IP Routing and Mobility

Cristina Hristea and Fouad Tobagi

Stanford University
{christea,tobagi}@Stanford.edu

Abstract. The original design of the Internet and its underlying protocols did not anticipate users to be mobile. With the growing interest in supporting mobile users and mobile computing, a great deal of work is taking place to solve this problem. For a solution to be practical, it has to integrate easily with existing Internet infrastructure and protocols, and offer an adequate migration path toward what might represent the ultimate solution. In that respect, the solution has to be incrementally scalable to handle a large number of mobile users and wide geographical scopes, and well performing so as to support all application requirements including voice and video communications and a wide range of mobility speeds. In this paper, we present a survey of the state-of-the-art and propose a multi-layer architecture for mobility in IP networks. In particular, we propose the use of extended local area networks and protocols for efficient and scalable mobility support in the Internet.

1 Introduction

In the broadest sense, the term mobile networking refers to a system that allows users to maintain network connectivity while moving from one location to another. Mobility is often associated with wireless technologies that require mobile networks to support continuous movement, at high speeds and for long periods of time. Recently, there has been an explosive growth in wireless devices with built in access to the Internet. In the near future, large numbers of mobile users will access the Internet for a variety of high-speed multimedia services. IP packet switching has become the standard towards which many networks are converging, including those in the telecommunication sector, as less efficient, less enabling circuit-switched technologies are abandoned. Although a lot of progress has been made, supporting mobility in IP networks is still a difficult challenge.

1.1 Challenges and Early Solutions

1.1.1 Duality of IP Addresses

The IP addressing scheme was designed and optimized for a stationary environment, which makes mobility difficult. With the introduction of mobile networking, IP addresses have acquired a dual significance. On one hand, they are expected to remain fixed during the course of a connection. An important reason for this is that, while, in principle, higher layers (above IP in the protocol stack) are supposed to be

S. Palazzo (Ed.): IWDC 2001, LNCS 2170, pp. 279–294, 2001.

independent of the IP layer, in practice they make use of the IP address for basic functionality. For example, the transport layer uses this address to establish and maintain connections. If the IP address is changed during the course of a session, the connection is lost and the session is terminated. Therefore, to maintain seamless connectivity during movement, IP addresses need to be kept fixed, transparent to changes in user locations. On the other hand, IP addresses need to change dynamically as users move, since they are used for packet routing and delivery. Routers in the Internet use IP addresses in the destination field of packets to identify the subnet where the user is located, and to obtain the MAC address of the user for final delivery of the packets. Moreover, typically routers in the Internet use address based filtering to discard packets whose source IP address are from outside the subnet. Therefore, the user IP address needs to change as the user changes location in order to conform to addressing at the new location[1].

Notice that mobile networking inside a subnet is not affected by the dual significant of IP addresses. Mobile users can roam inside a subnet without having to update their IP addresses. The reason why this is possible is because LAN switches learn the users location and can route packets to them quickly using this information.

One way to resolve the duality of IP addressing is to change the transport and application layers of the protocol stack in order to handle a dynamic IP address. A mobility solution at the TCP layer is proposed in [3]. Connection migration is performed to maintain connectivity for sessions in-flight at the time of move. For this solution to work, the mobile hosts, and fixed hosts in the Internet wishing to communicate with mobile hosts would need to be upgraded to the new versions of software. While upgrading the mobile hosts may be an easier task, upgrading all the hosts in the Internet is not a possibility. Furthermore, achieving good application performance with dynamic IP addresses remains a significant challenge. Simulation results show that significant disruption is incurred during migration; moreover the solution limits movement to a single end and also may apply to TCP applications only. Another transport layer solution, proposed in [18] suggests TCP be modified to use domain names instead of IP addresses. Again, the main disadvantage of this solution is that it does not integrate easily with the existing Internet, hence it could be prohibitively expensive to deploy.

The alternative solution to the problem of IP address duality is to allow hosts to maintain a fixed IP address as they move across subnets. In turn, this would require that routers propagate host-specific routes in the Internet. However, host-specific routing requires space in the routing tables proportional to the number of hosts, slows down the routing process and consumes potentially excessive bandwidth in the Internet.

1.1.2 Mobile IP

In the 1990's, the IETF designed a solution for mobility known as Mobile IP [4], which overcomes the duality of IP addresses without requiring that routers learn host-

[1] Note that, to resolve this duality, in Ipv6 a host is allowed to use two addresses. One address is used as a permanent identifier while the other address is used for routing purposes. The permanent address is included in the main Ipv6 header, while the routing address is inserted in a special-purpose extension header used for routing.

specific routes. Mobile IP solves the problem by allowing a single computer to hold two addresses simultaneously. The first address is permanent and fixed. It is the address that transport and application protocols use. The second address is temporary – it changes as the computer moves, and is valid only while the computer visits a given location.

A mobile host MH is assigned a permanent home address and a home agent HA in its home subnet. DNS maps the domain name of the host to its home address. When the MH moves to a foreign subnet, it acquires a temporary care-of-address COA from an advertised foreign agent in the subnet, and it registers its new address with the HA. The HA uses gratuitous proxy ARP to capture all IP packets addressed to the MH's permanent address and uses encapsulation to forward them to the mobile's current location[2].

There are two possibilities for the packets going back from the MH to the corresponding host CH. One choice is for the MH to send out un-encapsulated IP packets with the permanent home address of the MH as the source address and the address of the CH as the destination address. However, some routers in the Internet use address based filtering and discard packets from outside the subnet. To avoid this, the MH needs to encapsulate the packet using its COA as the source address and the address of the HA as the destination address. The HA decapsulates the packet and forwards it to the CH.

One thing to notice is that packets delivered via HA typically travel further through the Internet than they would if delivered by the optimal unicast route. Apart from increasing the round-trip delay observed by the communicating parties, this also affects other users by increasing the overall load on the shared resources of the Internet. A proposed mechanism, known as route optimization, attempts to fix this, by using binding updates, containing the current COA of the MH, from the HA to the CH. A CH with enhanced networking software can learn the temporary COA and then perform the encapsulation itself, sending the packet directly to the mobile host. This avoids the overhead of indirect delivery.

1.1.3 Industrial Solutions and Mobile IP

Mobile IP, or some variant thereof, is a popular solution adopted by the majority of industrial products offering IP connectivity to mobile users.

1.1.3.1 Ricochet

The Ricochet system from Metricom [17] implements a solution for IP mobility that is similar to the Mobile IP protocol. However, it is important to point out that Ricochet was designed more than a decade ago hence it predates the Mobile IP protocol. Wireless cells are connected to IP gateways and name servers that provide security, authorization and roaming support to users. At any given point in time, a user has three addresses: one IP address, which is fixed, and two layer-2 addresses: one is fixed and unique to that user, and the other is dynamic and unique to the cell where a user is located at that point in time. When a user first connects to the network, its request is validated by the local gateway and name server. If authorized, the

[2] Note that for Ipv6, the extension header plays the role of the encapsulation header in Mobile IP.

gateway provides the user with an IP address that identifies a permanent virtual connection between the user and the network. All Internet traffic for the user is tunneled through the gateway to which the user was originally connected. The gateway maps the IP address of the user to the layer-2 address of the user corresponding to the cell where the user is located. As the user crosses cells, the IP address it had acquired from the gateway remains fixed. However, the mapping of this address to a cell location changes to reflect the new location of the user. In essence, this gateway performs the function of a agent in Mobile IP, by providing the user with an IP address, and tunneling the traffic for that user to its most up-to-date location.

1.1.3.2 UMTS

One example of an industrial system that uses the Mobile IP protocol is the Universal Mobile Telecommunication System (UMTS), which is proposed in [10]. UMTS aims to provide IP level services via virtual connections between mobile hosts and IP gateways connected to ISPs or corporate networks at the edges of the mobile network.

Users are assigned domain names, which are used to identify the ISP that can be accessed to provide Internet connectivity to the user. When a user logs on, it is assigned an IP address by the gateway to which that ISP is connected, also known as the home gateway. A virtual connection is established, consisting of two segments: one segment connects the mobile and some foreign gateway (via the air interface), and another, connects the foreign gateway and the home gateway (via a protocol similar to Mobile IP). The virtual connection is maintained as long as the mobile remains on and the foreign gateway can be changed as the mobile roams from the coverage area of one gateway to another. One can think of the mobile as being linked to the home gateway via an elastic global pipe. To the external world, the mobile appears to be located at the home gateway because it is this gateway that provides the IP address for the mobile. This mobility model is similar to the Mobile IP protocol.

1.1.4 Other Challenges

Another challenge of mobile networking is to support multimedia applications with stringent performance constraints, such as low packet loss and high interactivity. As users move, handoff needs to take place between the user's old point of attachment to the network and the new point of attachment to the network. Handoff may require change of state, not only at routers in the network to which a user is immediately connected, but also at routers inside the Internet that deliver packets to that user. If the number of such routers is large, or the distance from the user to these nodes is large, this change of state can take a long time. An interruption in connectivity due to a slow handoff can cause packet loss, which can significantly lower the perceived quality of these applications by the user. Packet buffering is typically used to handle packet loss. However, packet buffering may result in excessive latency overhead. Real-time applications such as voice depend on packets being delivered at a constant rate and within a certain time budget. If at times of user movement, the network cannot ensure the timely delivery of packets, they become irrelevant and would need to be discarded, to the dissatisfaction of users. Therefore, to assist moving users and maintain the continuity of multimedia traffic, the solution needs to support fast handoffs. To achieve this, handoffs should not involve propagation of information over long distances (hence should be handled in the vicinity of the user location).

Also in order to achieve smooth handoffs, it may be necessary for the mobile user to be connected to multiple points of attachments (referred to as diversity in the literature). However, this becomes difficult as speeds increase and users move continuously across space, frequently changing their points of attachment.

Unfortunately, Mobile IP does not meet the challenge of fast handoff. Rather than attempting to handle rapid network transitions such as the ones encountered in a wireless cellular system, Mobile IP focuses on the problem of long-duration moves. Let us refer to movement that requires fast handoffs as micro-mobility. The reason why Mobile IP does not perform well for micro-mobility should be clear: after it moves to a new subnet, a MH must detect that it has moved, communicate across the foreign network to obtain a COA and then communicate across the Internet to its HA to arrange forwarding. Because it requires considerable overhead after each move, mobile IP is intended for situations in which the MH crosses subnets infrequently, e.g. when the MH remains at a given location for a relatively long period of time.

2 Accelerating Micro-mobility

Many researchers have investigated ways to improve Mobile IP by accelerating micro-mobility. Subnets in the Internet are grouped into domains. Inter-domain mobility is achieved using Mobile IP, while intra-domain mobility is achieved using techniques that are particular to each research scheme. Routers or switches inside the domain keep track of users and deliver traffic to them using their learning databases. To perform this function, these devices effectively implement host-specific routing or switching. Traffic between domains is exchanged via routers typically known as gateways. The basic idea is shown in Fig. 1.

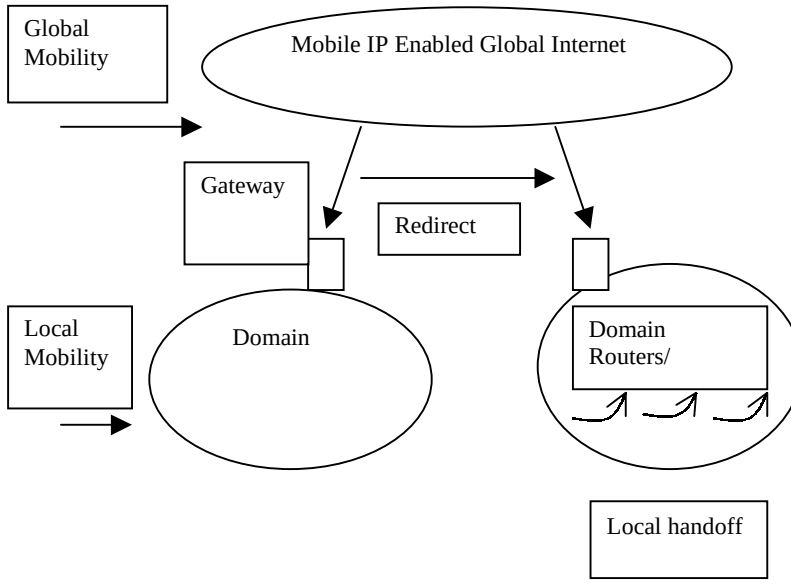

Fig. 1. Typical Micro-Mobility Architecture

2.1 HFA

In [1] hierarchical foreign agents (HFAs) are introduced to smooth out the handoff process when a mobile host MH transitions between subnets. This optimization is accomplished via hierarchical tracking of mobile hosts (MHs) by the foreign agents (FAs) and via packet buffering at FAs.

The FAs of a domain are organized into a tree structure that handles all the handoffs in that domain. The tree organization is unspecified and left up to the network administrator of that domain. One popular configuration is to have a foreign agent associated with the firewall to that domain be the root of the tree (also known as a gateway foreign agent or GFA) and all the other foreign agents provide the second level of the hierarchy.

An FA sends advertisements called Agent Advertisements in order to signal its presence to the MHs. An Agent Advertisement includes a vector of care-of addresses, which are the IP addresses of all its ancestors as well as the IP address of that FA. When an MH arrives at an FA, it registers the FA and all its ancestors with its home agent HA. The registration is seen and processed by the FA, all its ancestors and the HA.

When a packet for the MH arrives at its home network, the HA tunnels it to the GFA. The GFA re-tunnels it to the lower-level FA, which in turn re-tunnels it to the next lower level FA. Finally, the lowest-level FA delivers it to the MH. Therefore, an FA processing a registration should record the next lower-level FA as the other end of the forwarding tunnel.

Mobile IP route optimization extends the use of binding cache and binding update messages to provide smooth handoff via previous FA notification. However, tunneled packets that arrive at the previous FA before the previous FA notification are still lost. Such data loss may be aggravated if the MH loses contact with any FAs for a relatively long period of time. HFA includes an additional FA buffering mechanism. Besides decapsulating tunneled packets and delivering them directly to an MH, the FA also buffers these packets. When it receives a previous FA notification, it re-tunnels the buffered packets along with any future packets tunneled to it. Clearly, how much packet loss can be avoided depends on how quickly an MH finds a new FA, and how many packets are buffered at the previous FA. This in turn depends on how frequently FAs send out beacons or agent advertisements, and how long the MH stays out of range of any FA. To reduce duplicates, the MH buffers the identification and source address fields in the IP headers of the packets it receives and includes them in the buffer handoff request so that the previous FA does not need to retransmit those packets that the MH has already received.

While HFA helps reduce the overhead of handoff by handling handoff closer to the MH, it adds latency due to the need for packet encapsulation and decapsulation at every FA in the FA tree along the path from the CH to the MH. Moreover, scalability issues arise at the root FA and the FAs close to the root of the FA tree because of their involvement in packet tunneling for all the MHs of that domain. Finally, packet buffering results in latency overhead, while encapsulation still generates bandwidth overhead.

2.2 Cellular IP

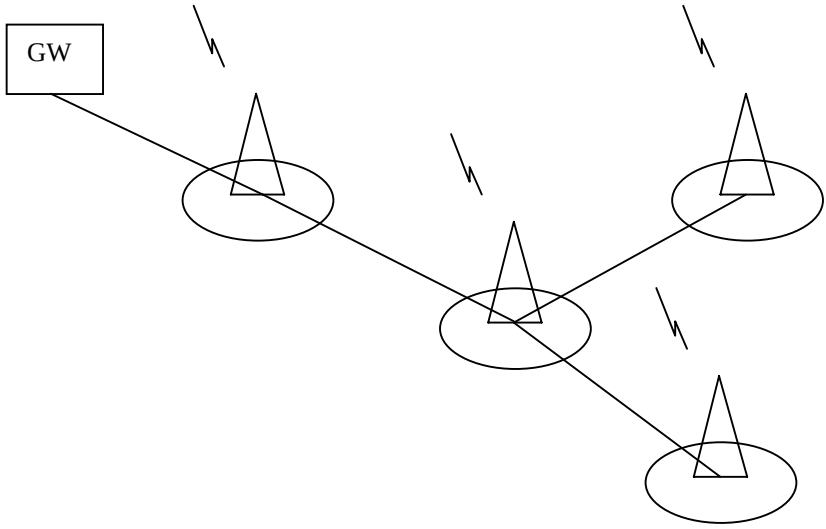

Fig. 2. A Wireless Access Network in Cellular IP. Base stations in an access network are interconnected by wired links. One gateway controls each access network.

Cellular IP access networks, depicted in Fig. 2, are connected to the Internet via gateway routers [9]. Cellular IP uses base stations for wireless access connectivity, for IP packet routing and for mobility support inside an access network. Base stations (BSs) are built on regular IP forwarding engines except that IP routing is replaced by Cellular IP routing. MHs attached to an access network use the IP address of the gateway as their Mobile IP care-of-address. The gateway de-tunnels packets and forwards them toward a BS. Inside a Cellular IP network, MHs are identified by their permanent home address and data packets routed without tunneling or address conversion. The Cellular IP routing protocol ensures that packets are delivered to the host's actual location. Vice-versa, packets sent by the MH are directed to the gateway, and from there, to the Internet.

Periodically, the gateway sends out beacons that are broadcasted across the access network. Through this procedure, BSs learn about neighbouring BSs on the path towards the gateway. They use this information when forwarding packets to the gateway. Moreover, when forwarding data packets from users to the gateway, BSs learn about the location of a user, and use that information to deliver packets sent for that user.

If a packet is received at a BS for a user that is unknown to that BS, a paging request is initiated by the BS. The paging request is broadcasted across a limited area in the access network called a paging area. The MH responds to the paging request and its route to the paging BS gets established. Each MH needs to register with a paging area when it first enters that area, regardless of whether it is engaged in communication or idle. Clearly, how fast paging occurs depends on the size of the paging area and on the efficiency of spanning tree traversal. A small paging area can

help reduce the latency of paging, however it increases the number of paging area required to cover a given area, which in turn increases the signalling overhead imposed on MHs.

We observe that the paging techniques in Cellular IP are similar to those existent in the Groupe Speciale Mobile system (GSM) [13]. Mobile users are located in system-defined areas called cells that are grouped in paging areas. Every user connects with the base station in his cell through the wireless medium. Base stations in a given paging area are connected by a fixed wired network to a switching center, and exchange data to perform call setups and deliver calls between different cells. When a call arrives at the switching center for a given user, a paging request for that user is initiated across all the cells in that paging area. If the user answers, a security check on the user is performed, and if the test passes, the switching center sets up a connection for that user.

Cellular IP supports two types of handoff: hard handoff and semisoft handoff. MHs listen to beacons transmitted by BSs and initiate handoff based on signal strength measurements. To perform a handoff, the MH tunes its radio to the new BS and sends a registration message that is used to create routing entries along the path to the gateway. Packets that are received at a BS prior to the location update are lost. Just like in Mobile IP, packet loss can be reduced by notifying the old BS of the pending handoff, and requesting that the old BS forward those packets to the new BS. Another possibility is to allow for the old route to remain valid until the handoff is established. This is known as semisoft handoff and is initiated by the MH sending a semisoft handoff packet to the new BS while still listening to the old BS. After a semisoft delay, the MH sends a regular handoff packet. The purpose of the semisoft packet is to establish parts of the new route (to some uplink BS). During the semisoft delay time, the MH may be receiving packets from both BSs. The success of this scheme in minimizing packet loss depends on both the network topology and the value of the semisoft delay. While a large value can eliminate packet loss, it however adds burden on the wireless network by consuming precious bandwidth.

Cellular IP specifies an algorithm to build a single spanning tree rooted at the gateway to the access network as we described above. A spanning tree is necessary for the broadcasting of packets, to avoid packets from propagating to infinity if the topology of the access network has any loops. However, because it uses only a subset of the links inside the access network, a single spanning tree can result in link overload if traffic in the access network is high. This can be a significant drawback of Cellular IP as high-density access networks supporting many Tb/s of traffic become possible to deploy. Moreover, a single spanning tree can be prone to long periods connectivity loss. Connectivity loss would make this technology unacceptable as a replacement to wired, circuit-switched technology for telephone communications. Finally, Cellular IP specifies an interconnect between base stations that has a flat hierarchy. As access networks cover more area and exhibit higher pico-cell densities, a flat hierarchy would result in latencies of packet traversal across the access network that are unacceptable.

The description of Cellular IP assumes that originally, each wireless cell (or even pico-cell) constitutes an IP subnet. Consequently, they propose that multiple wireless cells be grouped into one subnet to improve roaming between the cells of one subnet. However, this concept is not new. For example, the 802.11 standard uses Extended Service Sets (ESS) to interconnect multiple 802.11 cells within a single subnet. Cellular IP also proposes two protocols for configuration and routing in IP subnets,

however, LAN protocols already exist to accomplish these goals. For example, the algorithms for building a spanning tree and for learning as defined by the 802 standards are widely deployed and well known.

Nonetheless, it is clear that deploying wireless access networks as single subnets, like in Cellular IP is important for mobility. In this light, it becomes important to increase the size of IP subnets to the largest size possible in order to maximize their effectiveness in supporting IP mobility.

2.3 Hawaii

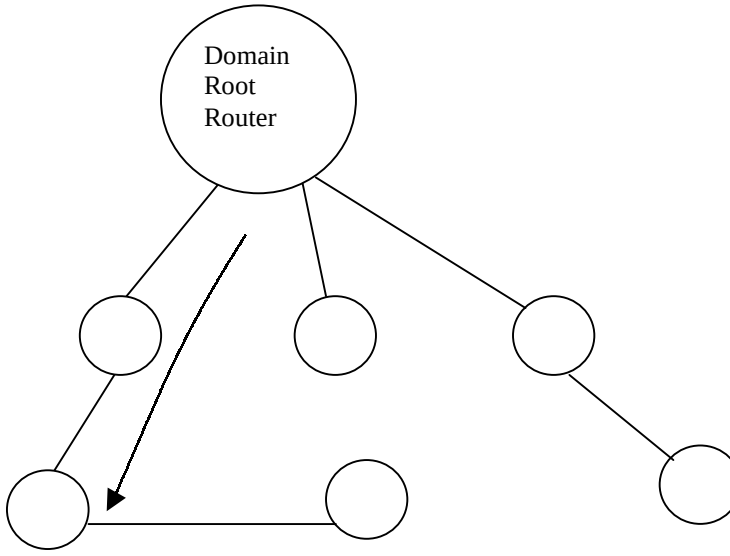

Fig. 3. Diagram of a Domain in the Hawaii Architecture. A domain root router acts as the gateway to each domain. Paths are established between the routers of a domain.

HAWAII segregates the network into a hierarchy of domains, loosely modeled on the autonomous system hierarchy used in the Internet [14]. The gateway into each domain is called the domain root router. When moving inside a foreign domain, an MH retains its COA unchanged and connectivity is made possible via dynamically established paths, as shown in Fig. 3. Path-setup update messages are used to establish and update host-based routing entries for the mobile hosts in selective routers in the domain, so that packets arriving at the domain root router can reach the mobile host. The choice of when, how and which routers are updated constitutes a particular setup scheme. HAWAII describes four such setup schemes, which trade-off efficiency of packet delivery and packet loss during handoff. The MH sends a path setup message, which establishes host specific routes for that MH at the domain root router and any intermediary routers on the path towards the mobile host. Other routers in the domain have no knowledge of that MH's IP address. Moreover, the home agent and communicating host are unaware of intra-domain mobility. The state maintained

at the routers is soft: the MH infrequently sends periodic refresh messages to the local BS. In turn, the BS and intermediary routers send periodic aggregate hop-by-hop refresh messages toward the domain root router. Furthermore, reliability is achieved through maintaining soft-state forwarding entries for the mobile hosts and leveraging fault detection mechanisms built in existing intra-domain routing protocols.

HAWAII exploits host-specific routing to deliver micro-mobility. By design, routers perform prefix routing to allow for a large number of hosts to be supported in the Internet. While routing based on host-specific addresses can also be performed at a router, it is normally discouraged, because it violates the principle of prefix routing. Furthermore, host-specific routing is limited by the small number of host-specific entries that can be supported in a given router. However, this concern can be addressed by appropriate sizing of the domain and by carefully choosing the routers that are updated when a mobile is handed off. One of the problems with the implementation of HAWAII is that a single domain root router is used. This router, as well as its neighbors inside the routing tree can become bottlenecks routers for the domain for two reasons: First, they hold routing entries for all the users inside the domain. Second, they participate in the handling of all control and data packets for that domain. Another disadvantage of HAWAII comes from its use routers as a foundation for micro-mobility support. With cells becoming smaller, it is possible that a larger number of routers would be needed for user tracking and routing in a given area; however, this can become prohibitively expensive.

2.4 Multicast-Based Mobility

Numerous multicast-based mobility solutions have been proposed [2,6,7]. In [2,6], each mobile host is assigned a unique multicast address. Routers in the neighborhood of the user join this multicast address, and thus form a multicast tree for that address. Packets sent to the mobile host are destined to that multicast address and flow down the multicast distribution tree to the mobile host. In [7], packets are tunneled from the home agent using pre-arranged multicast group address, to which a set of neighboring base stations in the vicinity of the mobile host adhere. The most significant drawback of these solutions is that they require routers to be multicast capable; this capability does not exist in the Internet routers of today and would need to be added. In essence, this solution requires that routers learn multicast addresses, in the same way that routers learn unicast addresses in the other schemes for micro-mobility that we discussed. Unlike LAN switches, routers are not designed to learn host addresses, and therefore they would need to be modified for this purpose. Other drawbacks of mobility schemes based on multicast routing are that they require unique multicast addresses to be used, which creates address management complexity and limits the addressing space.

2.5 Micro-mobility and LAN Switching

In all the solutions we presented, fixed IP addresses are used to track mobile users inside a domain. This is done via learning at base stations, routers or agents. Despite the use of IP addresses, which are hierarchical, the addressing structure within a domain becomes non-hierarchical, just like in a LAN. Consequently, these addresses

are tracked in the same fashion as layer-2 addresses in LANs. We make the observation that, in fact, these addresses are tracked in the same way as virtual channel identifiers in circuit-switched solutions such as ATM (employed in UMTS for the tracking of users by foreign gateways). In their original design, routers were not intended for performing tracking of individual host addresses, and consequently do not perform host-specific routing in an efficient way. It is unlikely that routers designs will be modified for this purpose. By design, layer-2 switches track host addresses, hence represent a more suitable solution for mobile tracking inside a domain.

3 A Multi-layer Infrastructure for Mobility

Our view is that an architecture to handle mobility must operate in a hierarchical fashion by providing functionality at multiple layers; namely, the MAC layer, the networking layer and DNS (or the "directory" layer). Each layer is suited for implementing mobility if specific circumstances are met. The MAC layer is ideal for delivering fine grain mobility inside homogeneous networks, by virtue of the fast, cost-effective switching technologies and the address learning schemes available at this layer. Similarly, the networking layer is best suited for implementing coarse grain mobility in cases where mobiles cross subnets and hence require new IP addresses to remain reachable, or when movement happens across heterogeneous networks where MAC layer addresses are incompatible. DNS can further support coarse grain mobility by maintaining an up-to-date directory of users and their IP addresses which can be used to simplify the operation of the networking layer. A description of this architecture as shown in Fig. 4. In the following subsections, we provide a more detailed description of the architecture. Readers who are interested in a complete description of the architecture are referred to [11].

3.1 Extended LAN

Over the past decade, we have witnessed tremendous developments in LAN technologies, such as increases in switch processing by a few orders of magnitude, and increases in link bandwidth and distances (owing to the fiber optics technology). These advances resulted in an increase in the size of LANs, and more recently, their deployment in metropolitan areas. We observed that such networks are well suited for providing mobility to portable IP devices. First, as mentioned earlier, mobile users can roam inside extended LANs without having to update their IP addresses. Secondly, the learning protocol implemented by LAN switches can be used to support diversity and adaptive routing. For these reasons, extended LANs are at the foundation of our network design for mobility.

Given the importance of LANs for mobility, it becomes paramount to answer the following questions:

1. How scalable are extended LANs in terms of number of users, user speed, application bandwidth and latency constraints?
2. What is the appropriate LAN structure and how large an area can it serve?

3. What is the protocol for tracking users in the LAN? How reliable is the LAN and what is its reconfiguration algorithm?

The answers to these questions need to take into account a variety of issues related to the wireless access networks, wired infrastructure, application traffic and requirements and user mobility. In order to minimize cost and maximize performance and reliability, the network design has to balance many parameters such as: processing power and storage capacity in the LAN switches, bandwidth across the wired links in the extended LAN, bandwidth and power consumption in the wireless cells. While a large extended LAN reduces the need for global mobility that can be inefficient, it also requires that the LAN support a larger number of users, and therefore increase the bandwidth requirements at the LAN switches and in the wireless cells in order to carry handoff control messages and user data.

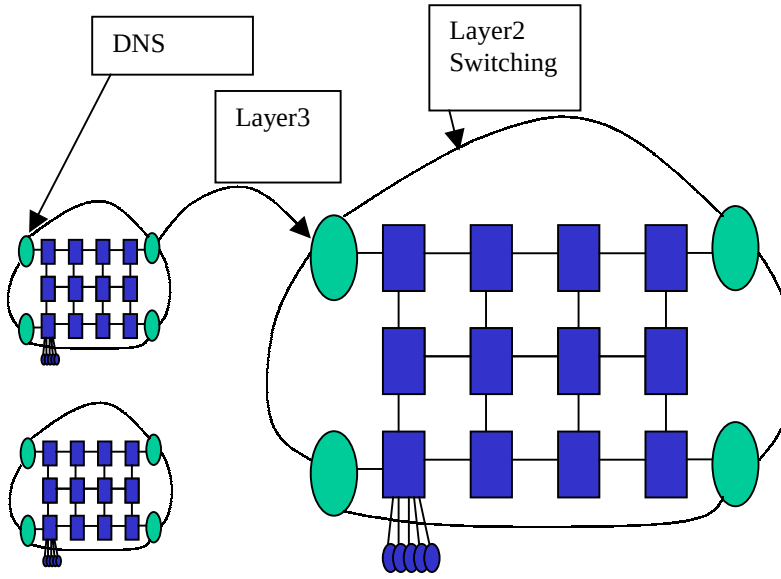

Fig. 4. A Multi-Layer Architecture for IP Mobility

3.2 Dynamic DNS

At the highest level of the protocol stack, dynamic DNS can be employed to track moving users as they change domains. The idea here is that DNS can behave like a directory that stores up-to-date, coarse-grain information on the location of mobile users. When the user enters a new domain and receives a new IP address, a DNS update is sent to ensure the most up-to-date mapping of the mobile's domain name to its IP address. Subsequently, communicating sessions that start following transition to a new domain can benefit from having the latest information and locate the user directly. However, sessions that are ongoing at the time of move cannot benefit from the directory update. This is because DNS lookups are not performed in the middle of

sessions for the purpose of renewing connectivity. Instead, a network layer solution needs to be devised for this purpose. By updating the DNS database, we minimize the need for network layer mobility support and allow for efficient routing for sessions which undergo domain name resolution prior to session establishment and which start following inter-subnet crossings.

For dynamic DNS to work properly, caches that store the mapping of domain names to IP addresses at the communicating nodes need to be either:

1. Binding caches, which guarantee that the latest mapping is given
2. Disabled
3. Have a low TTL value (under a few seconds).

The first choice may give better IP address lookup performance, particularly for slow mobility, however it can be expensive to maintain fast mobility and many users. The second option removes the cost of updating caches at the expense of lookup performance. The third option is a compromise between the first two options.

It is interesting to note the parallels that exist between dynamic DNS solutions and directories techniques in cellular telephone systems such as the Groupe Speciale Mobile (GSM) or the Personal Communication System (PCS) [13,15]. In the GSM and PCS systems, users are identified by a unique phone number. One important feature of the GSM/PCS system is the automatic, worldwide localization of users. The system knows where a user currently is, and the same phone number of valid worldwide. A hierarchy of databases consisting of Home Location Registers (HLRs) and Visiting Location Registers (VLRs) is used to track users. The HLR contains information about the current VLR of a user, and the VLR knows the switching center via which the user can be reached. The HLR/VLR databases are similar to the DNS directories in our solution because they store up-to-date information on the location of every user. Notice that this feature of GSM/PCS renders the mobility management scheme a challenging problem. While this approach eliminates system-wide paging, which vastly reduces the radio link signaling, it introduces remote database lookups that may incur a large amount of wired network traffic and long call setup delay. Much effort has been focused on exploring efficient location management techniques. Extensions to standard HLR/VLR schemes, such as partial replication and caching have been developed to improve wireless call setup performance [15]. We believe that these techniques may apply to our solution in order to achieve an efficient and scalable dynamic DNS implementation.

3.3 IP Mobility

The network layer is important for providing mobility when users roam between different administrative domains, different subnets within the same domain, and possibly between heterogeneous networks. The network layer solution has two components that should be used in combination in order to deliver wide-area mobility. One component requires that a tunneling protocol be used, such as Mobile IP to redirect sessions that are ongoing at the time of a move between different domains. A second component of the solution is using host-specific entries at the routers of a domain to track groups of mobile users inside that domain, as is done in HAWAII. The first solution is important in order to eliminate the problem with DNS-based

tracking that we outlined. The latter solution is important in order to improve the efficiency of roaming by giving users the ability to use a single IP address inside a large domain that extends beyond one subnet. To support a large number of users, routers that implement host-specific routing inside a domain should interconnect via a scalable fabric, and implement a scalable routing protocol.

4 A Case Study

A scalable LAN overlay is proposed to support mobile users in a metropolitan area network. A detailed description of this proposal can be found in [12]. As shown in Fig. 4, the extended LAN is implemented as a grid topology (e.g. the Manhattan Street Network). This is because the grid matches the topology of cities themselves - with the streets being rows and columns -, but also because the grid is scalable by virtue of its distributed nature. Wireless cells are connected to LAN switches in a hierarchical fashion. The hierarchy reduces the number of hops to be traversed when communicating between two access points in the grid, and therefore reduces latency. To connect to the Internet backbone, a scalable and distributed gateway router is necessary. The router needs to scale to support the aggregate traffic to and from all the cells in the LAN. For a large number of cells with many users, this bandwidth can become very large. For example, for a LAN supporting 2 million users, consuming 2 Mb/s each, the routing bandwidth is 4 Tb/s. Furthermore, the router must be physically distributed across many smaller routers to allow for load balancing at the links connecting the LAN switches to the subnet router.

A protocol is designed for the Manhattan Street Network that takes advantage of multiple links in the network and that balances the traffic load across all the links and switches in the LAN. The protocol works by partitioning switches into control and data partitions. Each control partition must have one or more switches in common with every data partition. Similarly, each data partition must have one or more switches in common with every control partition. For example, each row in the grid could be a control partition and each column, a data partition. A protocol similar to the Generic Attribute Registration Protocol (GARP) [16] is used to track users inside a given control partition according to the user location. Data packets for a given user are propagated along a given data partition (as given by the location where the data packet was first injected into the network) until the control partition for that user is reached and the packet delivered to the user.

This LAN design has a number of advantages. The LAN does not rely on a single spanning tree or root switch. This is important for scalability as the LAN extends to large geographical scopes. By exploiting control and data partitions, it minimizes the latency of user location updates without affecting the latency of packet routing inside the LAN. Finally, its operation relies on existing LAN switching techniques and protocols, which makes the solution simple, inexpensive and easy to deploy.

5 Migration Path

To transition to the mobile network of tomorrow, it may not be possible to design the supporting network infrastructure from scratch. Instead, support for mobility may need to be built on existing network structures, such as small subnets controlled by LAN switches and interconnected by IP routers with a small number of host-specific entries. Under such circumstances, one possibility is the use of Virtual Private Networks (VPN) to offer extended LAN connectivity across multiple small subnets. In order to support mobile users in the most effective fashion, the protocol to handle mobile users needs to be flexible enough to operate at different layers in the protocol stack, and versatile enough not to require changes in the implementation of the LAN switches and IP routers of that network. In particular, the protocols running on LAN switches should be based on existing 802 protocols, since they are implemented in hardware and therefore cannot be easily replaced or reprogrammed. The main challenge becomes how to use and optimize existing protocols for the purpose of efficient support for mobility.

6 Conclusions

This paper surveys the state-of-the-art in providing mobility support to mobile users in the Internet. In particular, emphasis is placed on micro-mobility techniques designed to accelerate Mobile IP. One observation is that all micro-mobility work in a similar way by requiring that network devices inside a given geographical area learn about the location of users and keep track of them as they move inside that area. The differences among these techniques are the type of device required to do the learning (it could be an IP router, Mobile IP agent or LAN switch) and the protocols for routing packets using the learning databases. This paper also presents an architecture for mobility, which exploits extended LANs, IP routing and dynamic DNS. One important feature of this architecture is its scalable and efficient LAN design, geared at optimizing IP mobility. By relying on existing technologies, and by virtue of working with Mobile IP, this architecture is also global, cost-effective, easily deployable and compatible with the Internet of today.

References

1. E. Perkins and K. Y. Wang, “Optimized Smooth Handoffs in Mobile IP”. Proceedings of the IEEE Symposium on Computers and Communications, Red Sea, Egypt, June 1999.
2. J. Mysore, V. Bharghavan, “A New Multicasting-based Architecture for Internet Host Mobility”. Mobicom 1997, Budapest, Hungary, September 1997
3. A. Snoeren and H. Balakrishnan, “An End-to-End Approach to Host Mobility”, Mobicom 2000, Boston MA, August 2000.
4. C. Perkins, “IP Mobility Support”, RFC 2002, October 1996.
5. S. Cheshire, M. Baker. “Internet Mobility 4X4”. Proceedings of the ACM SIGCOMM 1996, Stanford, CA, August 1996.

6. A. Helmy, “A Multicast-based Protocol for IP Mobility Support”, Second International Workshop on Networked Group Communication, Palo Alto, CA, November 2000.
7. S. Seshan, H. Balakrishnan, R. Katz, “Handoffs in Cellular Wireless Networks: The Daedalus Implementation and Experience”, Kluwer Journal on Wireless Networks, 1995.
8. J. Scourias and T. Kunz, “An Activity-based Mobility Model and Location Management Simulation Framework”, MSWiM, Seattle, WA, 1999
9. A. Campbell, J. Gomez, S. Kim, A. Valko, C. Wan, “Design, Implementation and Evaluation of Cellular IP”, IEEE Personal Communications, June/July 2000
10. “Naming, Addressing and Identification Issues for UMTS”, UMTS Forum, December 2000
11. Cristina Hristea and Fouad Tobagi “A Multi-Layer Architecture for IP Mobility”, in submission
12. Cristina Hristea and Fouad Tobagi “User Tracking and Routing in Metropolitan Area Networks with a Manhattan Grid Topology”, in submission
13. J. Schiller, “Mobile Communications”, Pearson Education Limited 2000
14. “HAWAII: A Domain-based Approach for Supporting Mobility in Wide-Area Wireless Networks”, R. Ramjee et al, Proc. IEEE International Conference on Network Protocols, 1999
15. “Efficient PCS Call Setup Protocols”, Y. Cui et al, IEEE Infocom, San Francisco, CA, 1998
16. IEEE 802.1d MAC Layer Bridging Standard.
17. http://www.metricom.com
18. “An Architecture for Content Routing Support in the Internet”, M. Gritter, D. Cheriton, USENIX Symposium on Internet Technologies and Systems, San Francisco, 2001