

Integrating QoS Support in TeleMIP's Mobility Architecture

Archan Misra, Subir Das,
 Anthony Mcauley, Ashutosh Dutta
 Telcordia Technologies, Inc.,
 445 South Street,
 Morristown, NJ 07960, USA
 archan@research.telcordia.com
 subir@research.telcordia.com
 mcauley@research.telcordia.com
 adutta@research.telcordia.com

Sajal K. Das
 Center for Research in Wireless
 Mobility and Networking (CReWMaN)
 Department of Computer Science
 and Engineering
 The University of Texas at Arlington
 P.O. Box 19015, Arlington, TX 76019-0015, USA
 das@cse.uta.edu

Abstract—The paper describes a Differentiated Services-based QoS architecture for next-generation wireless networks. The architecture is based on the two-level TeleMIP hierarchical mobility management scheme and integrates Bandwidth Broker-based admission control and resource provisioning for mobile nodes. The TeleMIP architecture is extended to satisfy the QoS requirements of a mobile node, while requiring it to specify its traffic profile only when it first moves into a domain. The paper explores alternative approaches for dynamically assigning Mobility Agents to a mobile node and evaluates their suitability for different service differentiation models.

I. INTRODUCTION

In this paper, we outline alternative approaches being considered for maintaining Quality-of-Service (QoS) guarantees in the TeleMIP (Telecommunications-Enhanced Mobile IP) mobility management architecture. TeleMIP [1], [4] has been designed as a hierarchical IP-based mobility management scheme for 3rd and 4th generation IP-based cellular networks. It uses Mobile IP [2] for global mobility management and the Intra-Domain Mobility Protocol (IDMP) [5] to manage mobility within the domain. IDMP's mobility infrastructure specifies the use of special nodes, called Mobility Agents (MA), which provide a mobile node (MN) with a globally reachable, stable point of attachment inside a domain. TeleMIP's two-level hierarchy reduces the latency of intra-domain updates, lowers the frequency of global signaling traffic, promotes efficient use of the IPv4 address space and supports the use of multiple mobility agents for redundancy and load balancing.

With the emerging and anticipated deployment of higher bandwidth packet-based technologies, such as GPRS [19] and UMTS2000 [18], cellular networks will not only support users with different bandwidth guarantees but also transport multimedia traffic with diverse QoS constraints. End-to-end QoS guarantees for such mobile users can only be ensured when consideration of such performance bounds are integrated into the mobility management architecture. TeleMIP's proposed approach

for supporting QoS guarantees is based on a hierarchical application of the Differentiated Services [9] approach, with QoS bounds decomposed into separate global and intra-domain components. Since node mobility can lead to rapid changes in the intra-domain traffic paths, intra-domain QoS is assured by a dynamic resource provisioning architecture. The intra-domain provisioning technique relies on the use of a centralized Bandwidth Broker (BB) to dynamically provision aggregate capacity between an MA and the nodes at the wireless boundary.

The paper first presents the Bandwidth Broker-based architecture for intra-domain admission control and resource provisioning. Whenever a mobile node (MN) first enters a domain, it is assigned one or more Mobility Agents (MA). The MN communicates its QoS requirements to the appropriate MA, which then negotiates with the Bandwidth Broker to perform admission control and, if necessary, resource provisioning within the domain. We also outline how IDMP's intra-domain registration messages have been extended to include additional QoS-specific information and specify additional IDMP control messages between an MA and candidate Subnet Agents (SA), located at the wireless boundary of the cellular domain. Such messages allow the QoS characteristics of an MN, such as the negotiated traffic rate and the specified conditioner parameters, to be transferred to the appropriate ingress nodes as the MN moves within the domain. TeleMIP's QoS architecture does not, therefore, require the MN to repeat QoS signaling and negotiation over the wireless interface during intra-domain movement.

A key feature of TeleMIP is the use of multiple mobility agents within the domain, with an MN being assigned an MA via a dynamic load-balancing algorithm. We subsequently investigate the appropriateness of alternative MA-assignment approaches for different service differentiation scenarios. For the less advanced differentiation model, where users are distinguished simply by different traffic rates, we suggest an assignment scheme whereby a specific MA handles only users belonging to a single rate class. Such an approach simplifies the admission control and provisioning functionality and also guards against possible bottlenecks at the MA in the forwarding path. For the more advanced differentiation model, where individual MNs have traffic belonging to multiple traffic classes, possibly

with diverse QoS guarantees, we discuss two MA-assignment techniques. In one approach, an MN is assigned multiple MAs, each managing mobility only for a specific traffic class. While such an approach leads to greater signaling overhead, it does provide the benefit of differential mobility management for individual classes and avoids complicating the packet forwarding process at the MA. In the alternative approach, a single MA handles traffic for all the service classes belonging to an MN. We shall see that, while reducing the signaling load, this scheme imposes additional classification functionality at MA and may lead to greater dynamicity in the resource provisioning architecture.

II. SCALABLE QoS IN MOBILE ENVIRONMENTS AND RELATED WORK

In this section, we present an overview of the TeleMIP architecture, as well as outline the two conventional approaches towards service differentiation and QoS provisioning in the Internet. We also briefly describe another interesting approach that has been recently proposed for managing QoS guarantees in future wireless IP networks.

A. Basic TeleMIP Architecture

TeleMIP was introduced in [1] as a two-level hierarchical mobility management architecture for next-generation cellular networks. Under TeleMIP, networks are partitioned into *mobility domains*, with each domain consisting of multiple subnets. A mobile node (MN) is identified with two care-of addresses. The global care-of address (GCOA) remains unchanged as long as the MN changes subnets within a domain, while the local care-of address (LOCA) identifies the MN's precise point of attachment within the domain, and accordingly changes with every change in subnet. To support this hierarchy, IDMP specifies a new functional entity called Mobility Agent (MA), which is similar to the Foreign Agents (FA) in conventional Mobile IP [2], but resides at a higher level in the network hierarchy. Global updates to the Home Agent (HA) indicate only the GCOA and are necessary only when the MN changes domains; intra-domain location updates contain the LCOA (possibly assigned by Subnet Agents (SA) present in every subnet) and are directed to the assigned MA. Packets from an external correspondent node (CN) are tunneled (either directly or via the HA) to the MN's GCOA. The MA intercepts these packets and then tunnels them to the MN's current subnet of attachment by using the LCOA. Additional details of the TeleMIP architecture and its advantages over Mobile IP for mobility management in next-generation cellular environments are provided in [1], [4], which also compare TeleMIP with alternative intra-domain mobility management approaches, such as Cellular IP [7] and HAWAII [8].

As stated earlier, TeleMIP essentially leverages Mobile IP for global (inter-domain) mobility management and IDMP for intra-domain mobility management. IDMP is specifically designed for intra-domain mobility with additional features [5] such as fast handoff and paging. The TeleMIP architecture proposes the use of distributed mobility agents and the assignment of MAs via dynamic load balancing algorithms. The basic TeleMIP architecture is only concerned with host connectivity and does not attempt to integrate QoS considerations in the mobility architecture. Figure 2 shows a logical layout of TeleMIP's functional

elements.

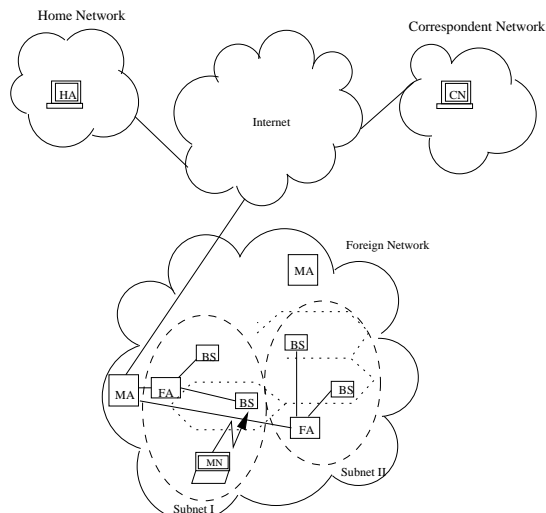


Figure 1: Abstract TeleMIP Functional Layout

B. Wireline IP QoS Architectures and Mobility

Architectures for supporting QoS guarantees on the Internet can be broadly classified into two groups, which differ in the technique for resource provisioning and the granularity of service differentiation. Both the approaches, however, suffer from drawbacks when used to provision resources for mobile hosts in wireless cellular environments.

The Integrated Services (Intserv) approach [11] uses the Resource Reservation Protocol (RSVP [13]) to explicitly signal and dynamically allocate resources at each intermediate node in the traffic path. Although not mandated, solutions using RSVP also assume a fine-grained differentiation model (where each flow, or at least each user) obtains individualized service guarantees and maintains separate reservation state at intermediate network elements. In the Intserv model, every change in an MN's point of attachment requires the generation of new RSVP messages to reserve resources on the new path. Even though modifications [17] can restrict such signaling only to the modified segment of the path, this approach incurs latency in QoS reestablishment, since signaling must be re-initiated and resources reserved (at least locally) at every change in an MN's point of attachment. Such an approach also requires the MN to initiate new RSVP messages for every change in subnet and can significantly increase the signaling load over the wireless interface.

The Differentiated Services (Diffserv) approach [9], on the other hand, uses a much coarser differentiation model where packets are classified into a relatively small number of classes at the network edge. This approach does not maintain any reservation state in intermediate elements and typically involves no dynamic signaling. QoS guarantees are provided implicitly by offline reservation of resources and appropriate marking of packets by traffic conditioners at the network edge. To provide performance guarantees in such environments, it is necessary to control the path of the offered traffic. Multi-Protocol Label Switching (MPLS) [15] is emerging as a standard approach to establish virtual paths for traffic flows in such an environment. The Diffserv approach to QoS typically involves pre-configuration of resources (for a specific class) along a path and the use of ad-

mission control algorithms at the edge to limit the offered load along that path. The rapid movement of an MN however leads to changes in the traffic path and makes it harder to devise effective admission control strategies under a static resource configuration regime.

The Bandwidth Broker (BB) architecture [3] has been proposed for introducing dynamic admission control and resource provisioning in the Diffserv architecture without requiring explicit end-to-end signaling. A BB maintains centralized information about the bandwidth reservations and current resource consumption for each traffic class. Network elements query the BB to determine if a new admission request can be satisfied without violating the associated performance bounds; the BB effectively replaces the end-to-end signaling model of RSVP with a per-domain decomposition approach. Integrating a BB-based dynamic resource allocation approach into the mobility management architecture can significantly improve the scalability of the QoS framework in cellular domains. TeleMIP's architecture for QoS support is based on the scalable Diffserv architecture and uses the Bandwidth Broker approach to dynamically allocate resources for MNs within the cellular domain.

C. Related Work

Relatively little work has been published on architectures that integrate QoS and mobility management in future wireless IP networks. [12] proposes an interesting QoS architecture, based on the Diffserv model, that splits the network into independent domains. The architecture uses a centralized QoS Global Server (QGS) to manage resources and perform admission control for all MNs in a domain. The QGS also interacts with the ingress nodes, called QoS Local Node (QLN), located at the ingress edge of the wireless interface to dynamically set up the conditioning and marking functions for individual MNs. Unlike our TeleMIP architecture, the approach does not separate intra-domain QoS management from global QoS management and does not consider the use of multiple MAs to handle multiple service classes.

III. TELEMIP QoS ARCHITECTURE AND IDMP ENHANCEMENTS

Based on our study of the limitations of the Intserv and Diffserv models, we believe that Intserv-based schemes, which require dynamic signaling of reservations along the new path, are likely to encounter significant scalability and latency problems in IP-based cellular domains. TeleMIP's QoS framework has, in fact, possesses the following two characteristics:

- An MN is required to specify its QoS requirements only during the initial registration in the domain. Traffic and service descriptors are transferred by elements within the wireline network to the appropriate ingress nodes during subsequent intra-domain movement.
- The Bandwidth Broker architecture is used to centralize the dynamic management of resources for different service classes. Mobility agents interact with the Bandwidth Broker to perform admission control and dynamic resource provisioning for mobile nodes.

TeleMIP splits the end-to-end QoS management into two distinct parts. The global QoS framework essentially uses the con-

ventional Diffserv/MPLS framework to manage resources between the MA and the global Internet, while the intra-domain QoS framework provides QoS support within the domain (between the MN and the MA). The MA essentially serves as a stable global "proxy" for the MN; our architecture requires all inbound (to the MN) and outbound (from the MN) packets to be routed via the MA. By adopting a two-level architecture, we localize the the dynamic component of network resource provisioning to nodes within the mobility domain. Global resources (between the MA and the Internet) will typically be reserved on an aggregate basis (trunk reservation), with the quantum of required resources being based on the operator's desired traffic handling capacity. Since the MAs are static nodes, the provisioning of global QoS is expected to be handled by the specification of conventional Service Level Agreements (SLAs) and static resource provisioning. Intra-domain resource provisioning, on the other hand, is more dynamic (on account of node mobility) and requires enhancements to IDMP as well additional functional elements in the TeleMIP architecture. The intra-domain resource allocation architecture is the focus of the rest of this paper.

A. Functional QoS Architecture

The functional architecture for QoS support in TeleMIP (via enhancements to the IDMP specifications) is shown in figure 2. We provide a brief description of the QoS-related functionality of each of the elements.

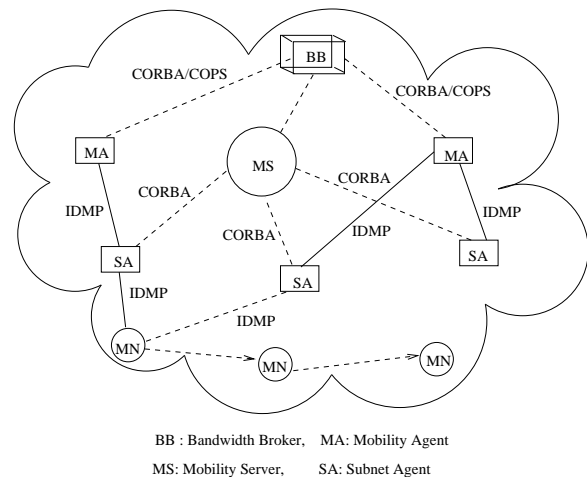


Figure 2: TeleMIP's QoS Signaling Architecture

- **MN:** The mobile node is responsible for indicating its QoS requirements (possibly for different classes) during its initial registration in the domain. Extensions are specified in the IDMP subnet-specific and intra-domain location update messages to characterize the MN's QoS requirements.
- **SA:** The Subnet Agents essentially behave as edge nodes for the cellular domain. For outbound traffic, the SA classifies packets into different classes, polices/ conditions them to ensure conformance to the negotiated profile and marks them appropriately. Since inbound traffic is typically policed at the MA (which acts as the logical ingress edge for the cellular domain), the SA simply forwards such traffic to the MN. The SAs are informed about an MN's traffic profile by the corresponding MA(s) using IDMP control mes-

sages. When an MN first moves into a domain, the serving SA is also responsible for querying the Mobility Server (MS) to obtain the identity of the appropriate MN(s). The relative resource allocation for different traffic classes on an SA can also be modified by directives from the Bandwidth Broker.

- **MA:** The MA delineates the boundary between the global and intra-domain QoS portions. For outbound traffic, the MA is responsible for collecting the aggregate traffic and conditioning it to ensure conformance to the statically-provisioned global traffic contract. For inbound traffic, the MA acts as the ingress point into the cellular domain. It is thus responsible for conditioning user traffic to ensure conformance to the negotiated profile and then marking it appropriately before forwarding to the LCOA. To perform admission control with a new MN, or to provision additional resources as existing MNs move within the domain, the MA interacts with the Bandwidth Broker. The MA will typically contact the BB to request aggregate bandwidth, rather than issuing reservation requests for individual MNs.
- **BB:** The Bandwidth Broker is the central entity for admission control and QoS provisioning in the cellular domain. Each domain typically has one BB, which maintains information about the resources allocated to different classes on different paths. Mobility agents typically request the BB for additional bandwidth (and other resources) on specific paths to accommodate additional traffic loads. The BB issues requests to internal domain nodes (including the SAs and the intermediate routers) to dynamically alter the resources allocated to different classes. The BB can also negotiate with BBs in other external domains for provisioning global QoS bounds.
- **MS:** The Mobility Server acts as the intelligent control entity that allocates mobility agent(s) to a new MN in the domain. When an MN moves into a domain, the serving SA requests the MS to assign one (or more, as we shall later) MA(s) to the MN. The MS is closely coupled with the BB and may often reside in the same network node. The key functional difference is that, unlike the BB, the MS is not concerned with the dynamic provisioning of resources, but rather, the dynamic allocation of MAs to best exploit the currently allocated resources.

While the interaction between the MAs and the SAs are specified as extensions to IDMP, we intend to use additional standard protocols for signaling between other entities. In particular, we are currently leveraging a CORBA-based implementation [16] for communication between the MAs and the BB. In the future, we plan to evaluate a COPS-based [10] interface for queries and responses between the MA and the BB. The BB also needs to communicate with the SAs, as well as other intermediate routers, to dynamically alter the reservation state at such nodes. In our initial prototype, we are hoping to use simple remote copy (rcp) to configure routers. In a more advanced implementation, we intend to use COPS as the interface between the BB (Policy Decision Point) and the SAs/routers (Policy Enforcement Points). Finally, an SA needs to interact with the MS to obtain the dynamic assignment of MAs to an MN. In our initial implementation, we intend to extend the CORBA-based in-

terface (between the MA-BB) to achieve this SA-MS signaling.

Figure 3 shows the example signaling flow for QoS support in the TeleMIP architecture. When the MN first moves into a new domain, it issues a local IDMP registration request, specifying the relevant QoS parameters, to the serving SA. The SA then contacts the MS (via the CORBA-based interface) to determine an MA that can handle the MN's QoS requirements. Once the MN obtains this MA address, it then performs an intra-domain registration with the designated MA. If the MA needs additional intra-domain resources to satisfy this request, it can contact the BB to request additional resources; the BB may subsequently use COPS or some other protocol to dynamically re-configure intermediate network elements. The MA can also modify the QoS requested by the MN in its RegistrationReply message. The MA also uses IDMP control messages to establish the appropriate conditioning/marking state in one or more SAs. (As we shall see shortly, the MA may multicast the MN's traffic profile to other SAs as well.)

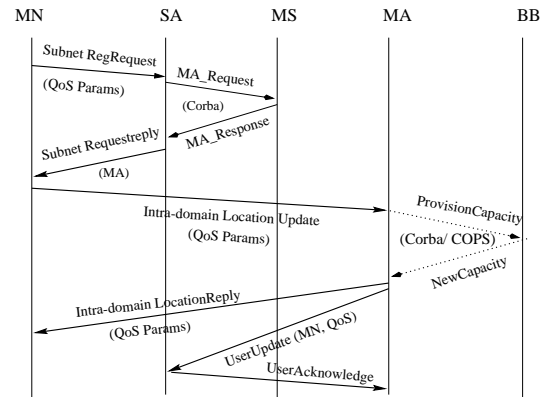


Figure 3: Example Signaling Flow for QoS in TeleMIP

B. Pre-Configuration vs. On-Demand Configuration

When an MN moves from its initial point of attachment within a domain, the new serving SA must be configured with the MN-specific traffic descriptor and conditioner parameters. Additionally, the MA may need to request the BB to reserve additional resources over the new path. This approach is called on-demand configuration and clearly, delays the continuous availability of QoS guarantees at the new location.

For services requiring lower latency in the re-establishment of QoS guarantees, the MA can adopt the pre-configuration approach. In this scenario, the MA can preemptively multicast the MN's traffic and QoS parameters to the set of neighboring (or candidate) SAs. (In the most extreme case, an operator may choose to broadcast such information to all SAs within the domain.) The new serving SA does not then need to interact with the MA to obtain QoS-related parameters and will then autonomously set up the traffic regulatory parameters as soon as the MN performs a subnet-specific registration, thereby significantly lowering the service interruption time. We plan to investigate both approaches in our experimental evaluation of this architecture.

C. Admission Control Choices

As the traffic path of an MN changes with every move, the network must ensure that adequate bandwidth is available over the new path to ensure conformance to the specified QoS bounds. The network operator can choose different admission control strategies for each service class to tradeoff the possibility of inadequate resource availability against higher network utilization. For example, given pre-configured bandwidth between an MA and the associated subnets (SAs), the MA could adopt a worst-case policy whereby the MA admits a new user only if every traffic path has adequate resources to meet the specified performance bounds, even if ALL users converge on the same subnet. Such an approach is extremely conservative, since MNs are likely to be distributed over all the subnets in the domain. On the other hand, an MA could also adopt alternative statistical approaches, whereby the MA uses mobility patterns to gauge the probability of service denial for different MNs. The MA could then admit a new user as long as the probability of service degradation (due to multiple users converging on a bottleneck path) did not exceed a specified bound. While such algorithms are likely to be service-specific, we intend to investigate such algorithms for a set of well-defined mobility patterns and traffic classes.

D. QoS for Mobile Nodes

In the next two sections, we shall study the utility of different MA-assignment techniques for different forms of service differentiation. It is thus important to clarify the various typical interpretations of the term ‘QoS’ in a cellular context.

In the simplest form, differentiation in IP-based cellular networks involves no QoS-related metrics at all; users are simply distinguished by differential bandwidth guarantees. Emerging packet-based wireless technologies (such as GPRS/CDMA2000) allow the capacity of the wireless link to be differentially allocated to different users. In this approach to differentiation, users subscribe to different customer profiles, which differ simply in the allocated traffic rate (traffic descriptor). The contract does not actually specify differential guarantees on specific QoS-related metrics, such as packet loss and delay. All packets from a user are treated uniformly within the network; service differentiation simply translates into the use of differential traffic rates during conditioning at the network edge. In the near term, operators can be expected to be more interested in this form of rate-based service differentiation.

Under a more advanced QoS model, users are distinguished not simply by the negotiated traffic rate, but also by the bounds associated with performance-related metrics, such as packet jitter or loss. An operator could be offering a variety of service classes, each possibly tuned for a specific application-profile and each possessing independent QoS bounds. The user profile for an individual MN could then have multiple bandwidth guarantees, one each for each separate class of traffic. Such a model requires nodes (such as SAs) at the network edge to classify packets from the same user into different classes. Furthermore, we shall see that the support of multiple service guarantees for a single user can complicate the TeleMIP signaling and resource allocation framework.

E. IDMP Message Extensions

In this sub-section, we briefly discuss the QoS-related extensions to IDMP. We simply present a logical description of each message and defer the detailed message formats to a separate draft. We first discuss the extensions to previously defined IDMP messages.

- The *Subnet RegistrationRequest* message has been extended to include QoS-specific extensions, such as the peak rate and desired packet delay. An MN typically sends such extensions only when it first moves into the domain; such information is used by the SA (and MS) to assign an appropriate MA to the MN.
- The *Subnet RequestReply* message has been extended to allow an SA to specify multiple MA addresses to an MN.
- Extensions to the *Intra – domain LocationUpdate* message allow an MN to specify the requested QoS parameters to the candidate MA.
- Extensions to the *Intra – domain LocationReply* message allow an MA to specify the assigned QoS parameters to an MN. The MA needs to include such extensions in its reply as it may not be able to satisfy the QoS parameters only requested by an MN.

Besides extending the existing IDMP messages, we have defined new messages for communication between the MA and the SAs. The *UserUpdate* message is sent by the MA to the relevant SAs. This message contains the QoS parameters assigned to an MN, as well as the MN’s permanent ID and its GCOA. A corresponding *UserAcknowledge* message has also been defined for communication from the SA to the MA.

IV. MA-ALLOCATION FOR USERS DIFFERENTIATED BY BANDWIDTH

In this section, we specify a candidate MA-allocation algorithm when users are distinguished simply based on their collective rate guarantees. All traffic from a user is assumed to belong to the same service class. As an example, consider a network where the subscriber is offered 3 different profile choices, Gold, Silver and Bronze with corresponding bandwidth guarantees of 192 Kbps, 64 Kbps and 16 Kbps respectively.

In this bandwidth-based differentiation scenario, the TeleMIP resource provisioning approach reserves an individual mobility agent (MA) for a specific subscriber class. MNs belonging to different subscriber classes are allocated to different MAs; of course, the architecture allows multiple MAs to be responsible for MNs belonging to the same traffic class. Furthermore, since all traffic from a particular MN belongs to the same subscriber class, each MN is associated with a single MA and is bound to a single GCOA. The MS simply uses the subscriber class information of an MN as an index to determine a candidate MA. Such an allocation strategy is shown in figure 4. MNs 1 and 4, which belong to service class Gold, are assigned *MA 1* by the MS, while MNs 2 and 3, which belong to service class Silver, are assigned *MA 2*.

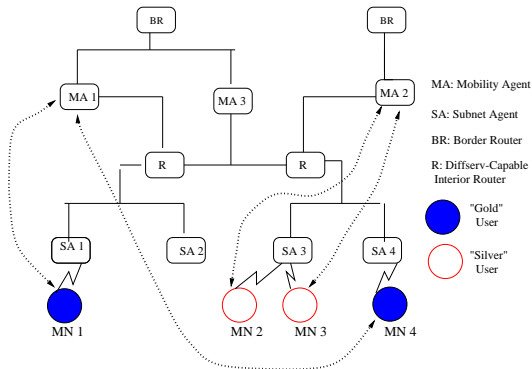


Figure 4: Example of ‘One Class-per MA’ Architecture

Such an MA assignment scheme is attractive for several reasons. Since each user associated with an MA has a common bandwidth rate, the MA can perform admission control for a new MN very easily. Given a specific level of reserved resources, the MN can simply determine the maximum permissible number of simultaneous users and reject a new request if the number of currently registered MNs equals or exceeds this threshold. If the MA needs additional resources on a specific path (via a request to the BB), it can determine the requisite capacity by directly considering the additional number of subscribers that it wishes to provision on this path.

This architecture also leads to scalable resource provisioning when each subscriber class not only has an explicit bandwidth guarantee, but an implicit service guarantee as well. For example, an operator may guarantee Gold users a maximum packet delay of 30 msec within the domain; Silver users can have a corresponding bound of 60 msec. Since each MA caters only to a single user class, it does not perform any additional classification for either inbound or outbound packets. For inbound packets, the MA simply tunnels the packet to the corresponding LCOA and marks the outer header with the PHB (per-hop behavior) corresponding to the associated traffic class. The ability to dispense with additional classification on both the inbound and outbound packet streams is very important. TeleMIP requires the MA to decapsulate and re-encapsulate every incoming packet; any further complexity induced by the classification mechanism could lead to delays in the packet forwarding process at the MA. For outbound packets, the MA simply marks the DS field in the packet headers appropriately. For example, an MA supporting Gold users may mark packets to indicate a request for the Expedited Forwarding (EF) [14] PHB, while an MA supporting best-effort users may choose only the default per-hop behavior from the global Internet.

Identifying an individual MA with a single user class also simplifies conditioning functionality at the subnet-level SAs. An SA does not need to classify each outbound packet from a user separately; it can simply mark outgoing packets based on the destination MA. For example, in our example in figure 4, packets destined for MN2 are marked with a higher priority, while packets marked for MN1 are marked with a lower priority.

V. MA-ALLOCATION FOR USERS WITH MULTIPLE QoS CLASSES

In a more advanced wireless architecture, traffic from an individual user could consist of packets belonging to different ser-

vice classes. Under this model, all VoIP packets from MN1 could be associated with traffic class 0 while all Web-based traffic could be mapped to traffic class 1. We consider two possible approaches in this case.

A. Multiple GCOA

In first approach, an MN is assigned multiple MAs, each serving traffic that belongs to a specific service class. This approach allows us to ensure that, as before, an individual MA caters only to a specific service class and consequently applies the same packet marking policy for all inbound (and also, outbound) packets. Thus, at least from the MAs perspective, this architecture offers the same provisioning and performance advantages enumerated earlier.

In this approach, when an MN first registers with a domain, the MS allocates it multiple MAs, one for each of requested service classes. The MN is informed of multiple MAs through the extensions specified in the IDMP *Subnet Request Reply* message and must perform an independent intra-domain location update with each MA separately. The MN is thus identified by multiple GCOAs, each corresponding to a different service class. Whenever the MN changes subnets, the MN must now perform multiple intra-domain location updates, one for each associated MA. The problem of QoS management for a single MN is thus effectively decomposed into a set of N independent QoS management problems, one for each of the N MAs that are associated with an MN. Figure 5 shows an example MA-allocation in this approach, where $MA 1$ and $MA 3$ manage mobility for service class 0, while $MA 2$ manages mobility for service class 1. MN 1 is associated with two MAs, $MA 1$ and $MA 2$, while MN 2 is also associated with two MAs, $MA 2$ and $MA 3$.

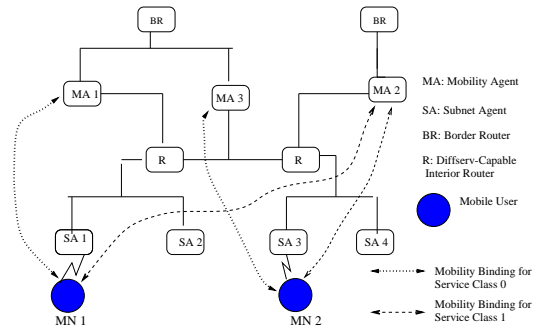


Figure 5: Multiple GCOA-based QoS Architecture

This architectural approach requires upgrades at the following IDMP nodes:

- **SA:** In addition to providing an MN with multiple MA addresses, the SA will now need to store multiple traffic profiles for a single MN. For outbound traffic (from the MN), the SA must first classify each packet into the appropriate traffic class and condition/ mark it using the class-specific traffic/ service descriptor.
- **MN:** In this multi-MA architecture, an MN essentially has to manage multiple global care-of addresses; each portion of its traffic stream is associated with a potentially different MA. Clearly, the MN will now need to issue multiple intra-domain location updates on every change in subnet, with each location update managing user mobility only for

a specific traffic class.

This approach offers MNs the significant benefit of being able to choose additional intra-domain mobility features, such as such as IDMP's fast handoff or paging support [4], on a per-class basis. For example, an MN may choose to activate fast handoff support only for VoIP traffic (class 0 handled by $MA 1$ in figure 5 and not for the Web-browsing traffic (class 1 handle by $MA 2$). The MN simply turns on the fast-handoff bit in its location update request to $MA 1$ and keeps it off in its signaling with $MA 2$.

This architecture is useful only if the global mobility management scheme can effectively handle multiple care-of address. TeleMIP thus requires Mobile IP to support multiple care-of addresses, with separate addresses being used for different traffic classes. While Mobile IP already allows an MN to register multiple care-of address with its HA, we are extending the global registration mechanism to associate each care-of address with a separate traffic class.

In comparison to the use of a single GCOA (discussed shortly), this architecture provides greater flexibility in the provisioning of network resources. However, this approach does lead to a larger signaling load, especially over the wireless interface, since the MN must perform N independent intra-domain location updates for every change in subnet.

B. Single GCOA

An alternative approach to supporting multiple traffic classes per user is to associate an MN with a single MA, as in conventional IDMP. For inbound traffic, the MA must now classifying packets into different service classes and appropriately mark them before tunneling them to the MN. Similarly, for outbound packets, the MA must mark packets for different service classes differently to associate them with different PHBs. This approach is functionally very similar to the classic Diffserv architecture, with the MA acting as a "logical edge" where packet classification and marking are performed. Since an MN is associated with only a single MA, it does not need to perform multiple intra-domain location updates on changing subnets. This approach thus clearly reduces the mobility signaling load over the wireless interface and is illustrated in figure 6.

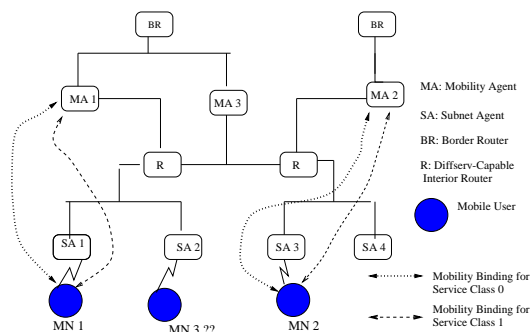


Figure 6: Single GCOA-based QoS Architecture

Although attractive at first glance, this approach can lead to several complexities, especially if a BB is absent in the domain and resource provisioning is done by purely static mechanisms.

- The need for additional classification by the MA for every incoming and outgoing packet significantly increases the processing complexity in the forwarding path. In the cur-

rent TeleMIP architecture, where the MA lies in both the data and control plane, such complexity could significantly reduce the forwarding capacity of an MA.

- Selective application of intra-domain mobility features, for specific traffic classes, is harder. As an example, if an MN activates fast handoff support, the MA will forward all packets (irrespective of the service class) via the proactive multicasting mechanism [4]. The MN, on the other hand, may desire fast handoff support only for VoIP traffic and not for non-real time Web traffic. A blanket forwarding technique may be undesirable in such a case, especially such fast handoff results in additional charges to the user. To permit selective activation of such mechanisms, we would need to specify additional signaling and processing functionality at the MA.
- Perhaps, most importantly, using a single MA to support multiple traffic types leads to a loss of flexibility in the support of multiple QoS bounds. Situations may arise in such cellular domains where a single MA is unable to satisfy the diverse QoS bounds of different service classes. To illustrate this case, consider figure 6 as before. Assume that the network is provisioned such that $MA 1$ and $MA 2$ can each support a maximum of 10 Mbps of traffic for class 0 and 20 Mbps of traffic for class 1. Also, assume that $MA 1$ is currently forwarding 9 and 18 Mbps of class 0 and 1 traffic respectively; similarly, $MA 2$ is forwarding 8 and 19.5 Mbps of traffic for class 0 and 1. If a new user ($MN 1$) now requests a service rate of 1.5 Mbps for each of the service classes, it should be clear no individual MA can satisfy both requests. A multi-MA approach, on the other hand, would associated the MN with two MA addresses, with class 0 traffic handled by $MA 2$ and class 1 traffic handled by $MA 1$. Constraining all user traffic onto a single path (a single MA) may be a significant limitation, especially in the cellular environment where node mobility can lead to high unpredictability in the link traffic loads. At the very least, a single MA-approach appears likely to cause more frequent invocation of the BB to provision additional resources along a specific path. This dynamic provisioning strategy can encounter unexpected instabilities if the frequency of such allocation requests is higher than the periodicity with which the BB determines the network state.

VI. CONCLUSION

This paper presented an architecture for supporting QoS guarantees in next generation IP-based wireless networks. Our approach integrates TeleMIP's two-level hierarchical mobility management with a Bandwidth Broker-based approach for dynamic resource provisioning and admission control. Service differentiation is based on a hierarchical Diffserv architecture, with dynamic bandwidth allocation used to provision resources for mobile users as they move within a domain. Subnet agents, located at the boundary interface, enforce specified traffic descriptors for individual users and mark packets to provide service differentiation. The Bandwidth Broker maintains a snapshot of the allocated resources within the domain and can dynamically reconfigure routers and edge nodes to alter the bandwidth allocated to different service classes. We indicated the various in-

interfaces/ protocols used in our architecture for communication between different TeleMIP nodes.

We have specified extensions to current IDMP messages that enable an MN to specify its QoS requirements and that allow an MA to transfer negotiated parameters to new ingress nodes (SAs), thus avoiding the need for repeated QoS signaling on every MN movement. We also investigated the use of dynamic MA-assignment algorithms to support service differentiation in a scalable manner. When users are distinguished merely by different bandwidth guarantees, we discussed the advantages of an allocation strategy whereby an individual MA caters to MNs only belonging to a specific rate class. In scenarios where a single user has traffic belonging to multiple traffic classes, we explored the possibility of either using different MAs to handle different classes or using a single MA to handle all the traffic classes. While the single-MA architecture results in lower signaling load over the wireless interface, the multi-MA architecture keeps the functionality of an MA simple and hence, avoids potential bottlenecks in the forwarding path.

Although this paper presents several scalable architectural alternatives for introducing QoS support in IDMP, we have yet to perform serious evaluation of each of these alternatives in prototype testbeds. We are currently working on completing our Linux-based implementation of TeleMIP and IDMP and plan to shortly integrate our implementation with that of the Bandwidth Broker. Also, while dynamic resource reservation techniques (a la Intserv) are unlikely to prove viable for cellular environments, it is necessary to practically investigate and demonstrate the limitations of this approach.

VII. ADDENDUM

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.

REFERENCES

- [1] S. Das, A. Misra, P. Agrawal and S. K. Das, "TeleMIP: Telecommunication Enhanced Mobile IP Architecture for Fast Intra-Domain Mobility", IEEE PCS Magazine, August 2000.
- [2] C. Perkins, "IP Mobility Support", RFC 2002, October 1996.
- [3] Internet-2 Qbone Bandwidth Broker, available at <http://www.merit.edu/working.groups/i2-qbone-bb/>.
- [4] A. Misra, S. Das, A. Dutta and S. K. Das, "Supporting Fast Intra-Domain Handoffs with TeleMIP in Cellular Environments", under submission.
- [5] A. Misra, S. Das, A. McAuley, A. Dutta and S. K. Das, "IDMP: An Intra-Domain Mobility Management Protocol using Mobility Agents", draft-mobileip-misra-idmp-00.txt, July 2000, Work in Progress.
- [6] S. Das, A. Misra, A. McAuley, A. Dutta and S. K. Das, "A Generalized Mobility Solution Using a Dynamic Tunneling Agent", to appear in proceedings of ICCCD2000, December 2000.
- [7] A. Campbell, J. Gomez, C-Y. Wan, S. Kim, Z. Turanyi and A. Valko, "Cellular IP", draft-ietf-mobileip-cellularip-00.txt, January 2000, Work in Progress.
- [8] R. Ramjee, T. La Porta, S. Thuel and K. Varadhan, "IP micro-mobility support using HAWAII", draft-ramjee-micro-mobility-hawaii-01.txt, July 2000, Work in Progress.
- [9] S. Blake, D. Black, et al, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [10] J. Boyle, et al, "The COPS (Common Open Policy Service) Protocol", RFC 2748, January 2000.
- [11] R. Braden, et al, "Integrated Services in the Internet Architecture: An Overview", RFC 1633, June 1994.
- [12] J. Chen, et al, "A QoS Architecture for Future Wireless IP Networks", to appear in Proceedings of PDCS 2000, November 2000.
- [13] R. Braden, et al, "Resource ReSerVation Protocol (RSVP): Version 1 Functional Specification", RFC 2205, September 1997.
- [14] V. Jacobson, K. Nichols and K. Poduri, "An Expedited Forwarding PHB", RFC 2598, June 1999.
- [15] E. Rosen, A. Vishwanathan and R. Callon, "Multiprotocol Label Switching Architecture", draft-ietf-mpls-arch-07.txt, July 2000, Work in Progress.
- [16] R. Talpade, et al, "A Bandwidth Broker Architecture for VoIP QoS", under submission.
- [17] A. Terzis, M. Srivastava and L. Zhang, "A Simple QoS Signaling Protocol for Mobile Hosts in the Integrated Services Internet", Proceedings of INFOCOM 1999, March 1999.
- [18] D. McFarlane, et al, "IMT-2000 Standards: Service Provider's Perspective", IEEE Personal Communications, vol.4, August 1997.
- [19] P. Rysavy, "General Packet Radio Service (GPRS)", GSM Today Online Journal, September 1998.