.

When a previously unknown packet arrives, the edge node will check for the SLA and policies stored for the particular MN being contacted. A central bandwidth broker is in charge of the policy management, and once it receives a request from an edge node, it checks its databases for the proper forwarding rules and returns them to the edge node. Adjusting the load created by best-effort traffic is vital. The common open policy service (COPS) is the protocol architecture for co-ordinating policies between network nodes [4,16].

This method can be used to adjust the load admitted into each service class, if the network is operating with aggregate service classes, and not per-flow, as with RSVP. This can decrease the network load and thus allow for smoother handovers, especially if the traffic belonging to the best-effort class is not consuming all leftover capacity. Therefore, there is enough bandwidth left to support moving terminals. For example, when a mobile moves, and while the resource allocations are not yet in place, the flows of the mobile could receive a special treatment for a period of time. During that time, the mobile and the network can set up the resources for the path.

The ARs should not need to make the primary policing decisions when the arriving load exceeds the capacity of the forward link. If we allow downlink traffic to flood the AN, mobility management schemes are affected. The bandwidth broker could be used to co-ordinate the AN resources and configure the gateways to drop excess traffic. In addition, a bandwidth broker can be used to prioritise existing resource allocation in view of incoming flows. This would be similar to existing cellular networks, where ongoing calls are not disconnected when a new client wants to make a phone call; if resources are not available, the new client just has to wait for resources to become available.

5.2. Coupling of micro-mobility and quality of service

The following investigation covers possible mechanisms by which the performance of reservation-based QoS, as defined in the IntServ architecture [5], can be enhanced for the domain internal routing structure. IntServ implicitly assumes that the route taken by a traffic stream across a network is reasonably stable for the duration of a reservation. The reservation is installed along the path using a QoS signalling protocol, the most widely adopted of which is RSVP. For simplicity, RSVP is used as an example protocol in the following discussion, but the concept can be extended to any other out of band soft state mechanism. When using IntServ with RSVP, changes to the path are handled by the soft-state nature of the architecture, and reservations are installed along the new path by periodic refresh messages. The installation of the reservation along the new route is not immediate, and the level of QoS received by a traffic flow can be temporarily reduced. In contrast, the routes in the mobile environment can be dynamic, changing every time the MN changes AR. Therefore, there is a need for fast re-establishment of paths and QoS reservations. If this is not supported, unacceptable disruption to the application traffic can occur every time the MN changes location.

In order to improve the behaviour of reservation-based QoS in the micro-mobile environment, the QoS and micro-mobility mechanisms can be coupled to ensure that reservations are installed as soon as possible, after a mobility event such as handover. In this study we present three levels of coupling over three different micro-mobility schemes. Here we present the key aspects of the three schemes relevant to this discussion, although a more deep classification can be found in Section 3.3:

- *Proxy agent architectures* [20,33,47] tend to employ tunnels, either a single tunnel or a hierarchy of tunnels, to forward traffic to the CoA allocated to a MN. The tunnel-based micro-mobility mechanisms add scalability to RSVP because reservations can be aggregated onto a single trunk link between mobility agents, and support for QoS aware routing is possible because it will simply effect the route the tunnel takes across the network.
- *The MANET-based scheme* considered in the following discussion uses MER-TORA [38,39] to distribute the routing information within the network. After handover, a host-specific route is inserted into the network, using route update messages, to ensure that traffic travelling to the MN can be routed to its new location.
- *Per-host forwarding schemes* use soft-state host-specific forwarding entries for each of the MNs within a domain. The entire domain has a special gateway that is the default route via which all nodes access the external network. Routing information is refreshed periodically, and updated immediately during handover to install the explicit route to the MNs new location.

The three scales of coupling presented for consideration are described in the following sections.

5.2.1. De-coupled

In the de-coupled option, the QoS and micro-mobility mechanisms operate independently of each other and the QoS implementation is not dependent on a particular mobility mechanism. The QoS reservations are installed using RSVP signalling and IntServ control service parameters, and routing information is distributed using either standard or specialised micro-mobility routing protocols. Changes in network topology are handled by the soft-state nature of the reservations.

Potential problems with this approach occur when the MN hands over to a different AR and the path to and from the MN changes. A section of the old reservation, up to the point where the path to the old AR (OAR) and the new AR (NAR) intersect, is no longer valid because the traffic flows to and from the MN are now travelling via different network nodes. This node can be referred to as a crossover router, similar to the crossover router concept used in some micro-mobility schemes, and is illustrated in Fig. 5.

In order to provide the required QoS to the MN's traffic streams, reservations are required along new paths to and from the MN's new location. These are installed by the refresh mechanism used by RSVP to maintain the soft-state reservation information. The refresh messages are generated periodically, and in the mobile environment there will be a disruption to the agreed QoS during the interval between the MN moving location and the generation of a refresh message. If the refresh message is generated before the route to the MN's new location has been completely propagated throughout the network, the reservation will be made along an incorrect route and not corrected until the next refresh message. The reservation may even be refused if the resources are not available along the incorrect path or the router cannot route the data to the required destination. This will occur every time the MN moves AR, which may be many times during one RSVP session, and can lead to poor overall QoS for an application.

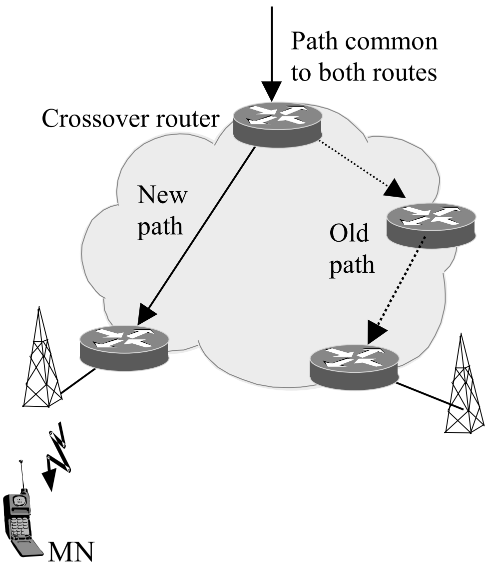

Fig. 5. Concept of a crossover router.

In addition, the reservation along the old path cannot be explicitly removed, and must be left to timeout, which is not the most efficient use of network resources.

These problems are common to all micro-mobility schemes.

5.2.2. Loosely coupled

The loosely coupled approach uses mobility events to trigger the generation of RSVP messages, which distribute the QoS information along new paths across the network. The RSVP messages can be triggered as soon as the new routing information has been installed in the network. This mechanism is the local path repair option, and is outlined in the RSVP specification [12] and has the effect of minimising the disruption to the application's traffic streams because there is a potentially shorter delay between handover and reservation set up. It also avoids the problem of trying to install a reservation across the network before the routing update information has been propagated. The latency for installing the reservation can also be reduced by localising the installation to the area of the network affected by the change in topology, i.e. between the crossover router and the NAR. The areas of the network affected by the topology change can have reservations installed across them almost immediately, instead of having to wait for the update to travel end to end, or for the correspondent node to generate a refresh message for reservations to the MN. In the case where the QoS must be re-negotiated, however, end-to-end signalling is required. The old reservation should be explicitly removed, freeing up unused resources immediately.

However, the loosely coupled approach requires additional complexity within the intermediate network nodes to support the interception and generation of RSVP messages when the router is acting as the crossover node. Another disadvantage is that bursts of RSVP signalling messages are generated after handover to install multiple reservations. This does not happen in the de-coupled case, because the reservation signalling messages are generated when refresh timers expire, not by the same triggering event.

In the *proxy agent architectures* the loosely coupled approach overcomes the problem of reservations being installed before valid routes to the MN are available by ensuring that the reservation is not installed until the registration information generated by the MN has propagated across the network. Reservations from the MN will not be installed until the acknowledgement of registration is received. This indicates that information concerning the CoA of the MN has been distributed in the network, and that a valid route to the MN's location is known. The crossover mobility agent can create reservations to the mobile.

In *MANET based schemes*, the loosely coupled approach associates the two mechanisms via triggering. For reservations from the MN, the receipt of a route update acknowledgement indicates that the explicit route to the MN's new location has been installed in the network, and causes the generation of the refresh messages to provide the fast re-establishment of the reservation. For reservations to the MN, the crossover router is responsible for generating the appropriate RSVP messages.

In *per-hop schemes* triggering is also used to perform the integration. For example, in HAWAII after a handover QoS signalling can be triggered once the new routing information has been distributed into the network and RSVP can make use of its routing interface to receive a path change notification. The reservation is installed in the network as soon as the route to the MN is stable without having to wait until the next timeout to send QoS messages.

5.2.3. Closely coupled

The closely coupled approach combines by using the same signalling mechanism to propagate the mobility and QoS information, either as an extension to the QoS/MM signalling protocol or via a unique QoS-routing protocol. This approach minimises the disruption to traffic streams after handover by ensuring that the reservation is in place as soon as possible after handover. However, instead of having to wait for an acknowledgement that the route to the MN is in place in the network, as with the loosely coupled approach, the QoS requirements for traffic flows travelling to the MN can be installed at the same time as the routing information. This avoids the problem of installing a reservation before valid routing information to the MN has propagated across the network, and also provides a means to install multiple reservations using one signalling message. This reduces the bursts of QoS signalling traffic sent across the network that occurs with the loosely coupled approach.

As with the loosely coupled strategy, the QoS reservation updates can be localised to the area affected by the topology change, unless end-to-end re-negotiation is required. The reservation along the old path can also be explicitly removed. However, the closely coupled approaches place requirements on the micro-mobility mechanisms to transparently carry opaque QoS information and additional complexity is required in the intermediate nodes. In some cases, additional micro-mobility messages are required to support this solution.

In the *proxy agents architectures*, this closely coupled version extends the loosely coupled strategy commented before for this scheme, with additions that support the opaque transport of QoS information in the registration messages. The addition of QoS information in the registration messages allows the MN to choose a mobility agent based on the available resources. This feature can provide some degree of traffic engineering within the network.

In the *MANET based scheme*, the closely coupled approach extends the loosely coupled solution so that the route update messages transparently carry opaque QoS information about traffic flows travelling towards the MN. The reservations are installed at the same time as the routing information, minimising the disruption to the traffic flows.

Finally, in *per-host schemes* the closely coupled approach is very similar to that commented for MANET schemes. This integration can be performed easily because both mechanisms rely on soft-state signalling mechanisms based on path set-up and refresh messages. The suggested most suitable way to perform this integration is to extend the micro-mobility protocol to opaquely carry IntServ objects to distribute the QoS control information at the same time as the routing data.

5.2.4. Comparison of approaches

Coupling reservations with micro-mobility mechanisms allow reservation set up delays to be minimised and packet loss reduced. Reservations along the new path can be installed faster because QoS messages can be generated as soon as the new route is established, reducing the disruption to the data flows. Also scalability and overhead are improved because a minor number of update messages are sent or they are localised to only the affected areas of the network.

Another advantage of coupling the two mechanisms is that it ensures that the request for a QoS reservation only occurs when there are valid routes to the MN in the network. Otherwise, the reservation will be installed along the incorrect route, and may be rejected if the resources along that route are not available, or if the route to the required destination is unknown.

The closely coupled approach requires support from particular micro-mobility mechanisms so that the opaque QoS information can be conveyed across the network. This has the consequence that the QoS implementation will be specific to a particular micro-mobility mechanism, and extensions to the micro-mobility protocol may be needed to support the required functionality. However, the closely coupled approach maintains consistency between the reservation and the routing information within the network, and can reduce the amount of signalling required to set-up multiple reservations.

The choice between whether to use the loosely coupled approach or the closely coupled approach is a trade-off between a QoS solution that is tied to a micro-mobility protocol and the performance advantage close coupling provides. The closely coupled approach potentially provides improvements in performance and efficiency, but at the expense of additional complexity and loss of independence from the underlying micro-mobility mechanism.

5.3. Changes to existing micro-mobility protocols

Based on the investigation of mobility mechanisms in Section 3 three types of protocols are chosen as candidates for integration with QoS protocols. The studied approach follows the most efficient classes of protocols (Section 3.3) and particular examples of protocols belonging to them. These are per-host forwarding schemes with the choice of HAWAII, MANET-based schemes with the choice of MER-TORA and PAA with the choice of HMIP. However, it should be noted that this is our preliminary choice and we plan to further refine it by proposing to merge some protocols, that is, their particular PDI-solutions, or propose some new mechanisms where necessary. Design is in progress of independent PDI-solution specifically concerning the path updates, handover management and support for idle mobile nodes. Although this will result in a proposal for new protocols our preliminary results indicate that there will be three of them. These are a combination of cellular IP and HAWAII [27], modified MER-TORA with extension such as paging support and flexible variant of Hierarchical mobile IP with the possibility of adjusting the number of proxy agents depending on the network scenario and topology. Regardless of this, our current choices for modelling integration of QoS and mobility management seem appropriate since the essence of the operation and interoperability issues will be almost identical. The keys PDIs that we will be studying are the development of path updates, handover management and addressing.

The MH may experience wide variations of QoS due to mobility. When a MH performs a handover, the AR in the new cell must take responsibility for allocating sufficient resources in the cell to maintain the QoS requested (if any) by the node. If sufficient resources are not allocated, the QoS needs may not be met, which in turn may result in premature termination of connections.

It is clear that when a node requests some QoS it is requesting it for the entire connection time, regardless of whether it is suffering handoffs or not. The currently proposed reservation protocol in the Internet, RSVP, implements so-called *immediate reservations* which are requested and granted just when the resources are actually needed. This method is not adequate to make guaranteed reservations for mobile hosts. To obtain mobility independent service guarantees a MH needs to make *advance resource reservations* at the multiple locations it may possibly visit during the lifetime of the connection.

There are a number of proposals for advanced reservations in the Internet community that can be classified into two groups, depending on the techniques they use:

- Explicit advanced reservation signalling,
- Admission control priority.

Those groups are not necessarily distinct, as both approaches could be used together. Admission control strategies are transparent to the mechanism using explicit advanced reservations, other than when a request is rejected.

5.3.1. Admission control priority

It is widely accepted that a wireless network must give higher priority to a handover connection request than to new connection requests. Terminating an established connection from a node that has just arrived to the cell is less desirable than rejecting a new connection request. *Admission control priority based mechanisms* rely on this topic to provide priorities on the admission control to handover requests without significantly affecting new connection requests.

The basic idea of these admission control strategies is to reserve resources in each cell to deal with future handover requests. The key here is to effectively calculate the amount of bandwidth to be reserved based on the *effective bandwidth* [18] of all active connections in a cell and the effective bandwidth of a new connection request.

There are a number of different strategies to do this:

Fixed strategy: one simple strategy is to reserve a fixed percentage of the AR's capacity for handover connections. If this percentage is high, adequate capacity will most likely be available to maintain the QoS needs of handover connections, but at the expense of rejecting new connections.

Static strategy: the threshold values are based on the effective bandwidths of the connection requests. There is a fraction of bandwidth reserved for each of the possibly traffic class. This fraction may be calculated from historic traffic information available to the AR.

Dynamic strategy: each AR dynamically adapts the capacity reserved for dealing with handover requests based on connections in the neighbouring cells. This will enable the AR to approximately reserve the actual amount of resources needed for handover requests and thereby accept more new connection requests as compared to in a fixed scheme. Such dynamic strategies are proposed and evaluated in Refs. [37,58].

Advanced dynamic strategy: this strategy assumes an analytical model where handover requests may differ in the amount of resources they need to meet their QoS requirements, and therefore it is more suitable for multimedia applications. A proposal for this strategy is described in Ref. [45].

This kind of admission control strategy can be used on statistically access control as the one performed on non hard guaranteed QoS provision, such as some DiffServ PHBs or controlled load on IntServ model. It is not enough for hard guarantees in all paths followed by a MN.

5.3.2. Explicit advanced signalling

Admission control strategies are not enough to accommodate both mobile hosts that can tolerate variations in QoS and also those that want mobility independent service guarantees in the same network. To obtain good service guarantees in a mobile environment, the MH makes resource reservations at all the locations it may visit during the lifetime of the connection. These are known as *advanced reservations*.

There are a number of different approaches for advanced reservation in the literature. We present here two of the most relevant for supporting IntServ (MRSVP [51]) and other for supporting DiffServ (ITSUMO approach [13]).

MRSVP

Mobile RSVP is an advance reservation protocol for supporting IntServ in a network with MNs. It introduces three service classes to which a mobile user may subscribe. All of them offer certain guarantees to the delay bounds for the mobile flows and all of them apply as long as the MN is conformed to its traffic parameterisation and mobility specification (MSPEC). These are: *mobility independent guarantees* (MIG) in which a mobile user will receive guaranteed service, MIP in which the service received is predictive, and *mobility dependent predictive* in which the service is predictive with high probability although there may be circumstances of high load when the service may be seriously degraded.

MRSVP considers a network architecture in which a MN can make advance resource reservation along the data flow paths to and from the locations it may visit during the lifetime of the connection. In MRSVP reservation model, a MN can make advance reservations from a set of locations, called MSPEC. Ideally, the MSPEC should be the set of locations the MN will visit while it participates in the flow. The advance determination of the set of locations to be visited by a MN is an important research problem, although several mechanisms have been proposed to approximately determine them by the network. Also, in many situations, a MN can specify its own MSPEC as part of the mobility profile. In any case, it can be assumed that the MN has acquired its MSPEC, either from the network or from its mobility profile, when it initiates a reservation. In the MRSVP reservation model, the MSPEC of a MN can be changed dynamically while the flow is open. In such a case, resources will be reserved at the newly added locations of the MSPEC only if enough resources are available on the data flow path to/from those locations.

Two types of reservations are supported in MRSVP: active and passive (Fig. 6). A mobile sender makes an *active* reservation from its current location and it makes *passive* reservations from the other locations in its MSPEC. Similarly, a mobile receiver makes an active reservation to its current location and passive reservations to the other locations in its MSPEC. On a link, active and passive reservations for a flow are merged. However, either of the active and passive reservations for the same flow on a link can be removed without affecting the other. To improve the utilisation of the links, bandwidth of passive reservations of a flow can be used by other flows requiring weaker QoS guarantees or best effort service. However, when a passive reservation becomes active (i.e. when the flow of the MN who made the passive reservation moves into that link), these flows may be affected. The resources of passive reservation are multiplexed among the different classes of users.

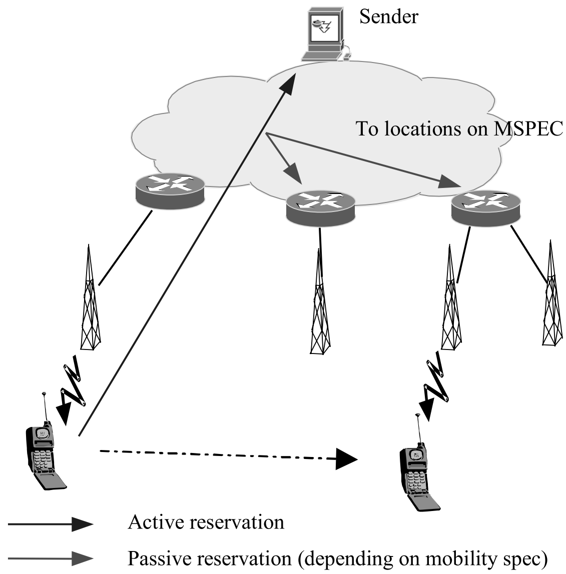

Fig. 6. MRSVP advanced reservations.

In MRSVP, a unicast packet is delivered to a MN by using the mobile IP routing protocol. In such a case, resource reservations for a MN must be established along the route determined by mobile IP. This implies that, when the MN is located in a foreign subnet and the unicast packets for the MN is delivered via its home agent by IPIP tunnelling, resource reservations must also be established over the tunnel (provided the routers on the tunnel are RSVP capable).

ITSUMO approach

As seen in Section 2.5, the ITSUMO approach has a different philosophy on advanced reservations. Although the MN itself has to explicitly request a reservation and specify a mobility profile, the advanced reservation is 'made' by global QoS server (GQS) on its behalf. Based on the local information and the mobility pattern maybe negotiated in the SLS, the QGS envisions how much bandwidth should be reserved in each QLN. The QGS then updates periodically the QLNs likely to be visited by a MN. However, it would seem more resource efficient if either a passive reservation (utilised for best effort traffic) or a "handover guard band" could be used.

The clear difference with the previous approach is that advanced reservation in MRSVP has to be signalled by the MN explicitly to every station according to its mobility pattern. This mobility pattern is known and processed by it. In the ITSUMO approach this information is updated periodically by the QGS, according to the mobility pattern informed by the MN but processed on the QGS. So it could be said that MN relies the explicit advanced reservation in the QGS (Fig. 7).

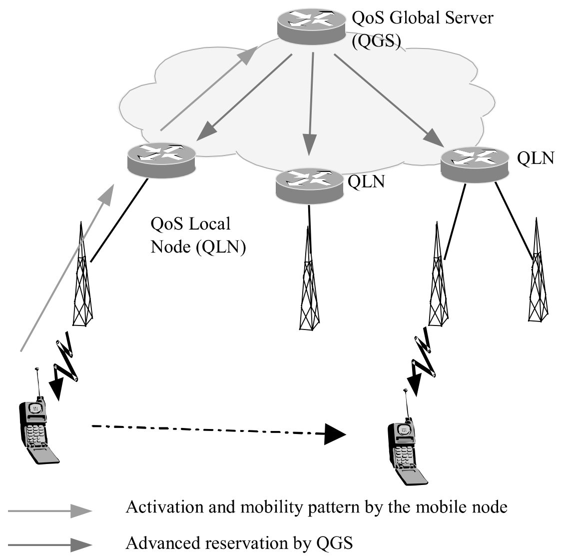

Fig. 7. ITSUMO advanced reservations.

5.4. Pre-handover negotiations

Allocating wireless link resources tentatively in a cell in order to provide better support for future incoming MNs raises questions on the actual use for these "mobility-support" resources. For example, some implementations reserve these resources and give them to the best-effort CoS, thereby providing better support of the background traffic. If several MNs arrive in the same cell, these resources are taken away from the background traffic and given to the new MNs. The background traffic may have adjusted to the larger amount of available resources and if the resources are dramatically reduced, the momentary service offered is turned upside down. In general, a network provider should not have very dramatically changing amounts of resources, but rather provide some gradually changing resource availability in order to let applications to adapt.

To counter the advance reservation scheme, the handover event and associated change to a new cell should be tied to actual resource availability in the new cell. When the network or the MN deem that a handover should occur, the MN could broadcast resource queries to the neighbouring ARs and request some indication of resource availability.

A context transfer protocol is under development in the IETF Seamoby Working Group. Context transfer during handover seems to be a good alternative for handling the resource co-ordination issue. The scheme implies that the MN would interact with two or more ARs at the same time in order to query for resource availability and make a reservation before doing the actual handover. This is possible mainly in "hot-spot" types of cellular networks, as for example, wireless LANs. In modern cellular networks context query and transfer would need to be handled, and the mobile informed about which cell it should handover to (mobile oriented, network-assisted handover). Alternatively, the network could perform the handover automatically (network co-ordinated handover). The question of a MN communicating with several ARs at the same time implies that several ARs cover the same physical area, thus the costs of such a network would be high.

Querying the resource availability in different access points requires support from the convergence layer between the IP layer and the link layer. The link layer would need to communicate the overall resource availability of an access point in order to let the IP layer to make a decision about a possible handover.

Initially context transfer would enhance handovers between access routers, allowing ARs to communicate, either directly or through the MN, the QoS context of a moving MN. A further refinement to the scheme would allow both AR and gateways to communicate the mobile's context during a handover. This would allow the time during which the mobile has no specific resources allocated to it to be reduced.

5.5. Solutions in third generation mobile communication systems

The currently evolving design of the third generation mobile communication systems (3G systems) aims to provide real-time multimedia services in wide area cellular networks [54]. These systems will include a packet-switched backbone (PSB) to carry data traffic in the form of IP datagrams, in addition to the traditional circuit switching for voice calls. MHs will connect to the radio access network (RAN) to gain access to the PSB. IP datagrams will be natively transported from the PSB to the Internet and vice versa, using the internet gateway serving node (IGSN) node as a gateway. As the standardisation of 3G systems evolves, more and more IETF protocols are incorporated into the architecture. UMTS Release 2000 considers the PSB as an IP backbone using the same protocols as IP fixed networks, while the RAN will use proprietary protocols. For the IP-based data transmission, this RAN is seen as a link layer.

Regarding mobility management and provision of QoS, 3G systems are still different from IP based fixed networks; three types of mobility are considered in 3G systems: terminal, personal and service mobility. Service mobility provides the same set of services regardless of the current point of attachment to the 3G network. Personal mobility allows users to receive their personalised service independent of their location in the network. This is an application level issue, therefore out of the scope of this article. As a brief summary, 3G systems specify their own personal mobility schemes. But lately SIP is being considered as a good candidate due to its natural interconnection with fixed servers.

The support for terminal mobility across different operators is a key requirement in 3G systems. To this end, the support of mobile IP is being considered with some proposed extensions [15]. In essence, the IGSN will act as FA supporting macro-mobility, while the movements of the terminal inside the universal terrestrial radio access (UTRA) are not visible outside the 3G network. UTRA has its own mobility support, which is considered link layer mobility for the IP transmission. Authorisation in the UTRA is also supported using the proprietary mechanism of 3G systems.

Regarding the provision of QoS, 3G systems will incorporate two new features with respect to 2G systems and their evolutions: support for user/application negotiation of UMTS bearer characteristics and standardised mapping from UMTS bearer services to core network QoS mechanisms, likely based in IETF QoS provision architectures. In 3G systems four traffic classes will be supported: conversational, streaming, interactive and background [22]. The main distinguishing factor between these classes is how delay sensitive they are.

5.6. Discussion

In this section we presented possible solutions for providing QoS in a mobile environment, as well as existing solutions used in future 3G environments. Each proposed solutions address different aspects of QoS provision, but when combined, may provide a complete QoS model for IP mobility.

The main problem with providing service quality in a mobile environment is following the MNs movement fast enough in order to provide nearly constant service. Different architectures have different amounts of QoS set up signalling, that is why, for example, DiffServ has a good behaviour in the mobile environment, but may not provide same guarantees as reservations-based schemes like IntServ. Also the distance between the communicating nodes affects the needed signalling and thus affects the service quality offered. A full support for flow differentiation on the whole end-to-end path may not even be available, which may make even the best mobile QoS schemes useless.

As discussed, an alternative to following the MN and setting up QoS after a handover would be to reserve resources prior to the handover and thus minimise the disruption caused to the packet flows. It also seems important that the QoS reservations have some level of coupling in order to ensure that they work reasonably effectively in mobile environments.

6. Conclusion

This paper first presented different architectures and protocols that provide QoS. We looked into IntServ, DiffServ, the combination of these, the RTP, INSIGNIA and ITSUMO. The benefits and shortcomings of each of these approaches were outlined, with the conclusion that none of them really provides both scalable and accurate QoS.

Then we went on and presented different ways to provide mobility to hosts. We studied macromobility through mobile IP and several micromobility schemes. Our novel analytical model for classifying and breaking-up of mobility protocols and their mechanisms supported the investigation of the integration of QoS and mobility protocols.

In the second part of this paper we discussed problems related to mobility and QoS. We deduced that the main problem in this field is following the movement of the MH fast enough to minimise the disruption caused to the QoS received by the application traffic flows. Also the depth of the handover signalling and the related QoS control affect the service outcome.

In Section 5 we studied solutions for the interoperability of mobility and QoS. We presented several schemes that provide parts of a total solution to mobile QoS. We discussed performing strict flow shaping at the network edge, coupling of micro-mobility and QoS protocols, modifying micro-mobility protocols themselves, advanced reservations, pre-handover negotiations and context transfer, and the 3G approaches.

It has become apparent that even though there exist several good partial solutions, we still need adaptive applications. Handovers, for example, still cause some disturbance to data streams. RTP can provide to this adaptability. The whole notion of end-to-end QoS still seems very distant. It is possible to provide adequate service to MHs in a private AN, but when the corresponding node is behind some wider public network, keeping the promised QoS becomes harder.

The third generation systems aim to provide high-speed mobile access. These architectures are however closed systems and highly controlled by government authorities and a few operators. In the future we will most likely see a growing number of wireless LANs covering hot-spot areas like shopping malls, airports and train stations. These networks can be based on open IP and architectures increasing the ease of deployment and minimising cost.

The IETF Working Group, Seamoby, is aiming to provide seamless mobility across ARs and even domains. The work of this group will hopefully lead to better mobility support, especially for the problematic multimedia streams. Part of the work done is on context transfer issues. In addition, QoS issues with the mobile IP are under study within the IETF, as well as an alternative QoS signalling protocol to RSVP.

The BRAIN and MIND projects, funded by the European Commission, are also studying an open IP-based mobile wireless network architecture. The projects propose and evaluate an architecture that provides support for mobile QoS, a framework for adaptive applications, and a well defined interface through which applications can request the level of service they require. The architecture has well defined interfaces to allow inter-operability among networks.

Acknowledgements

This work has been performed in the framework of the IST project IST-1999-10050 BRAIN, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues from Siemens AG, British Telecommunications PLC, Agora Systems S.A., Ericsson Radio Systems AB, France Télécom—CNET, INRIA, King's College London, Nokia Corporation, NTT DoCoMo, Sony International (Europe) GmbH, and T-Nova Deutsche Telekom Innovations-gesellschaft mbH.

References

[1] D.O. Awduche, E. Agu, Mobile extensions to RSVP, Proceedings of ICCCN'97, Las Vegas NV, September 1997, pp. 132–136.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An Architecture for Differentiated Services, Internet Engineering Task Force, Request for Comments (RFC) RFC 2475, December 1998.

[3] Y. Bernet, S. Blake, D. Grossman, A. Smith, An Informal Management Model for DiffServ Routers, Internet Draft (work in progress), February 2001 (draft-ietf-diffserv-model-06.txt).

[4] J. Boyle, R. Cohen, D. Durham, S. Herzog, D. Rajan, A. Sastry, COPS Usage for RSVP, Internet Engineering Task Force, Request for Comments (RFC) 2749, January 2000.

[5] R. Braden, D. Clark, S. Shenker, Integrated Services in the Internet Architecture: an Overview, Internet Engineering Task Force, Request for Comments (RFC) 1633, June 1994.

[6] Y. Bernet, Format of the RSVP DCLASS Object. Internet Engineering Task Force, Request for Comments (RFC) 2996, November 2000.

[7] Y. Bernet et al., A Framework For Integrated Services Operation Over Diffserv Networks. Internet Engineering Task Force, Request for Comments (RFC) 2998, November 2000.

[8] F. Baker, J. Heinänen, W. Weiss, J. Wroclawski, Assured Forwarding PHB Group, Internet Engineering Task Force, Request for Comments (RFC) 2597, June 1999.

[9] F. Baker, C. Ituralde, F. Le Faucheur, B. Davie, RSVP Reservation Aggregation. Internet Draft (work in progress), April 2001 (draft-ietf-issll-rsvp-aggr-04.txt).

[10] D. Black, Differentiated Services and Tunnels. Internet Engineering Task Force, Request for Comments (RFC) 2983, October 2000.

[11] Broadband Radio Access for IP-based Networks, IST-1999-10050, http://www.ist-brain.org.

[12] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Resource ReSerVation Protocol (RSVP)–Version 1, Functional Specification. Internet Engineering Task Force, Request for Comments (RFC) 2205, September 1997.

[13] J. Chen, A. McAuley, A. Caro, S. Baba, Y. Ohba, P. Ramanathan, A QoS architecture for future wireless IP networks, Twelfth IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS 2000), Las Vegas, NV, November 2000.

[14] A. Campbell, J. Gomez, S. Kim, A. Valko, C.-Y. Wan, Z. Turanyi, Design, implementation, and evaluation of cellular IP, IEEE Personal Communications 7 (2000) 42–49.

[15] S. Das, A. Misra, P. Agrawal, S.K. Das, TeleMIP: telecommunications-enhanced mobile IP architecture for fast intradomain mobility, IEEE Personnal Communications, August 2000.

[16] D. Durham, J. Boyle, R. Cohen, S. Herzog, R. Rajan, A. Sastry, The COPS (Common Open Policy Service) Protocol, Internet Engineering Task Force, Request for Comments (RFC) 2748, January 2000.

[17] K. El-Malki et.al., Low Latency Handoff in Mobile IPv4, Internet Draft (work in progress), May 2001 (draft-ietfmobileip-lowlatency-handoffs-v4-01.txt).

[18] A.I. Elwalid, D. Mitra, Effective bandwidth of general markovian traffic sources and admisssion control of high-speed networks, IEEE/ACM Transactions on Networking 1 (3) (1993) 329–343.

[19] P. Eardley, A. Mihailovic, T. Suihko, A framework for the evaluation of IP mobility protocols, Proceeding of PIMRC 2000.

[20] E. Gustafsson, A. Jonsson, C. Perkins, Mobile IP Regional Registration, Internet Draft (work in progress), March 2001 (draft-ietf-mobileip-reg-tunnel-04).

[21] R. Guerin, A. Oroa, QoS routing in networks with inaccurate information, IEEE/ACM Transactions on Networks 7 (3) (1999).

[22] H. Holma et al., WCDMA for UMTS, UMTS Services and Applications, Wiley, England, October 2000 (chapter 2).

[23] M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg, SIP: Session Initiation Protocol. Internet Engineering Task Force, Request for Comments (RFC) 2543, March, 1999.

[24] G. Huston, Next Steps for the IP QoS Architecture, Internet Architecture Board, Request for Comments (RFC) 2990, Internet Engineering Task Force, November 2000.

[25] V. Jacobson, K. Nichols, K. Poduri, An Expedited Forwarding PHB, Internet Engineering Task Force, Request for Comments (RFC) 2598, June 1999.

[26] D. Johnson, C. Perkins, Mobility Support in IPv6, Internet Draft (work in progress), July 2001 (draft-ietf-mobileip-ipv6-14.txt).

[27] C. Keszei, J. Manner, Z. Turányi, A. Valko, Mobility management and QoS in BRAIN access networks, BRAIN International Workshop, King's College, London, November 2000.

[28] Lee et al., INSIGNIA: An IP-based quality of service framework for mobile adhoc networks, Journal of Parallel and Distributed Computing 60 (4) (2000) 374–406.

[29] J. Mysore, V. Bharghavan, A New Multicasting-based Architecture for Internet Host Mobility, Proceedings of ACM Mobicom, September 1997.
[30] J. Metzler, S. Hauth, An end-to-end usage of the IPv6 flow label, Internet Draft (work in progress), November 2000 (draft-metzler-ipv6-flowlabel-00.txt).
[31] P. McCann, T. Hiller, J. Wang, A. Casati, C. Perkins, P. Calhoun, Transparent Hierarchical Mobility Agents (THEMA), Internet Draft (work in progress), March 1999 (draft-mccann-thema-00.txt).
[32] G. Montenegro, Reverse Tunnelling for Mobile IP, Internet Engineering Task Force, Request for Comments (RFC) 3024, January 2001.
[33] J. Malinen, C. Perkins, Mobile IPv6 Regional Registrations, Internet Draft (work in progress), March 2001 (draft-malinen-mobileip-regreg6-01.txt).
[34] A. Mihailovic, M. Shabeer, A.H. Aghvami, Multicast for Mobility Protocol (MMP) for emerging Internet networks, Proceedings of PIMRC2000, London, UK, September 2000.
[35] T. Narten, E. Nordmark, A.W. Simpson, Neighbor Discovery for IP version 6 (IPv6), Internet Engineering Task Force, Request for Comments (RFC) 2461, December 1998.
[36] K. Nichols, B. Carpenter, Definition of Differentiated Services Per Domain Behaviours and Rules for their Specification, Internet Engineering Task Force, Request for Comments (RFC) 3086, April 2001.
[37] M. Naghshineh, M. Schwartz, Distributed call admission control in mobile/wireless networks, IEEE Journal on Selected Areas in Communications 14 (1996) 711–717.
[38] G. O'Neill, S. Corson, Homogeneous edge mobility for heterogeneous access technologies, Proceedings of the IPCN 2000, Paris, France, May 2000.
[39] A. O'Neil, G. Tsirtsis, Edge mobility architecture–routeing and hand-off, British Telecom Technology Journal 19 (1) (2001) 114–126; Available at: http://www.bt.com/bttj/vol19no1/oneill/oneill.pdf.
[40] C. Perkins, Mobile IP, IEEE Communication Magazine 35 (1997) 84–99.
[41] C. Perkins, IP Mobility Support for IPv4, revised, Internet Engineering Task Force, Request for Comments (RFC) 2002, September 2000.
[42] C. Perkins, D. Johnson, Route Optimisation in Mobile IP, Internet Engineering Task Force, Internet Draft (work in progress), November 2000 (draft-ietf-mobileip-optim10.txt).
[43] D. Plummer, An Ethernet Address Resolution Protocol, Internet Engineering Task Force, Request for Comments (RFC) 826, November 1982.
[44] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, S.Y. Wang, HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks, Proceedings of the Seventh International Conference on Network Protocols (ICNP) 1999, pp. 283–292.
[45] P. Ramanathan, K.M. Sivalingam, P. Agrawal, S. Kishore, Resource allocation during handoff through dynamic schemes for mobile multimedia wireless networks, Proceedings of INFOCOM, March 1999.
[46] S. Seshan, H. Balakrishnan, R.H. Katz, Handoffs in cellular wireless networks: the daedalus implementation and experience, ACM/Baltzer Journal on Wireless Networks, 1995.
[47] H. Soliman, C. Castelluccia, K. El-Malki, L. Bellier, Hierarchical MIPv6 Mobility Management, Internet Draft (work in progress), July 2001 (draft-ietf-mobileip-hmipv604.txt).
[48] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, RTP: A Transport Protocol for Real-Time Applications, Internet Engineering Task Force, Request for Comments (RFC) 1889, January 1996.
[49] Z.D. Shelby, D. Gatzounas, A. Campbell, C. Wan, Cellular IPv6, Internet Draft, (work in progress), July 2001 (draft-shelby-seamoby-cellularipv6-01).
[50] S. Shenker, C. Partridge, R. Guerin, Specification of Guaranteed Quality of Service, Internet Engineering Task Force, Request for Comments (RFC) 2212, September 1997.
[51] A. Talukdar, B. Badrinath, A. Acharya, MRSVP: a resource reservation protocol for an integrated services packet network with mobile hosts, Proceedings of ACTS Mobile Summit'98, June 1998.
[52] A. Terzis, J. Krawczyck, J. Wroclawski, L. Zhang, RSVP Operation Over IP Tunnels, Internet Engineering Task Force, Request for Comments (RFC) 2746, January 2000.
[53] C. Tan, S. Pink, K. Lye, A fast handoff scheme for wireless networks, Proceedings of the Second ACM International Workshop on Wireless Mobile Multimedia, ACM, August 1999.
[54] B.H. Walke, Mobile Radio Networks, Networking and Protocols, Third Generation Cellular, Wiley, England, 2000 (Chapter 5).
[55] J. Wroclawski, A. Charny, Integrated Service Mappings for Differentiated Services Networks, Internet Engineering Task Force, Internet Draft (work in progress), February 2001 (draft-ietf-issll-ds-map-01.txt).
[56] J. Wroclawski, The Use of RSVP with IETF Integrated Services, Internet Engineering Task Force, Request for Comments (RFC) 2210, September 1997.
[57] J. Wroclawski, Specification of the Controlled-Load Network Element Service, Internet Engineering Task Force, Request for Comments (RFC) 2211, September 1997.
[58] O.T.W. Yu, V.C.M. Leung, Adaptive resource allocation for prioritised call admission over an ATM-based wireless PCN, IEEE Journal on Selected Areas in Communications 15 (1997) 1208–1225.

Jukka Manner received his M.Sc. in Computer Science with a minor emphasis in Industrial Management from the University of Helsinki, Finland. He is currently working towards a Ph.D. degree at the University of Helsinki. His research focuses on mobile and wireless communications and QoS in future generation mobile networks. He has actively participated in the European Commission sponsored IST projects "Broadband Radio Access for IP based Network" (BRAIN) and "Mobile IP based Network Developments" (MIND). He has published approximately. 20 articles in conferences and journals.

Alberto López Toledo received his Engineer Degree in Computer Science (1st Class Honours) from the University of Murcia, Spain, in 1999. He is currently working toward his Ph.D. in the Department of Communications and Information Engineering at the same university. In 1999 he joined the ISABEL group at the Department of Telematic Systems Engineering, Technical University of Madrid, as a visitor Research Assistant. Since then he has been involved in several national and international R&D projects in the field of wireless and mobile networks. His research interests include IP mobility, wireless networking and QoS provision in mobile environments.

Andrej Mihailovic received an electronic engineering degree (1996) and a M.Sc. degree in dynamic channel allocation for cellular telecommunications systems (1997) from King's College London, UK. Currently he is a Ph.D. write-up student and a research fellow in King's College London where he researches into IP networks, particularly mobility and multicasting aspects. He published various papers on protocols for solving issues related to IP mobility. His Ph.D. work was conducted in cooperation with British Telecom where he was a full time employee. He was also a member of the European Commission sponsored project BRAIN and is currently participating in the European Commission sponsored project MIND.

Héctor L. Velayos Muñoz was born in Madrid (Spain) in 1972. He obtained his Engineering degree on Telecommunication Systems in 1998 at the Technical University of Madrid (TUM). In 1998 and 1999 he was working at this university as networking researcher in both EU and Spanish funded research projects on the design and deployment of advanced data networks. During 2000 he was working for Agora Systems in the IST BRAIN project on radio systems beyond 3G. Currently he is working towards his Ph.D. at the Royal Institute of Technology, in Sweden. His main area of interest is the design of new architectures for wide area unlicensed wireless data networks.

Eleanor Hepworth is an Engineer at Siemens/Roke Manor Research, where she has worked since graduating with a BSc (Hons) in Computer Science from the University of Southampton. She has been involved in the development of third generation systems and QoS since 1999 and is currently participating in research concerning fourth generation architectures primarily focusing on QoS issues in mobile environments.

Youssef Khouaja was born in Tunis, Tunisia, on December 10, 1976. He received the degree of Computer Scientist engineering from Ecole Nationale des Sciences de l'Informatique, Tunis, in 1999. Since November, 1999, he has been preparing his Ph.D. degree from ENST (Ecole Nationale Superieure des Telecommunications), Rennes, at France Telecom R&D in the Terrestrial Broadcasting and Radio Access Networks Lab. His current work deals with IP based mobile protocols.