QoS-aware path selection in a B3G system

N. Nafisi[*], R. Ferrús[†], A. Gelonch[†], O. Sallent[†], J. Pérez-Romero[†], L. Wang[*], M. Dohler[*], H. Aghvami[*]
[*]Centre for Telecommunications Research King's College London,[†]Signal Theory and Communications Department UPC (Barcelona)
{nima.nafisi@kcl.ac.uk}

Abstract— **In the B3G system considered, which is based on the 3GPP UMTS architecture, the initial hypothesis made in this paper, is that the QoS management of the IP access network is taken in charge by a bandwidth broker. A pre-handover signalling for QoS-aware path selection is investigated with the goal of achieving the *"Always Best Connected"* paradigm. Furthermore, the scalability of the proposed scheme is analysed taking into account the overhead due to IP micromobility, QoS routing, CRRM and bandwidth broker.**

I. Introduction

The objective of the EVEREST project is to devise and assess a set of specific strategies and algorithms for access and core networks, leading to an optimised utilisation of scarcely available radio resources for the support of mixed services with end-to-end QoS mechanisms within heterogeneous networks beyond 3G.

In order to achieve this objective, this paper analyses the apparatus for the selection of a path, among the available paths offered through the various radio access technologies, which fulfils a set of QoS constraints. The approach chosen is centralised, based on the BB (*Bandwidth Broker*), as it is thought that a centralised QoS management of the BB can provide a simple interface between the IP QoS and the CRRM (*Common Radio Resource Management*). A hop-by-hop approach like RSVP [1] for the QoS admission control and reservation would a priori be more complex in order to choose among the CARs (*Candidate Access Router*) the best target AR in conjunction with the RRM.

First a signalling for QoS-aware path selection is discussed, then scalability issues related to the proposed architecture are examined.

II. Signalling for QoS-aware path selection

A. Problem description

In a mobile access network with heterogeneous RANs, the UE (*User Equipment*) has a high probability of being in the range of several APs (*Access Point,* following the terminology in [2]) with the same or different radio access technologies, as shown in figure 1. Moreover these APs are connected to ARs, which form a set of CARs from the IP handover perspective. The paths from each CAR to one of the gateway of the mobile access network may present different IP QoS parameters in terms of jitter, bandwidth and packet loss guarantees. The usual AP selection among a set of candidate APs is done based only on RRM parameters. Once this selection has been done, then the IP QoS reservation is carried out; and based on the successful output of this reservation (for instance in 3GPP the message "Activate PDP Context Accept") an end-to-edge communication path is set up from the UE to the gateway of the B3G access network. However, this approach for end-to-edge QoS and session establishment is not optimised as it considers IP QoS parameters only a posteriori. Other approaches are possible, as suggested in [1], [3], [4] and [5]. In these papers, prior to AP selection a coordination between the RRM entities of the RANs and the IP QoS management is performed. Thus, in a B3G system where the IP access network supports different radio access technologies and can therefore become a source of congestion, a QoS-aware path selection mechanism has to be provided in order to avoid the connection through an AP with not the required QoS guarantee in the access network part.

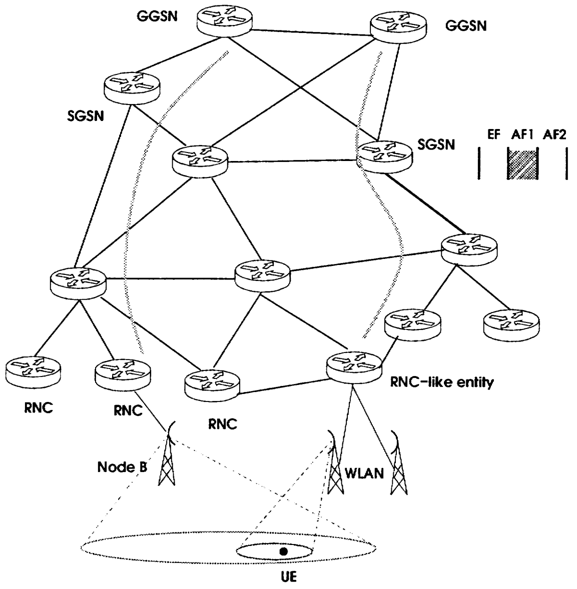

Fig. 1. Heterogeneous mobile access network with tight coupling

A closely related issue to the QoS-aware path selection is the QoS class mapping: radio classes (based on UMTS classes) and IP QoS classes (based on DiffServ). In a DiffServ domain, IP QoS is provided by the differentiation of aggregated flows, to which are associated DiffServ code-points. Based on the assumption of a DiffServ domain, one important issue is the mapping between UMTS QoS classes to DiffServ classes. It can be noticed that IP QoS parameters are only statistically

978-3-8007-2909-8/05/$20.00 ©2005 IEEE

guaranteed and per aggregated flow, whereas UMTS QoS guarantees are per flow. One possibility is a static mapping as suggested in the table I.

TABLE I

QoS CLASS MAPPING

Conversational	EF
Streaming	AF1
Interaction	AF2
Background	BE, AF3, AF4

However a static mapping is not an optimal way of providing end-to-edge QoS, because the QoS resources in the RANs and the IP domain are submitted to different constraints. In the RAN these constraints are related to the wireless medium. In the IP domain they are related to the traffic distribution in the network topology and to the user mobility. For instance let's consider a flow with the streaming class, which is mapped to the AF1 DiffServ class. After a handover to a new AR, it may happen that on the rerouted path the resources assigned to AF1 are not available, as shown in figure 1. Therefore the offered possibility is either to dynamically upgrade to an EF class, for which the IP trunk has enough capacity or to downgrade to a lower class (BE in the worst case), which decreases the overall QoS offered to the session.

In order to provide a coordination between the RRM and IP QoS management in terms of AP selection and QoS mapping, the existing 3GPP QoS architecture has to be enhanced. Nonetheless the enhanced architecture proposed remains compatible with the 3GPP specifications. First the 3GPP (releases 5 and 6) QoS architecture is reviewed, then the Everest QoS proposal is presented.

B. 3GPP QoS architecture

Release 5 introduces a new network domain called the IP Multimedia Subsystem (*IMS*). This is an IP network domain designed to provide an appropriate support for real-time multimedia services. The IMS enables SIP signalling, user authentication and IP end-to-end QoS signalling. The PDF (*Policy Decision Function*), standardised in 3GPP as part of the IMS, enables policy-based admission control in addition to the QoS admission control in the IP and radio domains.

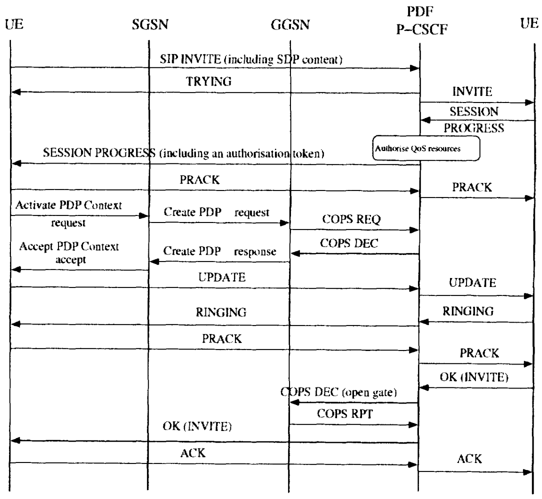

Fig. 2. Session establishment for IMS services

The figure 2 illustrates the required signalling at different levels (SIP, PDP context, PDF COPS signalling) for the establishment of an end-to-end session. The 3GPP QoS architecture is policy-based, and is aligned with IETF protocols (COPS). Two entities have been introduced in the 3GPP architecture: the PDF which is equivalent to a PDP (*Policy Decision Point*), and the PEP (*Policy Enforcement Point*) located at the GGSN. The PDF authorises a session based on the SDP (*Session Description Protocol*) information exchanged during the SIP signalling, and based on the user profile retrieved from the HSS (*Home Subscriber Server*). In figure 2, it can be seen that during the session establishment the successful completion of the SIP signalling is preconditioned by the local QoS reservation in the UMTS domain, which is accomplished by the PDP context signalling. Moreover the session establishment is also preconditioned by the policy admission control done by the PDF. The authorisation by the PDF to the GGSN is carried through the Go interface (based on the COPS protocol). Once the session has been authorised by the PDF, a COPS "DEC" message containing an authorisation token is sent by the PDF to the GGSN. This token is forwarded to the UE through a SIP signalling message, which is then reused in the PDP activation message. This token is finally used in the activation of the GGSN gate, which acts as a firewall for the sessions.

In summary, the 3GPP specifications on the end-to-end QoS architecture give merely a framework and a set of signalling for the end-to-end session establishment, including policy-based admission control (Go interface) and local QoS admission control mechanisms (PDP context signalling). These specifications leave intentionally the possibilities of using specific admission control algorithms and specific IP QoS control management schemes. The following subsection examines the Everest QoS proposal based on the use of a BB for the IP QoS control plane, and the interactions of the BB with the CRRM at the pre-handover phase.

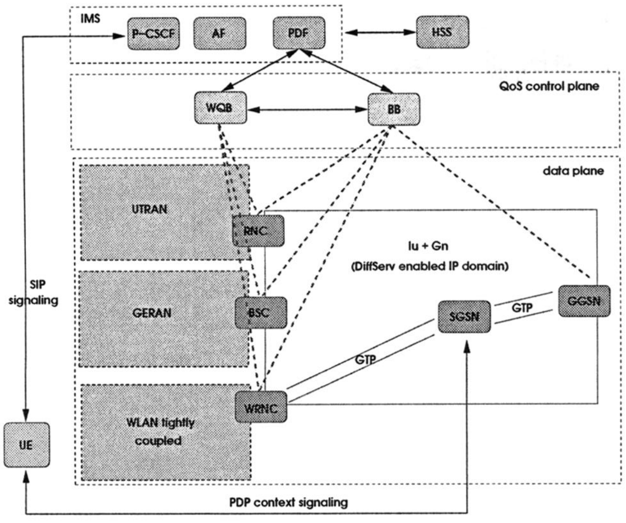

Fig. 3. Everest B3G QoS architecture

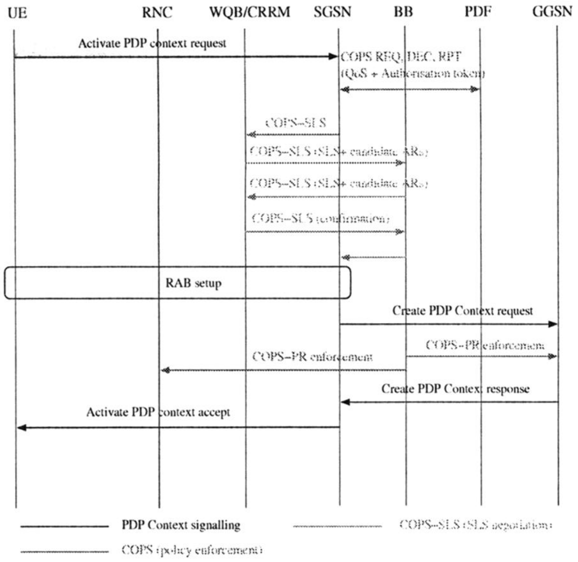
UE
RNC
WQB/CRRM
SGSN
BB
PDF
GGSN
Activate PDP context request
COPS REQ, DEC, RPT
(QoS + Authorisation token)
COPS–SLS
COPS–SLS (SLS+ candidate ARs)
COPS–SLS (SLS+ candidate ARs)
COPS–SLS (confirmation)
RAB setup
Create PDP Context request
COPS–PR enforcement
COPS–PR enforcement
Create PDP Context response
Activate PDP context accept
PDP Context signalling
COPS–SLS (SLS negotiation)
COPS (policy enforcement)

Fig. 4. Session establishment message chart

C. Everest QoS proposal

From release 5 of UMTS, an IP transport can be used for the signalling and data traffic in the UMTS access network, which encompasses the RAN (Iub and Iur interfaces), the Iu interface and the core network (Gn interface). At the layer 2 any suitable technology can be used, including the already deployed ATM technology. Furthermore, in the "all-IP" architecture of UMTS R5, it is specified that DiffServ should be supported in the different interfaces: Iur, Iub (TS-24.434, TS-25.426), Iu (TS-414) and Gn. For the control management of the IP data plane, the BB entity is chosen. Thus, the BB is in charge of the QoS in the IP domain confined between the RNCs and the GGSNs. The QoS control of the BB does not extend to the Node B (or AP), as it is supposed that a RNC and a Node B are connected through a point-to-point link, which is over-provisioned. The IP domain confined between the RNCs and the GGSNs has a partially mesh topology. Thus, RNCs can be connected to multiple nodes, as specified in the 3GPP TS-23.236-v5.2.0.

The interaction of the BB with the CRRM is done by the intermediary of a newly defined entity called WQB (*Wireless QoS Broker*). The WQB is the counter part of the BB for the radio access part. It has similar functionalities: an intra-domain communication interface (in order to configure the RRM functions at the RNCs and Node Bs), monitoring of the available resources, and admission control. Thus, the WQB entity includes all CRRM functionalities and in addition has a COPS-PR interface in order to enforce the policy-based admission control decision taken into the RRM entities (RNC, Node B) [6]. And finally the WQB presents an interface between the CRRM and the BB. This is illustrated in figure 3. Moreover the CRRM policy-based approach is proposed in 3GPP TR-25.891 as a feasible approach over which the WQB concept can be developed. The basic idea behind the CRRM policy-based approach in TR-25.891 is the standardisation of parameters and information exchange over an open interface between RRM and CRRM entities. This would enable the CRRM entity to provide CRRM policies to the RRM entities, thus allowing the traffic distribution in the network to be dynamically adjusted on the basis of a common strategy. In particular, the local RRM entities act as a master of the decisions but decision criteria is determined by the set of installed CRRM policies. A reporting information interface is proposed to achieve a twofold objective: CRRM is aware of network conditions from measurements received from local RRM entities at the same time as local RRMs can receive information of external cells from the CRRM. Upon such a basis, the CRRM approach considered here extends the CRRM policy-based in TR-25.891 by adding a decision support mechanism. In the case of a handover, the decision would be triggered by local RRM entities according to the installed policies. The CRRM gathers measurements per cell (for example the cell load) from RRM entities under its control. The handover decision is taken at the local RRM entity based on the prioritised list of candidate APs provided by the CRRM. This prioritised list is the result of CRRM information on other radio access technologies and also the information received from the BB. The pre-handover signalling between CRRM/WQB and the BB is illustrated in figure 4. It can be remarked that the information exchanged during this signalling contains a prioritised list of candidate AP, but also contains not only information on the SLS parameters [7] including the QoS class mapping.

III. Scalability

Until now, only the signalling necessary for the path selection between the BB and the WQB has been discussed. Nonetheless the use of a centralised QoS entity like the BB, in order to provide QoS-aware path selection, or the use of QoS routing imply a signalling overhead, which is analysed here in a qualitative manner. In this section the signalling between the BB and the IP mobility management during handover, is

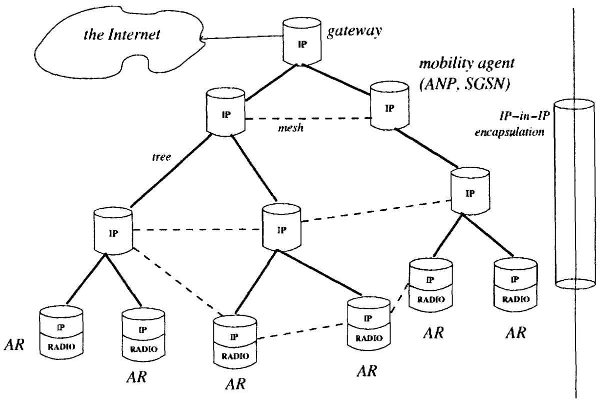

Fig. 5. Micromobility in the IP domain of UMTS

studied and more precisely the qualitative scalability of the proposed system encompassing the BB, WQB and IP mobility management. Here, it will also be shown that not only the BB allows an easy interaction with the WQB, but when a hierarchical structure is adopted for the BB, the WQB and the IP mobility management, QoS provisioning can be provided in a scalable manner.

A. IP micromobility and QoS routing

QoS routing has a central role in QoS-aware path selection during the handover process and which involves the different entities: WQB, BB and IP mobility management. QoS routing optimises the network resource utilisation and enable the selection of paths along which real-time applications like VoIP experience low delay and jitter. Here cross-issues between mobility management and QoS routing are examined.

In the IP domain of the access network, two alternative topologies might be chosen: mesh or tree-based (figure 5). Tree based topology has been used so far in cellular networks. In the Internet, the usual topology used is mesh based. In an IP mobile access network, the choice of the topology can have consequences on the mobility management and on the overall performance of the communication perceived by the user. In a mobile access network for improving the performance of the mobility management between domains, provided by Mobile IP, different micromobility protocols have been proposed. With Mobile IP, each time a mobile node changes its attachment point, a new tunnel has to be established, this incurs a delay, packet loss and signalling overhead. Thus micromobility protocols have been introduced, which are in charge of local mobility inside the access network, independently of the macromobility provided by Mobile IP. Micromobility protocols can be classified into two groups:

- tunnel-based protocols, for which the incoming packets are read at the mobility agent (ANP, SGSN), then a lookup of the UE is done in the visitor list, and finally the packet is tunnelled toward the appropriate router. The end of tunnel is the RNC to which the UE is attached. Examples of protocols based on this mechanism are: GTP, HMIP [8] and BCMP [9].
- host-based forwarding protocols, for which the incoming packets are forwarded based on a location database that maps host identifiers to location information, in a similar way as the precedent class, but here the incoming packets at the gateway are forwarded to a router which is one hop away. This process of forwarding continues until the packet reaches the mobile host. We can mention two examples for this class: Cellular-IP and HAWAII [10].

Even though host-based forwarding micromobility protocols could work with a mesh topology, the present Cellular-IP and HAWAII protocols have been designed only with a tree topology in mind; furthermore conceiving a host-based protocol which works with a mesh topology, would have disadvantages in terms of network overhead (signalling messages exchanged between all the neighbour nodes in the access network). On the contrary a mesh topology is easily handled by the tunnel-based class. Moreover, thanks to tunnelling the tunnel-based micromobility becomes an overlay above the layer three of the access network, where an interior routing protocol takes place. Thus, respectively with a tree topology the host-based forwarding protocols, and with a mesh topology the tunnel-based protocols, are more appropriate. Furthermore, in a mesh topology, there is the possibility of using a QoS routing. For the IP access network the following architecture is considered: a mesh topology and a tunnel-based micromobility protocol, like BCMP with a QoS routing protocol.

One final remark can be done on the interaction between QoS routing and mobility management. If on-demand QoS routing is used, then the IP-in-IP encapsulation used by the tunnel-based micromobility is unnecessary, as source routing is used by the on-demand QoS routing, i.e. in each data packet the IP header (in IPv6 the routing extension header is used) contains the list of the routers along the chosen path by the AR.

Thus, it has been shown that GTP or BCMP can interwork with QoS routing, which enables QoS-aware path selection in the IP access network. However the use of QoS routing implies a signalling overhead due to the dynamic link state advertisement, which can be decreased if several routing areas are defined in the IP access network instead of one. The utilisation of several routing areas makes the BB QoS control management overlay more scalable as explained in the following subsection.

B. Hierarchical structure

A hierarchical structure is a well known remedy to the disadvantages of centralised architectures for the QoS (BB) or the mobility management (tunnel-based approaches: GTP, BCMP). The figure 6 shows the envisaged approach for a hierarchical structure for the mobility agent, the BB and the WQB. The hierarchical or distributed structure for each of these entities is examined successively.

1) Mobility agent: The access network is divided into two parts: the so-called core domain and the edge domains, which are connected in a hierarchical manner to the core network, i.e. the edge domains are only inter-connected through the core

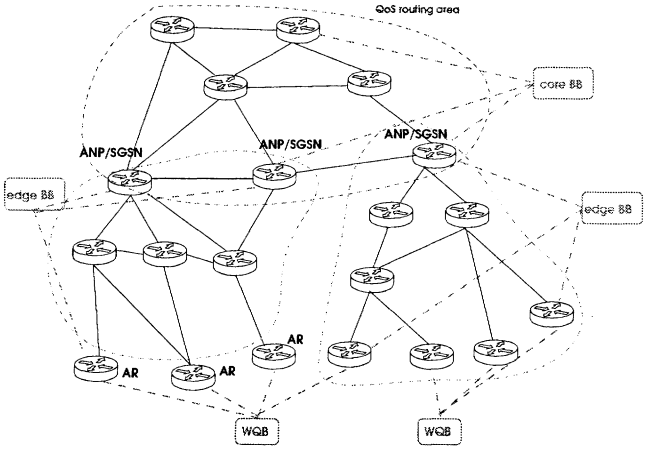

Fig. 6. Hierarchical architecture

domain. The routers interconnecting each edge domain to the core domain are called ABR (*Area Border Router*) following the hierarchical OSPF terminology [11]. One issue here is the placement of the ANP (or SGSN) in this topology. The closer the ANP is to the AR, the lower is the IP handover latency [10], nonetheless as in each domain QoS routing is considered the only possible position is at the ABR in order to avoid the subdivision of the QoS path into two separate paths. Thus, incoming packets at the ABR, where resides the ANP, are tunnelled or source routed to the corresponding AR. The same mechanism of tunnelling or source routing occurs between the gateway and the ANP on the fringe of the core network, as shown in figure 6.

2) Hierarchy of BBs: There is one BB in each edge domain and in the core domain, although several ANPs can be present in each edge domain in order to avoid a single transit point for all down-link packets (up-link packets do not need mobility management functionalities, therefore they can be source routed or tunnelled to a ABR with or without ANP functionalities). A hierarchy of BBs is a more scalable architecture compared to a single BB for the following reasons:

- there is not only a single BB for the whole access network to which all user requests are destined. A single BB per access network can become a bottleneck point.
- the BB monitors less routers.

However this increased scalability requires an additional signalling between the edge BBs and the core BB in order to establish an end-to-edge QoS session or to redistribute dynamically the resources in the core domain assigned to the edge BBs for load balancing purposes [12], [13]. Moreover signalling between edge BBs is also necessary when edge domain handover occurs and QoS contexts have to be exchanged between edge BBs.

3) WQB: The same scalability observations apply to the WQB. Instead of one WQB for the whole access network, several WQB entities can be envisaged. Each of these WQB would monitor a predefined set of RNCs and Node Bs. The set of RNCs monitored by a WQB does not have to belong to only one edge domain. On the contrary to the BB domain, which has to correspond to a routing area, the WQB set of monitored RNCs can overlap several routing areas. Furthermore, the size of a set of monitored RNCs is a dimensioning problem. As in case of the mobility agent and the BB, there is a trade-off between the scalability of the WQB (directly related to its domain size) and the overall QoS provided. For instance, the bigger the routing area is, the more alternative QoS paths can be offered nonetheless more signalling overhead is generated.

IV. Conclusion

In this paper, a pre-handover signalling for QoS-aware path selection has been presented for a B3G framework, which is compliant with 3GPP specifications. Furthermore, in order to facilitate the communication between the BB in charge of the IP QoS, a new entity (WQB), encompassing CRRM functionalities, has been defined. Finally the paper arises the scalability issues of the centralised architecture upon which the QoS-aware path selection mechanism relies. In order to analyse the scalability issues, cross-issues between IP micromobility and QoS routing are identified, and a more scalable architecture is proposed based on the use of hierarchical QoS routing areas in conjunction with a hierarchy of BBs.

Acknowledgements

This work has been performed in the framework of the EU funded project EVEREST. The authors would like to acknowledge the contributions of their colleagues from the EVEREST consortium.

References

[1] X. Fu, H. Karl, and C. Kappler, "QoS-Conditionalized Handoff for Mobile IPv6," *Networking2002, LNCS*, vol. 2345, pp. 721–730, 2002.

[2] J. Manner and M. Kojo, "Mobility related terminology," IETF RFC 3753, June 2004.

[3] J. Kakishima *et al.*, ""Plug and Access" for heterogeneous radio access environment," *WWRF10 New York*, Oct. 2003. [Online]. Available: http://www.wireless-world-research.org

[4] B. Moon and H. Aghvami, "Seamless Switching of RSVP Branch Path for Soft Handoff in All-IP Wireless Networks," *IEICE TRANS. COMMUN.*, vol. E86-B, no. 6, pp. 2051–2055, June 2003.

[5] N. Nafisi *et al.*, "Extending QoS policy-based mechanisms to B3G mobile access networks," *IST Mobile Communications Summit*, June 2004.

[6] J. Laiho and D. Soldani, "A policy based Quality of Service Management System for UMTS Radio Access Networks," *WPMC*, Oct. 2003.

[7] D. Goderis, D. Griffin, C. Jacquenet, and G. Pavlou, "Attributes of a Service Level Specification (SLS) Template," IETF Draft, Oct. 2003.

[8] H. Soliman *et al.*, "Hierarchical Mobile IPv6 mobility management (HMIPv6) <draft-ietf-mipshop-hmipv6-04.txt>," IETF Draft, Dec. 2004.

[9] K. Ceszei, N. Georganopoulos, Z. Turyani, and A. Valko, "Evaluation of the BRAIN Candidate Mobility Management Protocol," *IST Mobile Communications Summit*, pp. 889–895, Sept. 2001.

[10] A. Campbell, J. Gomez, S. Kim, and C. Wan, "Comparison of IP micromobility protocols," *IEEE Wireless Communications*, pp. 72–82, Feb. 2002.

[11] J. Moy, "OSPF version 2," IETF RFC 2328, 1998.

[12] Z. Zhuang, Z. Duan, and Y. Hou, "On scalable design of bandwidth brokers," *IEICE Transactions on Communications*, vol. E84-B, no. 8, pp. 2011–2025, Aug. 2001.

[13] E. Nikolouzou, G. Politis, P. Sampatakos, and I. Venieris, "An adaptive algorithm for resource management in a differentiated services network," *IEEE ICC*, pp. 2105–2109, June 2001.