A Novel Approach to Mobility Management

Ron Hutchins[a] *ron.hutchins@oit.gatech.edu*
Tracy Camp[b*] *tcamp@mines.edu*
Philip H. Enslow, Jr.[c] *enslow@cc.gatech.edu*
[a] Office of Information Technology, Georgia Institute of Technology, Atlanta, GA
[b] Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO
[c] College of Computing, Georgia Institute of Technology, Atlanta, GA

In this paper, we propose a novel approach to computer mobility. Our approach allows mobility to be rapidly deployed, as the networking infrastructure required for deployment is available off the shelf. Furthermore, a mobile node does not require modifications in order to use these mobile services. While our approach provides rapid deployment and supports both IP and non-IP protocols, only a subset of mobile usage scenarios are offered. In other words, our approach does not solve all the problems of mobility. We discuss the characteristics of mobility usage, and we list the scenarios our approach supports. We believe that the mobile usage scenarios supported by this method are some of the more common usage scenarios. We also believe that investigations into this method will provide more insights into network and mobility research.

1 Introduction

As technology has matured and computers have become more and more ubiquitous in our society, we have seen the growth in popularity of laptop computers and personal digital assistants (PDAs) like the Psion, the Pilot, the Newton, and the Sharp, in the workplace. Since these devices are portable, many of them are carried and used on airplanes, in cars, and in other non-traditional places. Consider, for example, the number of portable computers used on airplanes today, compared with the number that were seen a decade ago. Staying connected now goes beyond carrying a pager or even a cellular telephone. Email is remotely accessible around the world where Internet access is available, lending a consistency of information access that is not possible with our older communication methods. Though the current access mechanisms are somewhat crude and clumsy, the need to have continuous access to email, and other tools for real time collaboration, is growing.

Over the last few years, wireless and portable telephones and computers have shrunk in size, weight, and power consumption. The "Star Trek" model of communication seems more and more realizable. In fact, the current line of cellular equipment produced by Motorola approaches the form and mechanical functionality of the science fiction show's communicator. Wireless support is available today for portable computers though still very sparsely installed. Several new technologies on the market utilize the 900 MHz, 2.4 GHz, and 5.8 GHz frequencies in order to obtain unlicensed wireless connectivity. These multiaccess products allow LAN-like interconnections among untethered users, permitting these users to change interconnection points with little or no change in their low level configuration.

With the growing utilization of wireless communication, several models of deployment and management of these wireless "loops", both for telephony and/or data services, are being developed. This experience has helped form several models for mobile voice usage [18]. Models for mobile computing usage, on the other hand, are lacking. Intuition confirms the need for mobilizing computer communication, as mobile computing offers a broader set of services than those provided by cellular voice technologies alone. We need, however, to understand the models of mobile usage in order to develop applications that take advantage of the most common user scenarios. Furthermore, we need to deploy mobility rapidly, as the demand for mobile services continues to grow.

The initial models of mobile computing usage to date have been based on current cellular telephony experience, intuition developed by observing deployment of today's technology for mobility support, and current users' own predictions of their mobility needs. These models may or may not be accurate. We should use caution in setting the direction for mobility development and deployment based on this data. In this paper, we discuss an approach that allows mobility to be rapidly deployed on a large scale. Once deployed, we can document

*This work was supported in part by NSF Career Award NCR–9702449.

actual usage patterns, learn the habits and needs of mobile users, and then use this information to develop (or enhance) mobile applications.

1.1 The Wireless/Mobile Network Architecture

The motivation for mobility in computer communication directly follows from the growing mobility of our society: personal mobility, professional mobility, and military mobility. Along with the growing desire for mobility, numerous issues and problems have surfaced. Work has been done in the past in support of mobility, especially in the military domain [19]. This work has laid a solid foundation for today's emerging standards. These new protocols and standards are emerging at many layers of the OSI reference model.

- physical layer (L1) - the flexibility of wireless, both radio frequency (RF) and infra-red (IR) standards [13], and wired standards [12];
- datalink/medium access layer (L2) - the standardized architectures of point-to-point topologies such as cellular telephone endpoints and bridged wireless devices [12, 20];
- network layer (L3) - extending the point-to-points across multiple, politically controlled domains [16, 24, 25, 26];
- transport layer (L4) - producing and maintaining the end-to-end services in the face of multiple and differing error rates, wire speeds, and latencies [3, 4, 7], and supporting standards such as Dynamic Host Configuration Protocol (DHCP) [9] and Service Location Protocol (SLP) [31];
- application layer (L7) - whether the effects of mobility in applications are visible or not [26], and mobile applications such as the Coda file system [17].

Research is being done at each of the above layers to optimize system level functionality and some efforts show promise. But, because of today's sparse deployment of mobile computer communication systems, little hard data exists on mobile usage scenarios and movement patterns. One example is traces of telephone calls from vehicular traffic used for mobility calling and movement patterns are reported in [14]. This telephone data, however, may not be applicable to mobile computer communication. Although metropolitan and wide area mobile communications are being offered through several products such as the Ricochet network from Metricom, these implementations of wireless connectivity do not solve the mobile addressing problem.[1] Users of the system are restricted in their use of a permanent station identity.[2] Until the Internet mobility standards are complete and implementations of these standards are offered from major vendors, specific data on mobile usage patterns will be sparse, and applications that take advantage of common usage patterns will be slow to appear.

1.2 Broadcast Groups and Early Deployment

In this paper, we first consider local area networks, or LANs, mapped to datalink (L2) local broadcast groups [1]. We then propose a novel approach for the implementation of computer communication mobility. This approach, which uses existing standardized hardware and software without modification to participating hosts, offers rapid deployment of mobility for a subset of possible mobile usage scenarios and for a sparse population. Once a mobile infrastructure is deployed, we can then begin to observe how mobile users communicate. In this work, we assume mobile usage scenarios imply computer communication and services with broader requirements than voice services.

This paper is organized as follows. In Section 2, we propose using bridged broadcast groups for mobility, and we discuss the scalability problem that exists. Then, in Section 3, we consider the characteristics of mobile usage. We present the background on Mobile IP and the mobility enabling parts of IPv6 in Section 4. In Section 5, we describe a novel approach that provides rapid deployment of mobility, and we discuss the current status of this work on the Georgia Tech campus. Future plans and challenges are described in Section 6 and followed by our conclusions in Section 7.

[1] Ricochet uses dynamic IP address assignment for mobile nodes and the Post Office Protocol (POP) [23], a proxy mail server architecture, for users' email access. No permanent node address allocation is available except for commercial customers using the Metricom gateway. See `www.ricochet.com` for more details.

[2] The necessity of a permanent IP address for a host is not universally accepted, but the convenience of this permanent network layer identity is acknowledged by a large number of experienced computer users. While a fully-qualified domain name provides support for permanent end host identity, problems exist with this mechanism in the mobile environment today. Future work will discuss the use and meaning of an IP address in this context.

2 Exploring an Example

Let us consider as an example an IP-based, campus LAN environment implemented with equipment that creates a bridged network infrastructure.[3] By definition, broadcast packets are passed transparently across this bridged LAN and unicast packets are delivered without assistance from an L3 router. Broadcast in this context implies selecting destinations for packet delivery, not transmission of electrical signals, i.e., delivery of the same data, not the same signal, to multiple (all) stations transparently.

If wireless base stations are deployed to cover the bridged network such that transparent handoff is supported, a wireless mobile infrastructure is created. In this environment, no network layer address changes are necessary when mobile nodes relocate within the network. Node identity, based on a static mapping of IP address to host name via the domain name system (DNS) [21], is permanent. This identity function is preserved across location changes; that is, the bridged network supports any appropriately assigned IP address at any point on the local network through a flat allocation of network addresses [1].

In this environment, as long as a mobile node is within range of a radio base station attached to the bridged network, smooth, continuous operation is possible. Since a campus network is generally under a single political domain, access to services such as printing, news services, email services, and the Web are reasonably easy to architect. Many network operating system companies have made fortunes building small office support systems (albeit, not using wireless) by exploiting such a bridged local broadcast group environment, e.g., DEC's LAT [11] and Microsoft's NETBEUI [6]. From the perspective of the application, the bridged implementation of a local broadcast group solves most of the problems associated with mobility. Burst errors and latency, due to fading and retransmission, are not solved. However, these problems are due to the use of wireless technologies.[4] Adapting bridged broadcast groups for wireless access is trivial, but this technique is not necessarily scalable.

Very significant problems exist in the scalability of bridged networks. Generally, broadcast packets are used to implement some of the services across the bridged network (e.g., the Address Resolution Protocol or ARP [28]). Although datalink bridges limit the scope of unicast packets, there is no practical way to constrain broadcast packets other than by limiting the size of the broadcast domain. There is no specific recipe for maximum size of a local broadcast group; however, experience has shown that once critical mass is reached, generally in the range of several hundreds of end hosts, "broadcast storms" occur at some frequency.

There are several different causes for broadcast storms; most of which, however, are based on a similar situation. One host transmits an L2 broadcast packet which encapsulates an L3 header and data. This packet is received by all other hosts on the local network. The L3 address in this packet is incorrectly specified, which causes all receiving hosts to re-forward the packet as an L2 broadcast packet. Since the incorrect L3 address is not repaired before forwarding the packet, the process repeats for each of the received packets. This leads to an exponential growth in L2 broadcast packets on the local network, until mechanisms in the network layer (like TTL) dampen out these transmissions. Broadcast storms create problems for networks, as the unbounded load produced can cause service outages of several minutes' duration [1, 5]. In the late 1980's, the broadcast storm problem drove the subdivision of bridged networks into smaller broadcast domains. These local broadcast groups were then interconnected at the network layer by routers using hierarchical addressing (implemented via some form of subnetwork mask and broadcast address for each network interface on the LAN) to simplify routing tables. Since these routers did not forward broadcast packets, the critical mass of the large local broadcast domain was broken.

Moving beyond a single IP subnetwork solves the local broadcast group scalability problem, but breaks the mobility support desired. Current network engineering practice encourages designs that use multiple IP subnets and small local broadcast groups. This subnetwork hierarchy, which implies a fixed hierarchy of addresses as described above, creates problems with mobile computer addressing. Any node participating in a local broadcast domain whose host IP address is not assigned from the hierarchical address range specified for that local broadcast domain (from the same Class C block

[3] Although the mobility mechanisms described in this paper do not depend on IP, the ubiquity of the IP protocol suite in today's networks makes support for IP in mobile implementations necessary. Our examples describe an IP network for this reason.

[4] In circumstances where a mobile node may conveniently be disconnected from the network infrastructure during a move, using 10bT Ethernet with RJ-45 connectors will remove the impact of these problems. For example, consider students taking their laptops to/from classrooms and residence halls, going into sleep mode on the computer or powering down while they walk.

of addresses, for example) cannot participate in L3 communication with the other LAN connected hosts, including the router connecting that LAN to other networks, without modifications to its communication protocol stack. **It is this need to modify hosts in order to maintain a permanent host identity across changes in the address hierarchy that has presented the greatest challenge in implementing mobile computing protocols.**

Computer users who must stay connected when they travel have used different techniques in the past to deal with this problem. Dial-in serial line Internet protocol (SLIP) [29] or point-to-point protocol (PPP) [30] services permit a remote user to maintain his or her assigned IP address, and therefore the address-to-name mapping, by utilizing long distance telephone services. Using these cost per minute services, however, can get expensive. A dynamic implementation of the DNS system, augmenting DHCP services for temporary address assignment, has also proven useful for mobility support. (See [22] for a discussion of the tradeoffs in supporting dynamic addresses over the Internet.) This solution has problems for some mobile usage scenarios, due to the problems associated with maintaining DNS cache consistency across the Internet. Section 5.2 will discuss permanent and temporary addresses in more detail.

3 Characteristics of Mobile Usage

Much of the current mobile communication research attempts to solve the general problem of mobile usage, i.e., across all granularities of time and distance. Under this general model, a mobile user could maintain a TCP session across the local network or across the country, while traveling at speeds that would require changing the network attachment point every few hours or every few seconds. The complexity required to create this transparent interconnection across this range of distance and time is large. Since not every mobile usage scenario has such varied requirements, this level of complexity will be unnecessary for some users. In the long term, this level of generality in providing mobility may prove necessary, so work should continue in this area. In the short term, however, other solutions to the mobility problem should be examined, even if these solutions only enable the most common mobile usage scenarios. These alternate solutions should enable rapid deployment due to lower complexity requirements and the use of current off-the-shelf components in creating the infrastructure. Also, any implementation will create a breeding ground for improvements and advances in mobility and associated technologies.

A survey of current computer communication applications (such as email, news, the Web, and office automation tools) helps identify the parameters (and their extreme values) for mobile usage scenarios:

- distance traveled - the distance covered by the user's movements - a mobile user only covers a fixed, small distance while mobile (e.g., students on a college campus during classes) **OR** a mobile user travels an unbounded large geographical distance (e.g., a sales professional traveling over a regional territory). This parameter determines the extent of the necessary infrastructure deployment.

- duration at a single connection point - length of time the user stays connected through the same subnetwork point of attachment (SNPA) - a user travels while disconnected, or only maintains a connection when stationary (e.g., a professional working in an office and then migrating home before activating another connection) **OR** a user traveling at a velocity which mandates changing SNPA relatively quickly (e.g., moving through a picocell infrastructure or traveling in a car or airplane while connected).

- session length - total active session length - short active session length (e.g., a user connects, sends an email message, and disconnects) **OR** long active session length (e.g., a user downloads a large file).

One interesting relationship in the above parameters is the ratio of session length to duration at an SNPA. If the session length is shorter than the time between network connection point changes, a user may be mobile and yet **perceived as stationary.** We discuss this relationship further in Section 5.2.

4 Mobile IP and IPv6

Mobile IP [24], a standard for mobility in the IPv4 Internet, and the mobility enabling parts of IPv6 [8, 16] define mobility around four separate components:

- the mobile node - a mobile computer containing a modified communication protocol stack.

These modifications allow the mobile computer to mimic being on its home network while away from home.

- the home agent - a host on the mobile node's home network that proxies for (represents) the mobile node when it is away from home. The home agent maintains registration information for the mobile node, tunnels information for the mobile node to the foreign network, and responds to address resolution queries for the mobile node on the home network.

- the foreign agent - a host on the remote network that enables the mobile node to obtain local access to the "foreign" network and to register with its home agent. This function may be incorporated into the mobile node itself by using DHCP [9] to obtain a temporary local LAN address for the mobile node.[5]

- the correspondent node - any host, either fixed or mobile, that is communicating with the mobile node. Route optimization, an option for Mobile IP, can facilitate direct communication with the mobile node, rather than mandate that traffic be routed through the home agent to the mobile node (see [15, 27] for details). In other words, a change to the correspondent node is needed to enable it to communicate directly with a mobile node.

The best case scenario for Mobile IP occurs when the correspondent node has implemented route optimization. In this situation, the correspondent node may receive binding updates from the mobile node, i.e., packets which specify the SNPA (temporary address) of the mobile node. The correspondent node then updates its local routing information allowing packets to be routed directly to the mobile node bypassing the home agent.

Although Mobile IP implements the most general case of mobile usage, it requires modification of every host that plans to be mobile. Furthermore, with some current implementations of Mobile IP, an exchange of data between network managers on both the home and remote networks is necessary to set up the authentication required for the home agent [24] and DHCP services [9]. In the following section, we propose a new model for mobility support that will match the major performance characteristics of Mobile IP in an L2 implementation and simplify mobile usage for some mobile scenarios.

[5]The two addresses used by Mobile IP, one permanent and one temporary for the same network node, represent the core of a discussion on the meaning of an IP address in an IP internetwork (especially in relationship to mobility). The IP address may represent the permanent network layer identity of the node or it may identify the temporary SNPA.

5 A Novel Approach

Many of today's standard LAN protocols have been designed to utilize broadcast delivery. Aloha, Ethernet, token ring, and others rely on this feature for their functionality and simplicity. Since multicast addressing can be viewed as a subset of broadcast, an implementation of multicast communication provides a special case of local broadcast groups. The consistent feature of all these group implementations is that every station in the group receives the same packets (though not necessarily the same electrical signal[6]). Therefore, protocols such as ARP function properly, regardless of the geographic distribution of the broadcast domain. One new protocol that supports LAN emulation through distributed broadcast domains has been standardized by the ATM forum: LAN Emulation (LANE) [1, 10]. LANE was created to smooth the transition from legacy network standards, such as Ethernet, to end-to-end ATM services. Though it is still questionable whether ATM end-to-end services will prove to be generally viable, LANE has found a niche in many campus networks. On many campuses, LANE is used to tie Ethernet, token ring, and native ATM nodes together through an ATM campus backbone network.

With the current deployment of regional ATM networks and with LANE 1.0 available on many campuses, a fixed network infrastructure for rapid deployment of mobility support is realizable. We discuss further details of the LANE implementation in Section 5.1. In Section 5.2, we describe the similarities between Mobile IP and the LANE protocols. These similarities allow a LANE implementation on an ATM network to function as part of a mobility enabling infrastructure. (This infrastructure, as stated earlier, only provides services for a subset of the several possible mobile usage scenarios. For other mobile usage scenarios, modifications to the LANE implementation would be necessary. We address some of these modifications in Section 6.) In Section 5.3, we describe the current status of our proposed infrastructure pilot testing.

[6]L2 bridges allow store and forward delivery of a packet to all stations on (perhaps multiple) interconnected Ethernet segments. The electrical signaling path is broken to separate the collision domains.

5.1 The LANE Technology

LANE 1.0, besides allowing ATM attached hosts to directly participate in a logical IP subnet (LIS), maps legacy network protocols (e.g., Ethernet, etc.) to ATM. LANE implemented on network edge devices (ATM to Ethernet converters, basically) supports Medium Access Control (MAC) bridging over ATM [1]. LANE utilizes a LAN emulation client (LEC), a LAN Emulation Configuration Server (LECS), a LAN Emulation Server (LES), and a Broadcast and Unknown Server (BUS) to implement broadcast domains across ATM virtual circuit infrastructures.

A LEC (LANE client) is identified by a unique ATM address; the LEC is also associated with one or more MAC addresses, possibly mapped to hardware Ethernet ports on the LEC, which are reachable through the unique ATM address. The LEC provides a LAN service interface to any higher layer entity. The LES (LANE Server) fulfills the control function for a particular Emulated LAN (ELAN). For each ELAN, there is only one logical LES. If a LEC belongs to a particular ELAN, then the LEC has a control relationship with that ELAN's LES.

Each LEC is associated with only one LANE BUS (Broadcast-and-Unknown Server) in an ELAN. Each BUS is a multicast server that provides two services:

1. flood unknown destination address traffic, and
2. forward multicast and broadcast traffic to clients within a particular ELAN.

All ELANs in a particular administrative domain are served by one LECS (LANE Configuration Server). The LECS assigns each LANE client to an ELAN in the domain, by directing the LANE client to the LANE server that corresponds to the ELAN. If the LES is statically configured into the LEC, then the LECS is not necessary for the operation of the ELAN.

Communication between the LANE components is implemented with virtual circuits (VCs). VC connections are created from each ATM device to the LES/BUS for control and address resolution, i.e., MAC to ATM address mapping. Broadcast and multicast packets are transmitted by replicating them (flooding[7]) across multiple VCs; data between two LANE clients traverses a VC that interconnects these two LANE clients. We assume that each LEC has been preconfigured with the appropriate LES address and no communication to the LECS is necessary. Figure 1 illustrates communication in a LAN emulated with LANE 1.0 over an ATM backbone network. The following VCs, shown in the figure, are created:

1. VCs between LEC and LES for control of one-to-one communication (e.g. ARP requests)
2. VCs from LES to LECs for control of group communication (e.g. ARP replies)
3. VC from LEC to BUS for multicast send from a LEC (original multicast frame for replication to all LECs)
4. VCs from BUS to LECs for multicast forward (replicated multicast frames retransmitted to all LECs)
5. VC between two LECs for direct data exchange

[7] Flooding in this instance refers to the retransmission of a received data packet over multiple outgoing interface ports (physical or virtual) concurrently, implementing a type of broadcast functionality.

5.2 The Example Reconsidered

We now re-consider the example described in Section 2. ATM devices are utilized for backbone infrastructure across many campuses, and even across a few regions. At appropriate places on a campus, or on multiple campuses, standardized wireless base stations could be placed on the Logical IP Subnet (LIS) which utilizes LANE over the ATM backbone. Mobile computers that have the appropriate wireless equipment installed can then participate in the emulated LAN across the entire infrastructure. *To participate, no change in the mobile computer's configuration is required.*

In a network where hosts are connected via ATM interfaces utilizing LANE, a system of services can be architected so that major services (such as mail, news, web proxy, time, and name services) can appear on many different LANE subnets. With this scenario, much of the traffic of a subnetwork can be kept local to a geographically distributed LANE subnet. This localization of traffic can be used to optimize a mobile subnet at the expense of maintaining state for different LANE subnets on the machines providing services.

The LANE components map fairly closely to the Mobile IP components:

- the mobile node = a mobile computer without a modified communication protocol stack connected through a LEC edge device via Ethernet or a wireless interface (such as Wavelan or Proxim);

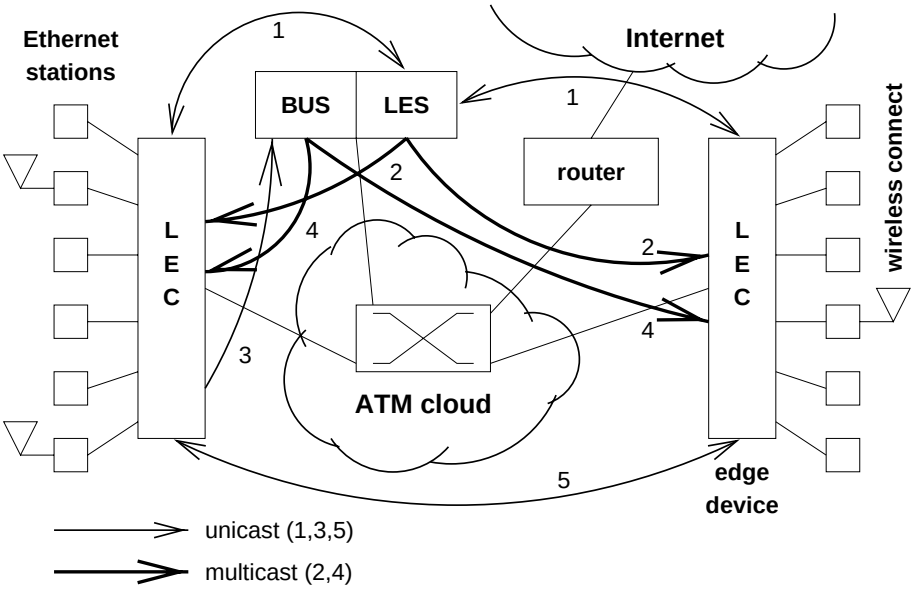

Figure 1: Virtual Circuit Mapping without LECS.

- the home agent = the LAN Emulation Server(LES), which manages security and performs ATM to MAC address mapping, and the Broadcast-and-Unknown Server (BUS) which participates in IP address resolution functions on the subnet;

- the foreign agent = the LAN Emulation Client (LEC), which manages bridge tables of MAC addresses; and

- the correspondent node = any host, fixed or mobile, that is communicating with the mobile node.

Mobile IP defines two network layer (L3) addresses for a mobile node: a permanent IP address (a network layer identity that is mapped to the fully qualified host name, e.g., x1.ferd.com, through DNS) and a temporary IP address (a subnetwork point of attachment - SNPA). LANE connected stations have a single L3 address mapped across SNPA changes to multiple L2 (ATM emulated datalink) addresses. This difference in IP address mapping is significant. To illustrate this difference, we compare Mobile IP to the ATM LANE 1.0 implementation of mobility. Consider a mobile node M which has implemented Mobile IP and has a permanent IP address of w.x.y.z issued by M's home network administrator, and a temporary IP address of a.b.c.d issued by some foreign network that supports Mobile IP. In this example, we assume that both w.x.y.z and a.b.c.d are drawn from a Class C address space. This mobile node M, in general, may communicate in three different scenarios[8]:

1. Mobile node M, either on its home network or on a foreign network, communicates with correspondent node P. P may be either mobile or fixed. Both M and P have permanent IP addresses assigned from M's home network. In other words, both M and P have IP addresses with the same network part, w.x.y.

2. Mobile node M, while away from home and attached to a foreign network, communicates with correspondent node P. P may be either mobile or fixed. P's permanent IP address has the same network address part as the temporary address assigned to M on the foreign network. In other words, M's temporary address and P's permanent address have the same subnetwork part, a.b.c in our example, and are thus on the same subnetwork.

3. Mobile node M, either at home or away from home, communicates with correspondent node P. P may be either mobile or fixed. Neither P's permanent (or temporary, if mobile) IP address bears a relationship to either M's temporary or permanent IP address. P is a station

[8]One other scenario exists, i.e., both M and P are mobile and both are on the same foreign network with temporary addresses from the same network space. This scenario is not relevant to this discussion and, therefore, ignored.

that bears no relation to M's home or foreign network (q.r.s, for example).

In Mobile IP, if the mobile node is away from home, all three scenarios above require active participation from the home agent, unless route optimization is implemented on the correspondent host and a cached address is available. In LANE, when both stations are in the same LIS (the first scenario above), a VC between the two LECs allows direct communication between the mobile node and the correspondent node, without intervention by the LES/BUS or a router. In other words, with LANE, all communication is handled via L2 data transfer in the first scenario above, even if the mobile node is away from home. In the second scenario, since there is no temporary IP address with the LANE implementation, a station must go through a router (L3) to communicate with the mobile node. This situation will change with coming protocols like Multiprotocol over ATM [2], which implements a "route once, switch many" model. In this approach, a "cut thru" path is learned after the first packet is routed, short circuiting the path between source and destination nodes on the same ATM infrastructure. When the third scenario occurs (i.e., the two communicating stations are on different logical IP subnets not interconnected via a common ATM infrastructure), the use of the "home" L3 router is necessary in LANE today. This situation is worse than that for Mobile IP, since a mobile node in Mobile IP can send traffic directly to the local station[9].

The above scenarios illustrate that the LANE implementation of mobility is optimized for communication between nodes in the same LIS. Past traffic patterns show that traffic sources usually send to destinations that reside on the same IP subnetwork (e.g., LANE subnetwork) [25]. Even where new applications like the Web change usage patterns, distributed storage, like Web caches, will bring significant amounts of this traffic back to local network attached devices. Thus, it is important to explore the performance of our LANE solution to the mobility problem.

Slow moving mobile nodes, and nodes that do not maintain a session during movement to another LEC will benefit from the LANE mobility approach. The complexity of the mobility infrastructure is contained in the emulated datalink foundation, which is built from the underlying ATM services. Since this is a standardized, off-the-shelf implementation, the cost to obtain mobility support should be much lower than the cost of using special purpose devices and software. Furthermore, since LANE is already supported on many campuses, there is little added management burden to support mobility on the existing infrastructure.

One constraint in the LANE implementation is due to the scaling problem discussed in Section 2- the emulated LAN must be "sparsely populated" with mobile nodes to avoid broadcast storms. (How sparse the emulated LAN must be needs to be determined in practice.) The possibility of virtual circuit depletion also constrains the size of the emulated LAN, due to the high rate of virtual circuit consumption with current LANE protocols. We believe, however, that the maximum size for this emulated LAN is "large enough" (i.e., encompassing several hundreds, possibly even thousands, of stations distributed across several tens or hundreds of SNPAs in a large geographical area) to permit broader work on mobility enabled applications. These emulated LANs can be overlaid, for example, on high performance national networks such as the National Science Foundation's Very high performance Backbone Network System (vBNS). In other words, these emulated LANs can be used to create several separate administrative domains, each of which provides mobility support for the several hundred mobile users attached. Though mobility between the different administrative domains is not solved directly, the utility of having several large deployments of mobile users should create the critical mass necessary to discover the habits and needs of mobile users.

The largest benefit from this proposed implementation is that no modification to mobile nodes and correspondent nodes is needed. This architecture, though not viable long term without changes, will support many of the common mobile usage scenarios discussed in Section 3. Specifically, preconfigured geographically distributed LANE subnetworks can transparently support applications from mobile nodes when the total session length (S) is shorter than the time between subnetwork point of attachment changes (T), that is, when the ratio of S/T is less than or equal to one. S could represent more than one completed session as long as no session is ongoing during a change in the subnetwork point of

[9] A Lane 1.0 Phase 2 Protocol implements redundant and replicated LES/BUS functions. This addition removes the single point of failure that resides in the LES/BUS but does not resolve the need for multiple entry/exit points to route traffic for a broadcast group. Route optimization requires the modification of every correspondent node or the implementation of IPv6 with the mobility enhancements on each correspondent node. Until this becomes common, the disadvantage of this single entry/exit point may be acceptable when utilizing LANE.

attachment. For example, a corporate staff member working at home can create multiple (perhaps long) sessions from that attachment point. The staff member can then close out any open sessions, move to the office (even if the office is in another region), and create multiple (perhaps long) sessions from the new attachment point while in his/her office. The staff member cannot, on the other hand, create a session that continues to function during movement, unless the movement is across ports on the same LEC device via a wireless interface.

5.3 The Present Situation

In the fall of 1997, Georgia Tech initiated a Student Computer Ownership program that has grown to more than 5000 student owned nodes. These student computers are connected by a LANE enabled Ethernet to ATM infrastructure. Recently, a regional implementation of the Internet2 initiative, called the Southern CrossRoads (SoX), began rapid deployment across thirteen Southeastern states and the District of Columbia. (See `www.internet2.org` and `www.sox.net` for details on these two initiatives.) The SoX network consists of high speed (at least STS-1) SONET links tying ATM switches together. It forms a virtual circuit based network that can support LANE subnetworks across multiple campuses. With students taking their laptop computers to class, and regional university staff taking their laptop computers to other campuses for network planning and applications research, the motivation to enable computer mobility across both the Georgia Tech campus and the region has risen greatly. With the SoX network, the infrastructure to provide this desired mobility support regionally is becoming available.

Georgia Tech's campus network backbone, GT-Net, currently supports more than 128 logical IP subnetworks. These logical IP subnetworks are implemented through LANE 1.0 broadcast groups. More than half of these subnetworks are distributed through the residence halls; the other half are deployed through different colleges and schools on campus. Over the past 15 months of operation, there have been few problems with these emulated LANs. Currently, we are deploying several wireless trials in buildings across Georgia Tech, and possibly between several campuses, to provide transparent mobility to users over the LANE subnetworks. These wireless additions will offer mobility services to users across the region, i.e., users able to connect to the wireless LANE subnetwork.

6 Future work

There are several issues to resolve to complete the functionality of our proposed implementation, i.e., utilizing LANE protocols to implement mobility. First, the mapping between MAC layer addresses and ATM addresses is kept in a Content Addressable Memory (CAM) on the LES. When a mobile device moves from one LEC to another, this CAM mapping must be updated. If this update is not completed quickly, then continuity of the session can not be provided. Some LANE systems implement a five minute cache flushing algorithm to remove old entries in the cache. On other equipment, a link status change on a port causes the cache to be flushed immediately. Early data captured during an example host move indicates that the address tables on this equipment are correct within approximately one minute after a move. This delay may be acceptable when users travel, for example, between work and home without an active session. However, for users that desire a continuous session during movement, or users requiring a cache update under one minute, this delay may be unacceptable. In other words, a user must currently disconnect for between one and five minutes when moving to another LEC. When a mobile node moves from one Ethernet port on an edge device (LEC) to another port on the same LEC (see Figure 1), only the local MAC bridge table entries must change. To update the bridge table, the cache entry must be deleted (1 - 5 minute timeout) or the mobile node must transmit a packet (standard bridge table management). Until this update is transmitted or the timeout event has occurred, packets destined for the recently moved station will be incorrectly delivered. Once a timeout has flushed the bridge cache, packets from new addresses are flooded to every port on the bridge in the normal way.

A second issue is virtual circuit handoff. This issue needs further investigation, as we want to generalize our implementation to include more mobile usage scenarios. The ATM forum has instituted a working group to create a Mobile ATM solution; this working group is addressing the virtual circuit handoff issue. Considering solutions to the CAM and VC handoff problems is a target for future work, as is the issue of the single router port handling the entire distributed subnetwork.

7 Conclusions

There are many desired mobile usage scenarios that are directly supported by our LANE/wireless de-

sign. Other scenarios are not easily implemented in this environment due to TCP's response to bursty errors being mistaken for network congestion, virtual circuit handoffs, and cache consistency across multiple LECs. Similar problems motivate research in both Mobile IP and Mobile ATM.

The major motivation behind this work is to remove some of the complexity of mobility from the end hosts and place a (probably small) burden for mobile support on the underlying network. Adding this support inside the network, however, could impact network scalability. To avoid scalability problems, Internet protocol designers and implementers have gone to great lengths to keep the network "dumb" and put the necessary intelligence in the attached hosts. This design has allowed the Internet to scale well, even as it experiences tremendous growth. Mobile IP and the mobility enabling parts of IPv6 require host changes, which allow network designers to continue to build "dumb" networks. In this work we investigate an alternative approach: adding mobility intelligence to the network infrastructure. Placing intelligence inside the network, at least around the edges of the backbone that provides mobility services, should enable mobility while keeping the scaling problem manageable. If the amount of information maintained inside the network necessary to enable this mobility can be kept compartmentalized in these edge devices, the scalability of the Internet will not be adversely affected.

We are not suggesting that the motivation for continued work on Mobile IP and Mobile ATM has changed. A solution to the general problem of mobility may be important in the long term. Our implementation approach is important for the interim period; that is, we want an inexpensive mobility infrastructure that can be deployed rapidly. On campuses where a significant number of students purchase laptop computers, the cost differential to modify software on a per host basis versus supporting mobility on the network itself is drastic. Since laptop mobility support is possible with off the shelf, standard protocols that are currently supported on campus networks, the time to obtain mobility support is greatly reduced. Once computers are mobilized on campuses, we can learn a lot about habits and needs of those mobile users. This information is vital for applications to take advantage of common mobile usage patterns.

Acknowledgement

We thank the anonymous reviewers for providing helpful suggestions that improved the quality of this paper.

References

[1] A. Alles. *ATM Internetworking*, Cisco Systems, May 1995.

[2] ATM Forum. Multi-Protocol Over ATM (MPOA), A Brief Description. http://www.atmforum.com/atmforum/library/mpoa.html, September 30 1998.

[3] A. Bakre and B. R. Badrinath. I-TCP: Indirect TCP for Mobile Hosts. In *Proceedings of the 15th International Conference on Distributed Computing Systems*, pp. 136-143, June 1995.

[4] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. Katz. A Comparison of Mechanisms for Improving TCP Performance over Wireless Links. In *Proceedings of ACM SIGCOMM*, pp. 256-269, 1996.

[5] D. R. Boggs, J. C. Mogul, and C. A. Kent. Measured Capacity of an Ethernet: Myths and Reality. *Computer Communication Review*, vol. 18, no. 4, pp. 222-234, August 1988.

[6] M. Brain. Network Communications Using the NetBEUI Protocol: Named Pipes and Mailslots for NT and Chicago. *Dr. Dobb's Journal*, pp. 82-87, October 1994.

[7] K. Brown and S. Singh. M-TCP: TCP for Mobile Cellular Networks. *Computer Communication Review*, vol. 27, no. 5, pp. 19-43, October 1997.

[8] S. Deering and R. Hinden. Internet Protocol, Version 6 (IPv6) Specification. *Request for Comments 1883*, December 1995.

[9] R. Droms. Dynamic Host Configuration Protocol. *Request for Comments 2131*, March 1997.

[10] B. Ellington. LAN Emulation Over ATM Specification - Version 1.0. *ATM Forum 94-0035R9*, March 1994.

[11] D. G. Hirsh. Flush Times for LAT. *DEC Professional*, vol. 9, pp. 60-64, February 1990.

[12] IEEE 802.3 Standard. *Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, New York, IEEE, ISBN 1-55937-555-8, 1996.

[13] IEEE 802.11 Standard. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, New York, IEEE, ISBN 1-55937-935-9, 1997.

[14] J. Jannik, D. Lam, N. Shivakumar, J. Widom, and D. Cox. Efficient and Flexible Location Management Techniques for Wireless Communication Systems. In *Proceedings of the Second ACM International Conference on Mobile Computing and Networking (MOBICOM '96)*, pp. 38-49, November 1996.

[15] D. Johnson. Scalable Support for Transparent Mobile Host Internetworking. *Wireless Networks*, vol 1., pp. 311-321, 1995.

[16] D. Johnson and C. Perkins. Mobility Support in IPv6. *Internet Draft*, draft-ietf-mobileip-ipv6-04.txt, November 1997.

[17] J. Kistler and M. Satyanarayanan. Disconnected Operation in the Coda File System. *ACM Transactions on Computer Systems*, vol. 10, no. 1, pp. 3-25, February 1992.

[18] D. Lam, Y. Cui, D. Cox, and J. Widom. A Location Management Technique to Support Lifelong Numbering in Personal Communications Services. *Mobile Computing and Communications Review*, vol. 2, no. 1, pp. 27-35, January 1998.

[19] B. Leiner, R. Cole, J. Postel, and D. Mills. The DARPA (Defense Advanced Research Projects Agency) Internet Protocol Suite. *Computers and Communications Integration: The Confluence at Mid-decade, INFOCOM '85*, April 1985.

[20] Lucent Technologies. *Wavelan Documentation and FAQ*, www.wavelan.com, August 1997.

[21] P. Mockapetris. Domain names - Implementation and Specification. *Request for Comments 1035*, November 1987.

[22] P. Mockapetris and K. Dunlap. Development of the Domain Name System. *Computer Communication Review*, vol. 25, no. 1, pp. 112-122, January 1995.

[23] J. Myers and M. Rose. Post Office Protocol - Version 3. *Request for Comments 1939*, May 1996.

[24] C. Perkins. IP Mobility Support. *Request for Comments 2002*, October 1996.

[25] C. Perkins. Mobile IP. *IEEE Communications*, vol. 35, no. 5, pp. 84-99, May 1997.

[26] C. Perkins and D. Johnson. Mobility support in IPv6. In *Proceedings of the Second ACM International Conference on Mobile Computing and Networking (MOBICOM '96)*, pp. 27-37, November 1996.

[27] C. Perkins and D. Johnson. Route Optimization in Mobile IP. *Internet Draft*, draft-ietf-mobileip-optim-07.txt, November 1997.

[28] D. Plummer. An Ethernet Address Resolution Protocol –or– Converting Network Protocol Addresses to 48 bit Ethernet Address for Transmission on Ethernet Hardware. *Request for Comments 826*, November 1982.

[29] J. Romkey. A Nonstandard for Transmission of IP Datagrams over Serial Lines: SLIP. *Request for Comments 1055*, June 1988.

[30] W. Simpson. The Point-to-Point Protocol (PPP). *Request for Comments 1661*, July 1994.

[31] J. Veizades, E. Guttman, C. Perkins, and S. Kaplan. Service Location Protocol. *Request for Comments 2165*, June 1997.