7-2001

# Application-Centric Analysis of IP-based Mobility Management Techniques

Archan MISRA
*Singapore Management University*, archanm@smu.edu.sg

Subir DAS
*Telcordia Technolgies*

Prathima AGRAWAL
*Telcordia Technologies*

**DOI:** https://doi.org/10.1002/wcm.21

## Citation

MISRA, Archan; DAS, Subir; and AGRAWAL, Prathima. Application-Centric Analysis of IP-based Mobility Management Techniques. (2001). *Journal of Wireless Communications and Mobile Computing*. 1, (3), 313-328. Research Collection School Of Information Systems.
**Available at:** https://ink.library.smu.edu.sg/sis_research/664

Published in Wireless Communications and Mobile ComputingOpen Access, Volume 1, Issue 3, July 2001, Pages 313-328.
http://doi.org/10.1002/wcm.21

# Application-Centric Analysis of IP-based Mobility Management Techniques

Archan Misra Subir Das Prathima Agrawal
Telcordia Technologies, 445 South Street
Morristown, NJ, USA
{archan, subir, pagrawal}@research.telcordia.com

*Abstract*— **This paper considers three applications- VoIP, mobile Web access and mobile server-based data transfers- and evaluates the applicability of various IP-based mobility management mechanisms. We first survey the features and characteristics of various IP mobility protocols, such as MIPv4, MIPv6, MIP-RO, SIP, CIP, HAWAII, MIP-RR and IDMP, and then evaluate their utility on an application-specific basis. The diversity in the mobility-related requirements ensures that no single mobility solution is universally applicable. We recommend a hierarchical mobility architecture. The framework uses our Dynamic Mobility Agent (DMA) architecture for managing intra-domain mobility and multiple application-based binding protocols for supporting inter-domain mobility. Thus, we recommend SIP as the global binding protocol for VoIP applications and MIPv4/MIPv6 as the global binding mechanism for the mobile server scenario.**

*Keywords*— **Mobility management, VoIP, Web access, Mobile IP, SIP, HAWAII, Cellular IP, IDMP, DMA, hierarchical.**

## I. INTRODUCTION

A variety of IP-based mobility management strategies have been proposed for next-generation IP-based cellular networks. In general, individual mobility management solutions, such as the cellular mobility management schemes for circuit-switched voice [1] or Mobile IP [2] based solutions for IP-based data networks, are largely designed for a specific application set. Next generation (3/4G) cellular networks, however, aim to offer a uniform, *application-independent* access and mobility infrastructure. Designing a mobility management solution that supports the diverse requirements of heterogeneous applications is a significant challenge- we shall see that no single approach to IP mobility applies uniformly across applications.

The key benefit of an IP-based mobility solution is its *independence of the underlying link and physical layers*. An IP-based mobility solution for packet-based cellular networks would not only enable seamless integration with the conventional Internet, but also insulate the management layer from future changes in underlying link and (largely proprietary) physical layer technologies. Any IP-based mobility scheme essentially resolves the location of a mobile node (MN) only up to the granularity of a *subnet*; additional link-layer mechanisms must be used to forward packets to the individual host. The interpretation of a subnet in the cellular context is architecture/operator dependent. Additional link-layer specific solutions may indeed be needed to manage the movement of an MN within a single subnet; such *micro-mobility* management approaches are however outside the scope of this paper.

In this paper, we evaluate various proposals for IP-based mobility management from an application-centric perspective. In particular, we consider three key application types:

- *Voice-over-IP (VoIP):* This refers to applications involving packet-based interactive IP telephony.
- *Mobile Web browsing/'Mobile-ecommerce':* This refers to the whole class of server-centric data applications, where a mobile user usually initiates interaction with static Web servers.
- *Mobile Server-based Data Transfer:* This is a more advanced data transfer scenario, where the MN is itself a server and a *correspondent node* (CN) is the initiator of a data transfer session.

We investigate the suitability of alternative mobility management strategies for each of these application sets. Literature on mobility protocols typically focuses on the capabilities and features of the individual protocols; application-centric analysis, on the other hand, is useful as it determines the utility of such protocols based on the capabilities *required* by an application. The three application types considered here exhibit significant differences in their mobility-related requirements. For example, VoIP applications require smooth and seamless handover during a session and also require mobile users to be continuously reachable. On the other hand, such seamless handover or continuous reachability is not very important for applications based on mobile access to Web, which primarily require roaming-based ubiquitous connectivity.

We shall first examine each of the three applications to determine their individual mobility-related requirements. Given the layered nature of the IP protocol stack, it is possible to implement mobility management solutions at various layers. Most proposed mobility solutions manage mobility (e.g., [2], [3]) (usually through dynamic addressing and seamless packet redirection) at either the network or application layers. IP-based mobility management proposals also use either a flat or a hierarchical mobility architecture, each of which has its inherent advantages and drawbacks. We shall see how the choice between a network layer and application layer solution, and between a flat and a hierarchical management architecture, depends on the specific characteristics of each individual application.

The rest of the paper is organized as follows. Section II shows how mobility management solutions may be classified on the basis of their functional and operational characteristics. Such a functional decomposition of each proposed solution allows us to devise hybrid architectures, which combine logically distinct elements from multiple mobility management schemes. In section III, we consider the mobility related service requirements of each of our applications in detail, especially in the context of the generic mobility feature set presented in section II. In section IV, we provide an overview of the various proposals for IP-based mobility management and analyze their support for the various mobility-related features. Such a feature-based analysis leads naturally to the contents of section V, which studies how the requirements of each application can be satisfied by devising a hybrid mobility architecture that combines aspects of the various proposed mechanisms. In general, we shall motivate the use of a two-level hierarchical IP mobility architecture, based on a common intra-domain mobility management architecture and characterized by application-specific global mobility solutions. Finally, section VI concludes the paper.

## II. CHARACTERISTICS FOR EVALUATING ALTERNATIVE MOBILITY SOLUTIONS

While different IP-based mobility solutions differ in their details, they all implement certain fundamental mobility-related functions. In this section, we first describe each of these basic functions; we shall later see that individual application scenarios may or may not require all of these functions. We then describe the various mobility-related features that are relevant from an application standpoint.

### A. Mobility Support Functions

Broadly speaking, the term "IP mobility management" refers to three logically distinct functions. Different approaches use different protocols and mechanisms to implement these three fundamental operations:

- *Configuration:* Configuration refers to the mechanism used to provide an MN an IP address that is topologically consistent with its current point of attachment. This is required to ensure mobility solutions without modifying (at least globally) IP's destination address-based packet routing mechanism. The configured address (or addresses) is typically transient (since it potentially changes as the MN moves), and is called a *care-of address* (CoA)[1] in this paper. A variety of protocols, such as DHCP [4], PPP [5], DRCP [6], Mobile IP [2] provide such configuration functionality.
- *Registration:* Node configuration must be followed by user registration. Registration mechanisms allow a foreign network to authenticate the user and to determine the user's service requirements (e.g., QoS needs). A variety of protocols, such as Mobile IP [2], RADIUS [7], DIAMETER [8], provide various forms of registration functionality.
- *Binding Update:* To allow an MN to be reachable, the MN must communicate its current address (CoA) to other nodes. Binding update mechanisms transmit this CoA either directly to the CN or to agent nodes that act as centralized servers or packet re-directors. Potential binding mechanisms include Mobile IP *Registration* messages or *Binding Updates* (BUs) or SIP [3] *Re-INVITE* messages.

### B. Features of Mobility Support

Feature-based classification of different mobility management techniques is another useful approach. We shall later see that each of the applications considered here requires support for one or more of these features. Accordingly, the suitability of a particular mobility scheme for a specific application can be studied by analyzing whether that solution supports the relevant set of features. The following set of characteristics can be used to distinguish between the different IP mobility management mechanisms:

- *Explicit Address Resolution/ Transparent Mobility:* Mobility solutions differ in whether they allow a mobile user (or node) to retain a permanently assigned address in a location independent fashion. Many applications do not have an explicit query phase; the CN simply transmits packets to a permanently assigned address and is unaware of any change in the MN's point of attachment. Many other applications, such as VoIP, however have an explicit signaling mechanism, where the CN dynamically resolves the MN's currently configured IP address. Mobility management techniques that do not provide transparent packet redirection are appropriate only for such applications, which can dynamically bind addressing information during an explicit query phase.
- *Ubiquitous Mobile Node Reachability:* Most of the proposed mobility management schemes use some binding protocol or update mechanism, whereby the MN keeps appropriate network nodes aware of its currently configured address. A binding protocol essentially ensures that the MN is *always locatable*: a CN can initiate communication. Applications such as VoIP clearly require such bidirectional locatability. However, future mobility scenarios may often involve applications (e.g., mobile access to the Web) where sessions are initiated by the MN alone. A binding protocol may be unnecessary in such cases.
- *Fast and Reliable Handoffs:* Any mobility solution typically gives rise to transients when an MN changes it's current point of attachment. Incoming (*in-flight*) packets forwarded to the old point of attachment could be lost until the MN has refreshed its configuration information and issued appropriate binding updates. While current cellular systems use elaborate handoff mechanisms to minimize this transient loss, IP-based mobility management mechanisms have traditionally not supported fast handoffs for seamless intra-session mobility. An application could require support for either low-latency or low-loss intra-session handoff mechanisms.
- *Paging:* Paging support reduces an MN's power consumption, since an idle MN does not need to perform address configuration and issue binding updates on every change in the point of attachment. Paging support has been an intrinsic part of the traditional cellular mobility management architecture. In contrast, many IP-based mobility solutions do not provide paging functionality and require the MN to re-configure at every change in the subnet of attachment. While paging is typically a device-specific (rather than an application-specific) requirement, the choice of application does affect the applicability of alternative paging schemes. The need to broadcast a solicitation to resolve the ambiguity in the MN's current point of attachment introduces delay in the packet forwarding process. Different applications

---

[1] Throughout this paper, we use the term CoA to refer to any such transient, mobility-related address. This is different from the traditional Mobile IP-specific connotation of the term "care-of address".

may possess different thresholds for the acceptable delay; clearly, such thresholds could impact the applicability of alternative mobility management schemes. Moreover, nodes running applications that have solely MN-initiated sessions may simply forego paging, since continuous reachability is not needed.

## III. APPLICATION-SPECIFIC MOBILITY REQUIREMENTS

In this section, we discuss the mobility-related requirements individually, for each of our three chosen applications. We shall see that the required features and the acceptable bounds on various performance metrics can indeed vary greatly across different application sets.

### *A. VoIP Mobility Requirements*

Unlike other traditional IP applications, conventional telephony has two separate planes, *control* and *transport*, that differ in their service requirements. Accordingly, an IP-based mobility management architecture must support different performance objectives for control and transport packets.

- *Control Traffic:* Control or call-setup messages are typically less delay-sensitive: the average latency for a call setup (within the continental US) is around $2.5$ secs [9]. The control channel is however, expected to exhibit high reliability; the mobility management mechanism should accordingly ensure that in-flight control packets are re-directed to a MN's current point of attachment with negligible loss.
- *Data Traffic:* Interactive voice applications require the 'lip-to-ear' one-way delay to be bounded to less than $150$ msecs [10]. While voice packets are not extremely loss sensitive (most codecs can tolerate up to $\sim 1-2\%$ loss rates), such losses should not result in loss of consecutive voice samples. Accordingly, an ongoing voice session requires fast handoffs, with the service interruption due to mobility-related transients ideally restricted to $\sim 20-30$ msecs.

We can thus see that *fast handoff for voice packets* and *low in-flight loss for control packets* are key requirements of a mobility architecture for VoIP applications.

Voice applications would ideally require both user and terminal mobility. Any voice conversation is usually preceded by a call setup phase, where the call (session) may potentially be re-directed to the callee's current mobile node (or equivalently, the current IP address). Such explicit signaling allows the MN (user) to be associated with a dynamically associated IP address. Furthermore, VoIP uses UDP as the underlying transport protocol and hence, can dynamically change the destination address associated with an ongoing connection. Accordingly, the MN does not need to maintain a constant IP address (CoA), even within a single session. Tunneling or other packet re-direction mechanisms, which make any change in the MN's CoA transparent to the CN, are accordingly not mandatory.

VoIP users require continuous reachability. Accordingly, the mobility management mechanism must include a global binding mechanism that ensures that the MN's current location is available at some centralized database. To match the significant savings in power consumption provided by current cellular paging schemes, the IP-based mobility solution must also provide a flexible paging mechanism. The paging operation for an idle MN is typically invoked by the first packet from a CN. Since such a packet is typically a control (call setup) packet, the paging latency must be bounded to ensure conformance to the acceptable bounds on the call setup delay.

Finally, since VoIP traffic is typically periodic, non-adaptive and has a fixed traffic rate, the mobility mechanism must provide seamless QoS support. Most VoIP codecs generate packets periodically at a constant rate during a talk spurt and employ silence suppression to reduce bandwidth consumption during a silent spurt. Provisioning based on aggregate voice traffic will realize significant multiplexing gains, since studies show that such aggregate traffic is fairly smooth. Accordingly, the QoS support mechanism need not be concerned with dynamically reserving resources for individual flows. Rather, the mobility architecture should provide a mechanism to reserve aggregate resources for voice traffic and dynamically change the reservation levels on a specific path as the constituent voice sources change their point of attachment.

### *B. Web Browsing/M-Commerce Applications*

While support for VoIP will indeed be critical to the successful deployment of IP-based cellular networks, it is important to realize that Web-based data retrieval applications (maps, stock information, weather reports, email) have been the primary driver for the introduction of packet-based wide area connectivity. With the predicted boom in mobile e-commerce applications, it is important to devise a scalable mobility architecture that supports the requirement of such applications.

A key feature of this application set is that it is primarily *pull-based*, with the MN (user) retrieving data from (usually static) servers. Such MN-initiated applications do not require continuous locatability; accordingly, there is no need for a binding protocol that stores the MN's current CoA in centralized servers for possible retrieval by CNs. In contrast to VoIP, such browser-based data applications typically do not have hard bounds on the delay or jitter experienced by individual packets. As such applications use TCP as the transport protocol for reliability, they are more sensitive to packet losses. Even moderately high loss rates can lead to undesirable TCP transients (such as slow start) and considerably increase the application response times. Accordingly, the mobility management mechanism must provide for low-loss handoffs during an ongoing session.

Such TCP-centric applications require network-layer mobility management mechanisms that provide transparency to the overlying transport layer. To avoid the need to reset an ongoing TCP connection, the mobility management technique must implement a re-direction mechanism that allows a CN to address all packets for a single session to a single destination address. However, this need for a stable intra-session address must be distinguished from the need to maintain a permanently assigned global address. In fact, for such pull-based applications, the MN does not need to maintain a unique permanent address and can indeed use different addresses for different sessions. It should also be noted that application layer protocols such as HTTP [11] can transparently deal with address re-configuration by automatically initiating a new TCP connection. However, such connection re-establishment carries performance penalties and should preferably be avoided.

While such TCP-based applications are inherently adaptive, mobile users may indeed like differential bandwidth guarantees. Since different users may subscribe to different QoS guarantees, the mobility management mechanism must allow an MN to specify its QoS profile as part of the registration process. The mobility framework should also preferably minimize the need for explicit QoS re-negotiation at every change in an MN's subnet of attachment.

### C. Reliable Data Transfer for Mobile Severs

Client-server based applications in current networks are characterized by *largely static* servers: since the address of the server is essentially constant, clients can often be manually configured with the corresponding addresses. This connectivity paradigm may, however, change as mobile nodes, such as laptops and PDAs, become progressively more powerful, and as packet-based high speed cellular networks become more widely deployed. Future scenarios could include the advent of *mobile servers*, whereby data and other services are resident on mobile nodes. A popular example of this might be an MN that acts as a repository for publicly available audio and video files; CNs retrieve such files for local use as needed.

To allow seamless support for clients that are statically configured with a fixed server address, the mobility solution must allow CNs to communicate with the MN (server) using a fixed permanent address. Such support is essentially provided by defining transparent re-direction mechanisms that re-route packets, addressed to the MN's permanent address, to its current point of attachment. As in the mobile Web access case, any changes in the MN's CoA should be transparent to the CN. This is especially important in the bulk-transfer case, since bulk-transfer applications such as ftp, do not provide for automatic re-establishment of TCP connections. Since such applications are, from an MN's viewpoint, *push-based* (client nodes initiate connection requests), a global binding protocol that ensures continuous locatability is clearly necessary.

Fast handoff and paging support is probably less critical in such application environments. Bulk data transfer applications are not sensitive to the delay or jitter experienced by individual packets. Power conservation may also not be important for conventional hosts, such as laptops. However, the mechanism should minimize the loss of in-flight packets due to mobility-related transients, since such losses are very detrimental to TCP performance. As in the mobile Web access scenario, the mobility framework must allow individual nodes to specify their individual QoS profiles. There is however an important, but slightly less obvious, distinction between the mobile Web access scenario and the case of a mobile server: a server is more likely to have multiple concurrent sessions, each catering to a different client and each with possibly different QoS characteristics (based on the individual CN's QoS requirements). It is thus absolutely vital to ensure that the (server) MN does not need to perform expensive QoS renegotiation for every session on every change in subnet.

Table I provides a synoptic view of the mobility-related features required by each of our three applications.

| **Feature** | **VoIP** | **Web Browsing** | **Mobile Server** |
|---|---|---|---|
| Continuous Reachability | Yes | No | Yes |
| Permanent IP Addressing | No | No | No |
| Fast Intra-Session Handoff | Yes | No | No |
| Low Loss Handoff | No | Maybe | Yes |
| Paging Critical | Yes | No | No |
| Paging Latency Acceptable | Maybe | Yes | Yes |

Table I: Application-Specific Mobility Features

## IV. ALTERNATIVE MOBILITY MANAGEMENT SCHEMES

In this section, we survey the various proposals for IP-based mobility management. Since most of these proposals have been extensively discussed in literature, we shall provide only a high level overview of each scheme. The advantages and drawbacks of each proposal will be evident when, in the next section, we consider the suitability of the solutions for each of our three target application scenarios. We shall also show how a hierarchical architecture makes all such proposals more scalable and robust and allows us to use a common intra-domain mobility solution for all the three application sets under consideration. We shall then introduce and provide a brief functional description of our Dynamic Mobility Agent (DMA) architecture for intra-domain mobility management.

### A. Mobile IP

Mobile IP (MIP) [2] is the standard solution for supporting IP mobility. Mobile IP allows an MN to roam across different subnets and change its point of attachment, while maintaining the ability to be addressed by its permanently assigned home address. On every change in subnet, an MN obtains a new temporary *care-of address* (CoA), which is topologically consistent with its current point of attachment. This new CoA may be provided either by Foreign Agents (FA) present in every subnet, or via independent configuration protocols, such as DHCP. The MN then informs a Home Agent (HA), located on its home subnet, of this CoA; the HA is then responsible for re-directing packets addressed to the MN by intercepting them and tunneling them to this CoA. Such packet re-direction gives rise to triangular routing, since packets from the CN travel first to the HA before being re-directed to the CoA.

Triangular routing leads to high overhead in networks, especially when the amount of re-directed traffic is high[2]. Such a condition is true either when the number of MNs is large (true for future cellular networks) or the packet traffic rate is high (true for data transfers). Accordingly, a route optimized version [24] of Mobile IP, MIP-RO, has been proposed. In this version, the HA transmits binding updates, on behalf of the MN, to the CNs indicating the MN's current CoA. The CN can then tunnel all packets directly to the MN's CoA, avoiding the need for intermediate re-direction at the HA. This process, however, requires the HA to have pre-established security associations with the relevant CNs, since all such updates must be authenticated to guard against incorrect re-direction.

A slightly modified form of packet re-routing has been specified in the IPv6 version [12] of Mobile IP, MIPv6. In this pro-

---

[2] Experiments and estimates [25] indicate that eliminating triangular routing might lead to about $50\%$ savings in bandwidth consumption.

posal, the binding update is transmitted directly by the MN to the current CNs, whose addresses are stored in a list at the MN. The CN can then send the packets directly to the MN (using a routing header instead of encapsulation), thereby eliminating triangular routing. While the direct transmission of binding updates reduces the latency of the update process, MIPv6 has other drawbacks associated with the size of the header and the need for an MN to generate separate binding updates for each CN.

### *B. Session Initiation Protocol (SIP)*

In contrast to MIP's network-layer mobility management solution, SIP [13] is an application layer protocol for creating, modifying and terminating multimedia sessions with one or more participants. It has gained wide acceptance as the means to set up and maintain sessions for multimedia and telephony services over IP networks. SIP is essentially a peer to peer protocol between user agents (UAs); however, specialized servers called SIP proxies are used to significantly extend the functionality and scalability of the signaling architecture. SIP provides a user-level approach to describe the media capabilities (such as codec type and rates) of sessions, to dynamically change the media capabilities during a session and to re-direct specific application-layer sessions to other terminals/devices. SIP is a very flexible control mechanism, especially for streaming applications, which typically use UDP as the underlying transport protocol.

From a mobility management perspective, the basic SIP specifications provide *user mobility* by allowing an UA to dynamically alter the association between a mobile user and his/her current IP address. A SIP user can use the $SIP-REGISTER$ method to inform a server or other UAs of his/her current IP address; this mechanism provides for inter-session mobility since a SIP CN UA must first query the appropriate SIP servers to resolve the SIP User ID to a currently valid address. Since the user ID is host-independent, a mobile user can be dynamically associated with any mobile node. To allow a user, using a specific MN, to change subnets *during* a session, [3] defines a technique by which the MN uses the SIP $Re-INVITE$ method to inform the CN UA of the new address. The UA at the CN is then responsible for migrating an ongoing multimedia session to the new IP address. The signaling flow is very similar to Mobile IPv6 (with SIP Re-INVITEs replacing MIPv6 Binding Updates), except that the mobility is managed at the application layer.

While proposals for SIP based mobility provide an interesting mobility management alternative to MIP, the flat nature of the management hierarchy poses problems in both cases. Clearly, the signaling load is essentially unchanged since the MN must generate global SIP-REGISTER and Re-INVITEs, or MIP binding updates, for every change in subnet. Moreover, if the CN is located far away, the update latency associated with a SIP Re-INVITE or a MIP binding update can be high. Finally, SIP or MIP provides no paging solution–an MN must obtain configuration information and generate global bindings for every change in subnet.

### *C. Hierarchical Mobility Protocols*

The problems of high update latency and large global signaling load are common to any non-hierarchical mobility management solution. A variety of IP-based hierarchical mobility management approaches have been defined. Hierarchical mobility management is, of course, not a new concept: the current cellular system uses a two-level hierarchy, with the MSC managing mobility within the currently visited domain and the HLR provided a centralized point for resolving the domain currently serving the mobile phone. Packet-based hierarchical solutions for IP networks, however, provide several additional challenges and design alternatives.

Hierarchical IP-based management techniques essentially group IP subnets into *mobility domains* and minimize the scope (and hence latency) of most updates by localizing them to within the domain. Mobility and appropriate packet re-direction within the domain is managed *locally*, by one or more nodes within the domain. Nodes outside the domain are aware only of the current domain where the MN is located. The precise point of attachment within the domain is known only by nodes within the domain. Two alternative approaches can be used to manage intra-domain mobility:

- *Route Modification Approach:* While modifying routing tables on a per-host basis is clearly impractical on a global scale, it may not be infeasible in a local domain. The routing-based solutions essentially ensure that the MN retains a single CoA throughout its sojourn inside a domain; explicit host-based routes are used to route packets to the MN's precise point of attachment.
- *Multi-Address Approach:* In the multi-addressing approach, an MN is associated with multiple CoAs; each CoA simply resolves the MN's location at the lower level in the hierarchy. At each level of the hierarchy, the packets are re-directed to the next CoA.

#### C.1 CellularIP, HAWAII

Cellular IP (CIP) [14] and HAWAII [15] are two route-modification approaches for network layer intra-domain mobility management. Although both approaches currently use Mobile IP as the global (inter-domain) mobility management protocol, they can be extended for use with alternative global protocols. Both CIP and HAWAII implicitly assume a tree-like intra-domain topology[3], with the root of the tree defining the ingress point into the domain. CIP associates an MN with a single CoA, belonging to the ingress node, which is called the Gateway (GW), and uses the MN's permanent home address as the unique identifier inside the domain. Packets tunneled to the CoA are decapsulated at the GW and then forwarded by host-based routing tables using the permanent home address in the inner destination header. In contrast, HAWAII uses the co-located mode of Mobile IP and assigns a unique CoA to each MN; a CN can then directly transmit a packet to this CoA. (Currently, however, HAWAII is specified only for use with MIP, in which case global tunneling is unavoidable). The root of the tree intercepts the packet and then forwards it (based on this unique CoA) via host-based routing tables.

While host-based route modification schemes are plausible for intra-domain management, host-based routes do have cer-

---

[3] CIP and HAWAII can be defined for non-tree topologies only via additional network management techniques, which essentially define a logical tree (per MN or per groups of MNs) over the physical topology.

tain drawbacks (see [16] for a detailed discussion). In particular, host-based routes may lead to high route re-establishment overhead in case of node failures at higher levels in the domain hierarchy. In fact, analysis shows that while the number of routing updates required due to a potential node failure can be low on average, it can be very high in the worst-case. Both HAWAII and CIP, however, have a key advantage: *they do not require additional tunneling inside the domain.* This may be a significant advantage for applications with smaller packets, where encapsulation leads to a sharp increase in the transport overhead.

### C.2 Hierarchical CoA Techniques

Several proposals [17], [18], [19] use two CoAs to define a two-layer hierarchy. We shall shortly discuss our mechanism, called the Intra-Domain Mobility Management Protocol (IDMP) in further detail. The Mobile IP Regional Tunnel Management mechanism (MIP-RR) [17] provides a hierarchical extension to MIP, with a Gateway Foreign Agent (GFA) providing a care-of address that remains valid throughout a mobility domain. [19], on the other hand, defines an approach, specifically tuned for IPv6, that uses intermediate agents to provide an MN a stable, global CoA. However, these protocols are specifically designed as extensions to MIP and do not currently support mechanisms such as fast handoff and paging. We next present our DMA architecture, which is a two-level hierarchical architecture that is independent of MIP.

Table II provides a summary of the configuration, registration and global binding protocols associated with each of the proposed solutions discussed here.

| **Proposal** | **Configuration** | **Registration** | **Global Binding** |
|---|---|---|---|
| MIPv4 | MIP (FA)<br>DHCP | MIP+ MIP Registration<br>AAA | MIP BU |
| MIPv6 | DHCP<br>Auto-config | MIP+<br>AAA | MN-CN BU |
| SIP | DHCP<br>DRCP | AAA<br>SIP Server | SIP REGISTER,<br>SIP Re-INVITE |
| CIP | MIP (FA) | MIP+AAA | MIP |
| HAWAII | DHCP | MIP+AAA | MIP |
| DMA | DRCP<br>IDMP | IDMP+<br>AAA | MIP<br>SIP |

Table II: Alternative Proposals and Mobility Functions

## *D. DMA Architecture*

The DMA architecture is a two-level, hierarchical mobility management architecture and was introduced in an initial form in [16]. Mobility within the domain is managed by using the *Intra-Domain Mobility Management Protocol* (IDMP) [18]. The architecture specifies a new node, called the Mobility Agent (MA), that resides at a higher layer in the network hierarchy (than individual subnets) and that provides an MN with a stable point of attachment. Specialized agents, called Subnet Agents (SA) are present at each subnet to provide support for functions such as paging and fast handoffs. Each MN essentially uses two separate CoAs for managing intra-domain and inter-domain mobility:

- *Local Care-of Address (LCoA)*: This is similar to MIP's CoA as it identifies the MN's present subnet of attachment. This address however has only local (domain-wide) scope; on every change in subnet, the MN informs its MA of the new LCoA.
- *Global care-of address (GCoA)*: This address resolves the location of the MN only up to the granularity of the domain and remains unchanged as long as the MN stays in the current domain. Packets routed to the GCoA are intercepted by the MA and then tunneled to the LCoA.

This architecture has two key distinguishing features:

- The intra-domain mobility protocol is distinct from the global binding protocol. Accordingly, the approach allows individual MNs, to use MIP, SIP or any other mechanism to inform remote nodes (HA, CN, SIP server etc.) of changes in the GCoA. Also, intra-domain authentication and security is completely distinct from global authentication.
- The architecture assumes the presence of multiple MAs in a domain and uses load-balancing algorithms to distribute the mobility load across MAs. More importantly, load-balancing is combined with dynamic resource provisioning and MA-SA signaling to provide MNs with QoS assurances without the need for QoS re-negotiation on every change in subnet.

Figure 1 depicts the functional layout of IDMP, and shows how the packet is forwarded to the MN via two successive CoAs. The DMA architecture essentially uses the Differentiated Services (diffserv) [26] framework for assuring QoS within the domain. The architecture requires the MAs to interact with a Mobility Server, which implements the load balancing algorithm, and a centralized Bandwidth Broker (BB), which is responsible for dynamically reserving resources for the various traffic classes. Details on the integrated QoS support in the DMA architecture are available in [20].

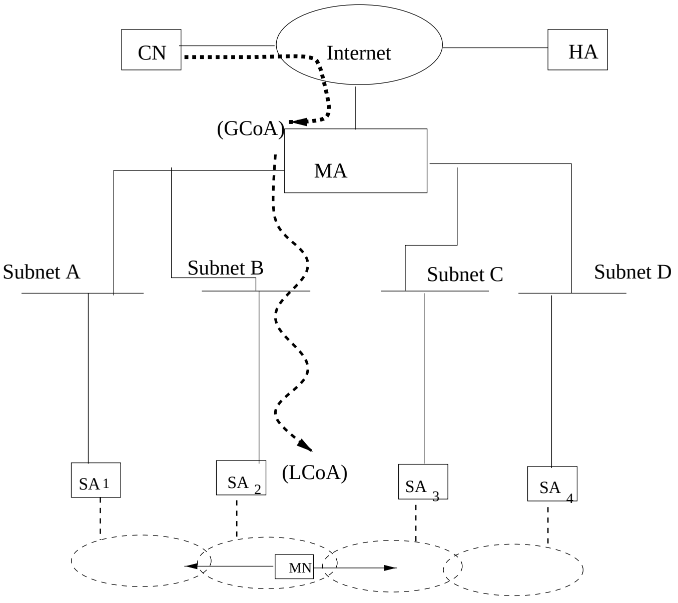

Figure 1: IDMP Logical Elements & Architecture

### D.1 IDMP Fast Handoffs and Paging

Both fast handoff and paging support in the DMA architecture use some form of multicasting and are logically represented in Figure 2. In basic IDMP, the latency of the handoff process (and hence the duration of service interruption) equals the time needed for the MN to inform the MA of its new LCoA. For nodes requiring faster handoff in IDMP, the MN transmits a $MovementImminent$ message to the MA, whenever it senses (via layer-2 triggers) the possibility of movement. The MA then proactively multicasts inbound packets, for a limited duration, to the SAs that are neighbors of the MN's current point of attachment, where such packets are temporarily buffered. If the MN subsequently registers with one of these neighboring SAs, the new SA will forward such cached packets immediately after the subnet-level registration is complete, thereby eliminating the delay and losses associated with the transmission of an intra-domain location update . Figure 2 shows the use of pro-active multicasting (solid lines) to support fast handoff for neighboring SAs, $SA_1$ and $SA_3$. Since IDMP's paging process is very similar to the fast handoff mechanism, we refer the reader to [18] for complete details. In the paging mode, an $Idle$ MN does not perform any location update or registration as long as it stays within a Paging Area (PA), comprising multiple subnets. On receipt of an incoming packet for an $idle$ MN, the MA buffers it and multicasts a $PageSolicitation$ to the MN's current PA, requesting the MN to re-register at the MA with a new and currently valid LCoA. Figure 2 shows the transmission of a $PageSolicitation$ (dashed lines) to the MN's current PA, namely $PA_2$, comprising subnets B, C and D.

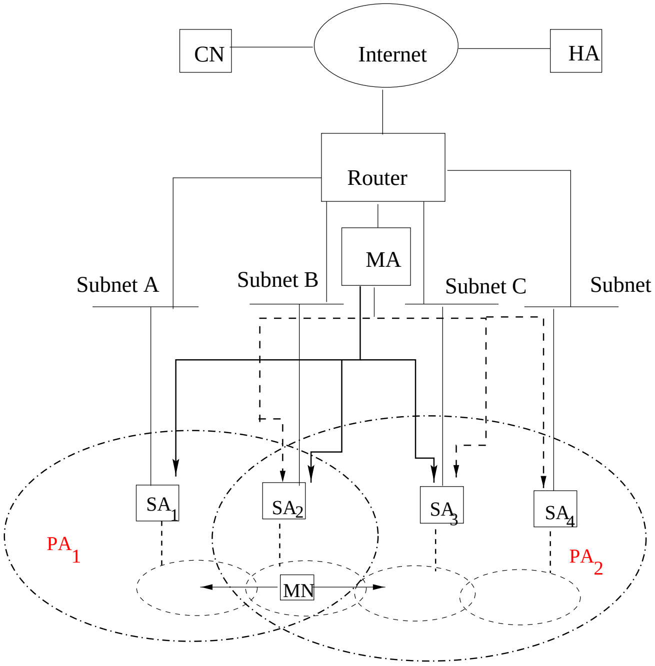

Figure 2: IDMP Fast Handoff/ Paging

The two-layer multi-CoA approach does not suffer from the scalability concerns associated with the route-modification approach. However, the multi-CoA approaches do require the intermediate agent to handle all traffic destined to an entire group of MNs (e.g., all MNs managed by a single MA). Moreover, the multi-CoA approach requires this agent to decapsulate and re-encapsulate all incoming packets; this not only results in higher transport overhead, but could also lead to processing bottlenecks at the MA. Although DMA's use of multiple MAs alleviates the potential for bottleneck, such traffic concentration may turn out to be a scalability issue. In general, both the route modification and multi-CoA approaches are legitimate solutions for managing intra-domain mobility. There is very little experimental or analytical evidence on the relative performance of these two approaches. We believe that detailed simulation studies of the two approaches are needed to determine their comparative performance under realistic topologies and workloads.

## V. APPLICATION-DEPENDENT CHOICE OF MOBILITY MANAGEMENT SCHEMES

In sections III and IV, we have analyzed the mobility related requirements of each of the three applications, and the performance characteristics of the typical IP mobility solutions. We now study the appropriateness of these mobility management strategies on an application-specific basis. After studying each application independently, we shall see how universal mobility support can be achieved by defining a two-level mobility management hierarchy, with a common protocol and architecture for managing both intra-domain QoS and intra-domain mobility, and multiple binding protocols for optimal management of inter-domain mobility.

### A. *Mobility Management for VoIP*

VoIP applications require a user to be always locatable by a CN; accordingly, a global binding protocol must be used to make the MN's current CoA available to the individual CNs or network servers.

A flat Mobile IP architecture is not suitable as a mobility management solution for several reasons. Firstly, MIP leads to significant latency in the handoff process, since the update must reach the HA before packets are correctly tunneled to the new CoA. Moreover, the basic scheme leads to too much overhead, especially in the MN-MN communication scenario, where packets suffer from quadrilateral routing (triangular routing in both directions). The need for global updates (over possibly multiple hops) not only increases the latency of communication, but, in IP networks characterized by possible packet losses, also significantly increases the mean time before a binding update is reliably received. To understand the drawback of such multi-hop binding updates, consider an arrangement whereby the CN, HA and MN are three vertices of an equilateral triangle, with each side corresponding to $S$ separate hops. Let each hop have a probability $p$ of packet loss and result in a delay of $d$ msecs. The probability of the transmission of a successful binding update over any side of the triangle is

$$P_{succ} = (1-p)^S. \tag{1}$$

Since each update (the first one or subsequent retransmissions) is successful with probability $P_s$, the probability that the update takes exactly $K$ transmissions is $P_S^K$:

$$P_S^K = (1-P_s)^{K-1}P_s. \tag{2}$$

An update traversing $S$ hops incurs a delay of $S*d$ msecs. Assuming that each retransmission is generated at an interval of $S*d$ msecs, we see that if $X$ denotes the random variable indicating the time till a successful transmission, the cumulative distribution $F_S(kd)$ is given by:

$$F_s(kd) = Prob\{X \leq k * S * d\} = \sum_{i=1}^{k} P_s^k. \tag{3}$$

For the MIP-RO case, however, a successful transmission requires two independently successful transmissions, MN to HA and HA to CN. Thus, the probability of a successful transmission in exactly $K$ transmissions is

$$P_{S,MIP-RO}^{K} = \sum_{i=1}^{K-1} P_S^i * P_S^{K-i}, \tag{4}$$

whence we can derive the cumulative distribution $F_s^{MIP-RO}(.)$.

Figure 3 shows the distribution of $F_s(.)$ for a hypothetical operating condition, where $d$, the per-hop delay is 10 msecs, $S = 5$ hops and $p$ is either $0.01$ or $0.05$. The figure shows that the probability of relatively large mobility-related transients is not insignificant, especially in networks with reasonably high packet loss probabilities.

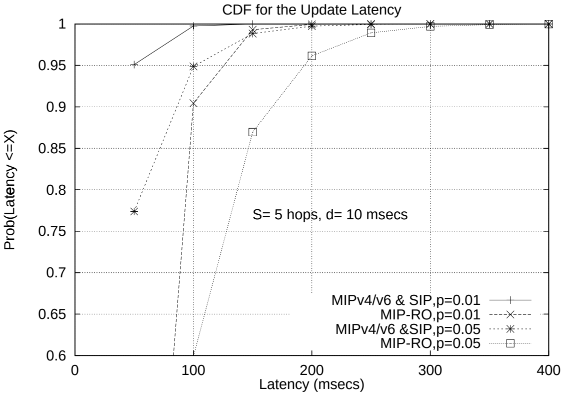

Figure 3: Update Latency Distribution for MIP/SIP Schemes

MIP is also unsuitable for VoIP due to the high packet overhead. VoIP packets are typically small; the 20 byte encapsulation header in MIPv4 or the 16 byte routing header in MIPv6 can significantly increase the payload overhead[4]. Moreover, MIP provides no support for paging, a feature commonly available in all voice-based cellular networks.

SIP can however, serve not just as a control protocol for VoIP applications, but also as a useful part of the overall mobility management solution. Since VoIP calls have an initial explicit call establishment phase, there is no need to maintain a permanent home IP address. Since the VoIP packets are transmitted over UDP, the SIP UA at the CN can simply transmit a packet directly to the CoA indicated in the SIP Re-INVITE message. The direct transmission of a binding update from the MN to CN provides for the elimination of triangular routing without incurring the latency of MIP-RO. Finally, by avoiding encapsulation, SIP allows an efficient mobility solution where the IP header is kept to a minimum. While SIP-based mobility management has several advantages, it still suffers from the drawbacks of a flat management architecture presented earlier (in fact, the distribution of the update latency for SIP is identical to that for MIPv6). Also, SIP does not provide any paging support.

### A.1 Proposed Solution

To remove these drawbacks, we propose a *combined two-level mobility management hierarchy, with the IDMP-based DMA architecture used to manage local mobility and SIP used as the global binding protocol*. Under this solution, the MN generates a SIP Re-INVITE (or a SIP REGISTER), with the GCoA as the advertised address, only when it changes domains. The solution proposes to use IDMP's Globally Co-located (GC) mode, where each MN obtains a unique GCoA. *Accordingly, the CN does not need to tunnel packets to the GCoA and can forward them with any encapsulation over the global Internet.* Since the bulk of the location updates are intra-domain and restricted to the MA, the handoff latency and global signaling load is significantly reduced. Moreover, VoIP nodes can activate the IDMP fast handoff procedure to further reduce the interruption associated with an intra-session handoff. The duration of service interruption with IDMP fast handoff is O(10-20)msecs and essentially includes the latency associated with performing a new subnet configuration (new LCoA) at the new subnet. The per-user buffering employed in IDMP's fast handoff process practically eliminates the loss of in-flight packets. Typical VoIP applications use a playback buffer that allows for variability in the delay of individual packets. Experiments indicate that audio quality suffers no perceptible loss as long as packets are not lost but simply delayed. Accordingly, using IDMP serves to significantly improve the IP-based mobility management of VoIP traffic.

IDMP provides voice applications with another significant benefit: paging. Since cellular phone users may roam over multiple subnets while in an idle state, IDMP's intra-domain paging mechanism allows an MN to save significant signaling power. Moreover, IDMP's paging operation is ideally suited to IP telephony applications where the initial packets from the CN are call-establishment (control) messages. By buffering such packets, the MA prevents packet losses until the MN responds to the paging request. IDMP's paging delay, which includes the time for propagation of the PageSolicitation broadcast, the subsequent subnet configuration and finally the intra-domain location update, can be bounded by O(100)msec in practical systems. This delay is well within acceptable thresholds for the call establishment latency. IDMP's dynamic QoS provisioning architecture is also suited to re-establishing the QoS profile for an MN without the need for explicit signaling at every change in subnet.

The IDMP-based solution has one drawback– the need for tunneling VoIP packets to the LCoA inside the domain. When the LCoA is co-located, this tunneling leads to a reasonably large percentage loss of bandwidth even over the final air-interface. Of course, intra-domain routing based solutions, such as Cellular IP or HAWAII, do not have this drawback since no additional intra-domain tunneling is required. However, this tunneling is the current price to pay for a scalable and robust intra-domain management protocol. Such tunneling can be eliminated if the MA is permitted to simply replace the GCoA in

---

[4] As an example, consider a G.711 VoIP packet, with a payload of 80 bytes (20 msec packetization delay). The normal IPv4 packet has a header (UDP+RTP+IP) of 40 bytes; IP-in-IP encapsulation adds a further 20 bytes. The payload efficiency for the encapsulated packet is thus $80/140 \sim$ 57%.

the packet header with the LCoA; however, this may cause end-to-end security and authentication mechanisms to fail. Also, proposals under investigation, which minimize the header overhead via robust header compression techniques, may alleviate this drawback of intra-domain tunneling.

### A.2 Signaling Flow for IDMP+SIP Solution

Figure 4 shows the message flow (both IDMP and SIP) for this hybrid mobility management approach, when an MN first moves into the domain. The signaling also includes the messaging exchanged between IDMP nodes (including the Mobility Server (MS)) and the QoS provisioning elements (such as the Bandwidth Broker (BB)) to dynamically provision resources for the VoIP node. To authenticate the MN, the MA may also need to interact with AAA servers; this signaling is however not shown in the figure.

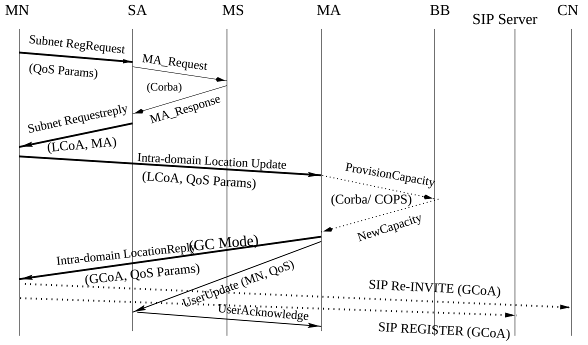

Figure 4: Signaling Flow for VoIP Mobility

For nodes supporting primarily VoIP traffic, we have shown how a *combination of application layer and network layer mobility support* provides the most scalable solution. Mobility within a domain is transparent to the application layer (the GCoA stays constant) and is managed through a secondary address (LCoA) at the network layer. The network layer is also used to support features such as seamless handoff and paging. SIP is used as the application layer protocol to provide a global mechanism for both terminal and user mobility. The use of SIP allows the applications to control the user response to mobility across domains.

### *B. Mobility Management for Web-browsing*

Universal Web-based access to the Internet is likely to be a hallmark of next-generation networks. Roadwarriors will access the network from a variety of places, such as airports, hotels, shopping malls and sports complexes. In this model of network access, a user simply roams to different access networks and connects to the network via his/her own node or a temporarily borrowed node. For example, a user may rent a device in an airport and access the network via either the cellular or a wireless LAN infrastructure. In such scenarios, each LAN provider may be a different operator; in the absence of a fixed association between an MN and a user, each operator may independently need to authenticate and authorize a user. Such users may or may not exhibit mobility across subnets during one session; while it is possible that a user walking in an airport may change the subnet of attachment, it may safely be assumed that the user will never change domains during such a movement. The exact nature of intra-session mobility in this case gives rise to two different solutions for mobile Web access.

### B.1 Proposed Solution for Intra-Subnet Mobility

We first consider the case where the user does not change subnets during a single active session. This may occur, for example, when a single subnet spans an entire airport terminal; intra-session mobility in this case is confined to possible changes in the link-layer point of attachment and is handled by appropriate layer-2 micro-mobility techniques. In such a scenario, node configuration and user registration are the two important IP mobility management functions. User registration is critically important, since a service provider charging for such connectivity would definitely need to verify the user's identity and his/her service requirements in a device-independent fashion. Since Web-browsing applications do not require continuous locatability, a binding protocol is unnecessary. In fact, even dynamic packet rerouting is not required, as the IP address of the MN does not change for the entire duration of a single session.

To provide configuration information, such as a valid IP address, to an MN for a specific session, we can use configuration protocols such as DHCP[4] or DRCP [6]. Traditional mobility solutions typically integrate the registration functionality with the binding mechanism; for example, Mobile IP (with its newly defined AAA interface) can support node authentication and authorization along with mobility binding. For nodes that do not require a binding protocol, such integrated functionality imposes extra overhead in the overall registration process. Simpler and specialized registration protocols may be used to allow such Web-browsing users, who do not have MIP client on their devices, to access the network. Several approaches and protocols, such as IEEE 802.1X [21], [22] and BURP [23], are currently being researched. While BURP offers a uniform and distinct registration protocol at the application layer, IEEE 802.1X offers port-based access based on the IEEE802.11 layer-2 technology. However, both the protocols need to be further investigated to determine their applicability and interworking capability with existing AAA protocols such as RADIUS and DIAMETER.

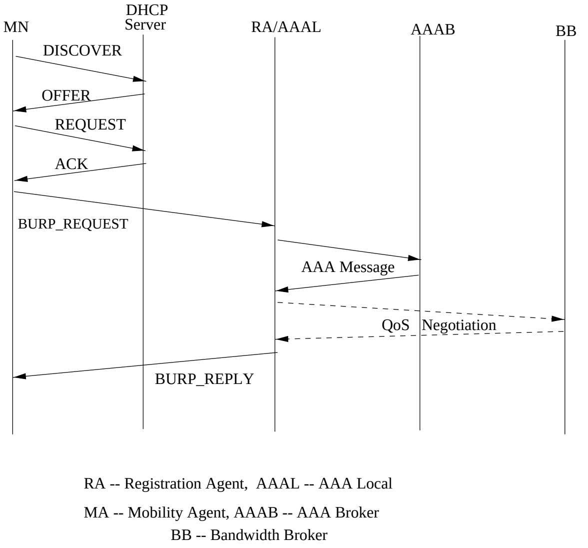

Figure 5: Signaling Flow for Mobile Web Access Using BURP

Figure 5 shows the signaling flow for mobile Web access using BURP. DHCP is used as the configuration protocol and provides the MN with a valid IP address (CoA) as the mobile user attaches his/her device to the network. A single CoA (a flat addressing architecture) is adequate in this case, since the MN does not need to issue any global or local binding updates. As part of the initial configuration parameters, DHCP also provides the BURP client (running on the MN) with the address of the BURP Registration Agent. The BURP client then sends a registration request (BURP_REQUEST) to the RA which in turn replies back (BURP_REPLY) to the client after proper authentication. During this process, RA will first contact the local AAA (AAAL) server running DIAMETER or RADIUS, and then a broker or home AAA server as necessary. BURP registration agent and local AAA are co-located in our model. Finally, dashed lines in the figure indicate that BURP RA may interact with QoS provisioning elements (such as Bandwidth Broker) to satisfy the user-specific QoS requirements.

### B.2 Proposed Solution for Inter-subnet Mobility

We now consider the case where an MN roams across multiple subnets during a single session. Such a case might occur, for example, when an airport terminal is partitioned into several distinct IP subnets; a user may change subnets even while roaming inside the airport terminal. To ensure transparent packet redirection for an ongoing session during such a subnet change, an intra-domain mobility solution, such as the DMA hierarchical mobility architecture, is sufficient. Since an MN is the sole initiator of a session, universal locatability of the user is not necessary; there is consequently no need for a global binding protocol. The independence of DMA from a specific global binding protocol is a key advantage in this situation. To retrieve data from the global Internet, the MN uses its GCoA as the source address. Accordingly, as long as the MN stays within the domain, the correspondent node (Web server) is unaware of the node mobility and simply transmits all packets (without encapsulation) to the GCoA. Such packets are intercepted by the MA and then forwarded via encapsulation to the MN. Encapsulation is necessary to preserve end-to-end security and authentication, both of which are expected to be standard components of any secure transactional model. While the DMA approach does incur the encapsulation overhead inside the mobility domain, this overhead (20 bytes for the IPv4 case) is not very significant for such data applications, which typically have large packet sizes.[5] While IDMP provides support for fast handoff, Web browsers are unlikely to request such support, since the application is not sensitive to variations in the delay of individual packets. The MN can also optionally use IDMP's paging support. However, the service model assumes that the MN is the sole initiator of all traffic sessions. Accordingly, paging support is not essential, since the MN does not need to be always locatable. An idle MN can simply discontinue the use of IDMP and re-initiate a new intra-domain registration as a new user ”logs on”.

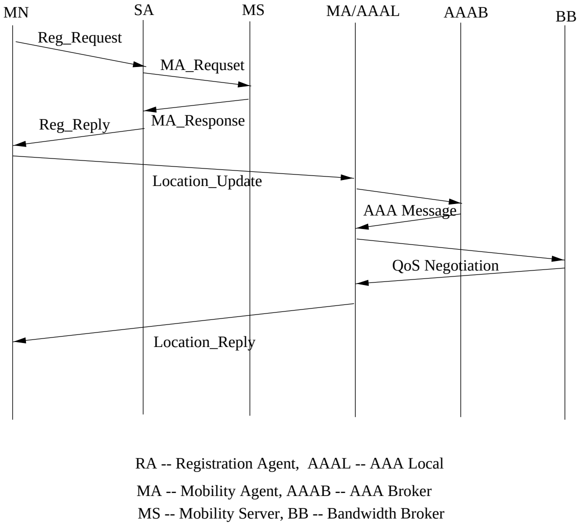

Figure 6: Signaling Flow for Mobile Web Access using IDMP

Figure 6 shows the typical signaling associated with the mobile Web application in this case. The figure assumes that the MN uses IDMP not only to obtain the configuration parameters, but also to specify the QoS requirements and register with the network. In such a model, we assume that the MA will interact with standard AAA protocols (e.g., RADIUS, DIAMETER) for authentication and with the BB for QoS provisioning. In contrast to figure 4, a global binding update message is absent, since all mobility during a single Web session is confined to the single domain. If the MN indeed changes domains during a session, the MN will obtain a new GCoA and the user will have to re-establish another session.

### C. *Mobility Management for Mobile Servers and Bulk Data Transfer*

In our mobile server scenario, we need the MN (server) to be always locatable via a permanent (home) address. Accordingly, a global binding protocol is necessary to store the MN's current CoA in centralized servers. Since the CNs are not assumed to perform an explicit query (for functional compatibility with current Internet hosts), the mobility management scheme also requires *transparent re-direction*: changes in the CoA of the MN must be invisible to the CN. Reliable bulk transfer applications use TCP as the transport mechanism. To provide seamless connectivity for ongoing TCP sessions, a network-layer mobility solution must be used that makes node mobility transparent to the upper layers.

MIP is thus a plausible mobility management approach for such bulk transfer applications. In fact, MIP was initially designed to provide ubiquitous and seamless network connectivity precisely for such TCP-based, non-real time applications. However, MIP was also designed for a predominantly static network architecture, where only a small fraction of the total hosts exhibited mobility. Node mobility will, however, be a fundamental feature of next-generation cellular networks. Given the potential for billions of mobile devices, MIP's flat architecture can lead to a significantly high global signaling load. Moreover, we have also seen that the need to transmit binding updates globally (potentially over a large number of hops) can lead to a significant transient period, where the CN loses connectivity to the MN.

[5] For most Web-based *pull* applications, it is the MN which generates small sized packets (TCP Acks). Such outbound packets do not need to be encapsulated as they travel directly to the (stationary) server.

### C.1 Proposed Solution

A two-level mobility management scheme, *which combines IDMP with MIP*, solves almost all the shortcomings associated with the base MIP solution. We have presented such a solution as part of the TeleMIP mobility management architecture [16], whereby the MN sends a global MIP registration packet only when it changes domains and obtains a new GCoA. This registration message specifies the GCoA as the MIP care-of address; packets are accordingly tunneled to the MN's designated MA by the HA. Since the HA will always tunnel packets to the GCoA, the MN uses IDMP's Mobility Agent (MA) mode for global addressing, whereby multiple MNs share the same GCoA. The DMA architecture is used for intra-domain mobility management, with the MA decapsulating inbound packets and then forwarding them (via re-encapsulation) to the MN's current LCoA. By reducing the frequency of multi-hop global binding updates, this hierarchical solution not only reduces the global signaling load but also significantly decreases the loss probability of an individual binding update packet. The DMA architecture also allows an MN to utilize IDMP's paging mechanism and conserve power in an idle state.

Our mobility solution to the mobile server application scenario uses MIPv4. The alternative MIP mechanisms, MIP-RO and MIPv6 could also be used. Both these mechanisms remove the overhead of triangular routing in the global Internet, since the CN now sends packets directly to the MN's CoA (IDMP's GCoA). In the mobile server scenario, triangular routing may not be a significant drawback, at least from a bandwidth overhead standpoint. The bulk of the data flow is principally from the MN towards the CN; such packets can indeed be transmitted directly to the CN. Inbound traffic (towards the MN) is typically small and consists principally of acknowledgement packets. Moreover, both MIP-RO and MIPv6 suffer from certain drawbacks associated with this specific application scenario. MIP-RO requires an upgraded CN that is aware of the server's mobility; the CN must not only accept binding updates and then tunnel packets for the MN to its currently registered CoA. If the CN indeed possesses this capability, then MIP-RO can certainly be used to provide a more direct routing of packets to the MN. A solution combining DMA with MIPv6, on the other hand, requires the MN to send individual binding updates to each of the currently active CNs, whenever it changes domains. Since a server node could potentially have a significant number of active CNs, this can lead to a large signaling load at the MN, especially over the first-hop wireless interface.

### C.2 Signaling flow for the TeleMIP solution

Figure 7 shows the signaling flow when a mobile server node (MN) first moves into an IDMP domain. One of the key differences with MIP-RR is the complete separation of the intra-domain and global update mechanisms. In the MIP-RR mechanism, the Gateway Foreign Agent (GFA) acts as a relay in the MIP registration process and is, hence, required to understand the MIP registration semantics. In our approach, the global MIP registration message is generated directly by the MN- the IDMP agents, such as the MA and the SA, are unaware of this registration mechanism. As stated earlier, such a separation makes IDMP functionality completely independent of the alternative global binding solutions and allows a common intra-domain management infrastructure to support multiple global binding solutions, each of which may be appropriate for a specific application scenario.

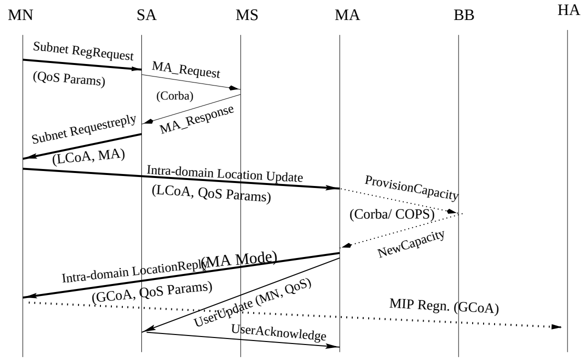

Figure 7: Signaling Flow for Mobile Servers

## VI. CONCLUSION

Next-generation IP-based cellular networks are expected to provide an integrated access and mobility management infrastructure for a variety of applications, with significant differences in their traffic profiles and service requirements. In this paper, we have investigated the merits and demerits of various IP-based mobility management schemes for three such application scenarios: VoIP, mobile Web access and mobile server-based data transfer.

The three applications differ in their need for various mobility-related features, such as continuous reachability, mobility transparency, fast handoffs and paging. We analyzed various IP mobility management proposals, such as MIP, SIP, CIP and HAWAII, and demonstrated why a hierarchical management architecture is important, especially as the number of mobile nodes increases. We also discussed the two alternative approaches to intra-domain mobility management. While the route-modification approaches (HAWAII, CIP) do not require any additional tunneling inside the domain, the multi-CoA approaches (MIP-RR, DMA, HMIP) do not need to maintain host-specific routes. As part of a hierarchical mobility solution, we described our Dynamic Mobility Agent (DMA) architecture, which allocates specialized agents called Mobility Agents dynamically to MNs and which uses load balancing algorithms and dynamic resource provisioning techniques to define an integrated QoS architecture.

Differences in the mobility requirements of each of the three applications considered imply that no single mobility solution performs best in all three cases. For the VoIP scenario, we see how combining the DMA intra-domain mobility architecture with SIP-based global mobility management provides a solution that minimizes the global signaling load, prevents unnecessary transport overhead and provides flexible support for features such as fast handoff and paging. The mobile Web access scenario, on the other hand, is a pull-based application that does not need the MN to be continuously locatable and accordingly does not require a global binding protocol. Accordingly, we see how support for configuration and registration is adequate for a wide variety of access scenarios. For scenarios where node movement during a single session is restricted to a single domain, the DMA approach can be used to manage intra-domain mobility in a manner transparent to the global Internet. Finally, we considered the case of mobile servers, where CNs retrieve data from an MN. In this case, we see that a MIP-based network layer mobility solution provides significant benefits by making node mobility transparent to the upper layers. As in the other application scenarios, a hierarchical mechanism using DMA as the intra-domain mobility solution provides a more scalable and robust solution.

While our analysis provides guidelines for a preferred mobility management approach, several questions remain unanswered. Perhaps most importantly, there has been no serious analysis of the relative merits of the route-modification and multi-CoA alternatives for intra-domain mobility management, especially in terms of their signaling overhead and scalability. It may well turn out that the route-modification approach is more appropriate for smaller and medium-sized domains. As we have pointed out in the introduction, all the management schemes discussed here essentially resolve the MN's location up to a subnet-level granularity. Depending on the penetration of IP into the cellular infrastructure, additional layer-2 mechanisms may be necessary to manage micro-mobility (across different access points within a subnet). Further research is necessary to decide how to effectively integrate the capabilities of such layer-2 mobility management mechanisms with IP-layer and above mobility solutions.

## REFERENCES

[1] K. S. Hellstern, G. Pollini and D. Goodman, “Network Protocols for the Cellular Packet Switch”, Proceedings of IEEE Vehicular Technology Conference, 1992.

[2] C. Perkins, “IP Mobility Support for IPv4, revised”, draft-ietf-mobileip-rfc2002-bis-02.txt, IETF, July 2000.

[3] E. Wedlund and H. Schulzrinne, “Mobility support using SIP,” Proceedings of Second ACM International Workshop on Wireless Mobile Multimedia, ACM/IEEE, August 1999, pp. 76–82.

[4] R. Droms, “Dynamic Host Configuration Protocol”, RFC 2131, IETF, March 1997.

[5] W. Simpson, “The Point to Point Protocol (PPP)”, Internet STD 51, July 1994.

[6] A McAuley, S. Das, S. Baba and Y. Shobatake, “Dynamic Registration and Configuration Protocol”, draft-itsumo-drcp-00.txt, IETF, July 2000, Work in Progress.

[7] C. Rigney, S. Willens, A. Rubens and W. Simpson, “Remote Authentication Dial In User Service (RADIUS)”, RFC 2865, IETF, June 2000.

[8] P. Calhoun, et.al, “DIAMETER Base Protocol”, draft-calhoun-diameter-17.txt, September 2000, Work in Progress.

[9] AT&T, “AT&T sets the standard for network reliability”, http://www.att.com/network/standrd.html, October 2000.

[10] Hammer Technologies, “Quality of Service Testing in the VoIP Environment: A Primer,” White Paper, http://www.hammer.com/ p_paper.htm, June 1998.

[11] J Fielding, J Gettys et al, Hypertext Transfer Protocol- HTTP/1.1, RFC 2616, IETF, June 1999.

[12] D. Johnson and C. Perkins, “Mobility support in IPv6”, draft-ietf-mobileip-ipv6-12.txt, IETF, April 2000, Work in Progress.

[13] E. Schooler, M. Handley, H. Schulzrinne and J. Rosenberg, “SIP: Session initiation protocol,” RFC 2543, IETF, March 1999.

[14] A. Campbell, J. Gomez, C-Y. Wan, S. Kim, Z. Turanyi and A. Valko, “Cellular IP,” draft-ietf-mobileip-cellularip-00.txt, IETF, January 2000, Work in Progress.

[15] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan and L. Salgarelli, “IP micro-mobility support using HAWAII,” draft-ietf-mobileip-hawaii-01.txt, July 2000, Work in Progress.

[16] 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, pp. 50-58.

[17] E. Gustafsson, A Jonsson and C. Perkins, “Mobile IP Regional Tunnel Management”, draft-ietf-mobileip-reg-tunnel-03.txt, July 2000, Work in Progress.

[18] 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, IETF, July 2000, Work in Progress.

[19] H. Soliman, C. Castelluccia, K. El Malki and L. Bellier, “Hierarchical MIPv6 Mobility Management”, draft-soliman-mobileip-hmipv6-01.txt, IETF, September 2000, Work in Progress.

[20] A. Misra, S. Das, A. McAuley, A. Dutta and S. K. Das, “Integrating QoS Support in TeleMIP's Mobility Architecture”, to appear in Proceedings of IEEE International Conference on Personal Wireless Communications, December 2000.

[21] “IEEE Standards for Local and Metropolitan Area Networks: Port based Network Access Control, IEEE draft 802.1X/D8, October 2000.

[22] P. Congdon et.al., “IEEE 802.1X RADIUS Usage Guidelines”, draft-congdon-radius-802.1x-05.txt, IETF, November 2000.

[23] S. Das, Anthony Mcauley and B. Patil, “Basic User Registration Protocol (BURP)”, Minutes of the AAA-WG, IETF-48, Pittsburgh, July 2000, Work in Progress.

[24] C. Perkins and D. Johnson, “Route Optimization in Mobile IP”, draft-ietf-mobileip-optim-09.txt, IETF, Work in Progress, February 2000, Work in Progress.

[25] R. Jain, T. Raleigh, C. Graff and M.Bereschinsky, “Mobile Internet Access and QoS Guarantees using Mobile IP and RSVP with Location Registers”. Proceedings of IEEE International Conference on Communications (ICC), June 1998.

[26] S. Blake, et al, “An Architecture for Differentiated Services,” RFC 2475, IETF, December 1998.