Vertical Handoffs in Wireless Overlay Networks

Mark Stemm and Randy H. Katz
{stemm,randy}@CS.Berkeley.EDU
Computer Science Division, Department of Electrical Engineering and Computer Science,
University of California at Berkeley, Berkeley, CA 94720-1776.

Abstract

No single wireless network technology simultaneously provides a low latency, high bandwidth, wide area data service to a large number of mobile users. Wireless Overlay Networks — a hierarchical structure of room-size, building-size, and wide area data networks — solve the problem of providing network connectivity to a large number of mobile users in an efficient and scalable way. The specific topology of cells and the wide variety of network technologies that comprise wireless overlay networks present new problems that have not been encountered in previous cellular handoff systems. We have implemented a vertical handoff *system that allows users to roam between cells in wireless overlay networks. Our goal is to provide a user with the best possible connectivity for as long as possible with a minimum of disruption during handoff. Results of our initial implementation show that the handoff latency is bounded by the* discovery time*, the amount of time before the mobile host discovers that it has moved into or out of a new wireless overlay. This discovery time is measured in seconds: large enough to disrupt reliable transport protocols such as TCP and introduce significant disruptions in continuous multimedia transmission. To efficiently support applications that cannot tolerate these disruptions, we present enhancements to the basic scheme that significantly reduce the discovery time without assuming any knowledge about specific channel characteristics. For handoffs between room-size and building-size overlays, these enhancements lead to a best-case handoff latency of approximately 170ms with a 1.5% overhead in terms of network resources. For handoffs between building-size and wide-area data networks, the best-case handoff latency is approximately 800ms with a similarly low overhead.*

1. Introduction

Wireless networking is becoming an increasingly important and popular way of providing global information access to users on the move. Current technologies vary widely in terms of bandwidths, latencies, frequencies, and media access methods. Despite this heterogeneity, most existing wireless network technologies can be divided into two categories: those that provide a low-bandwidth service over a wide geographic area and those that provide a high bandwidth service over a narrow geographic area. While it would be desirable to provide a high-bandwidth service to mobile users at all times, this is unlikely. Wireless local area networks only provide limited coverage, and a mobile host equipped only with a wide-area network interface cannot exploit existing high-bandwidth infrastructure, such as in-building wireless local area networks or wired networks. No single wireless network technology simultaneously provides a low-latency, high-bandwidth, wide-area data service to a large number of mobile users.

Our solution is to use a combination of wireless networks to provide the best possible coverage over a range of geographic areas. A mobile device with multiple wireless network interfaces has many ways of accessing the wired infrastructure through alternative wireless subnets. For example, a typical user may move from her office, where her personal digital assistant (PDA) or laptop is connected via an in-room infrared network, to elsewhere in the building, where it is connected via a building-wide radio frequency (RF) network. The same user may then move outside, where her connectivity is via a wide-area data network, and then into another building which is connected via a different building-wide RF network. This combination of wireless network interfaces, spanning in-room, in-building, campus, metropolitan, and regional cell sizes, fits into a hierarchy of network interfaces which we call a *wireless overlay network* structure.

We have implemented a *vertical handoff* scheme that allows a mobile user to roam among multiple wireless networks in a manner that is completely transparent to applications and that disrupts connectivity as little as possible. For example, when the above user leaves her office, her PDA performs a vertical handoff from the in-room infrared (IR) network to the in-building RF network. Our system makes this completely transparent to applications running on the PDA. The only artifact of the handoff visible to an application is the quality of the device's connection to the wired infrastructure (increased/decreased bandwidth, increased/decreased

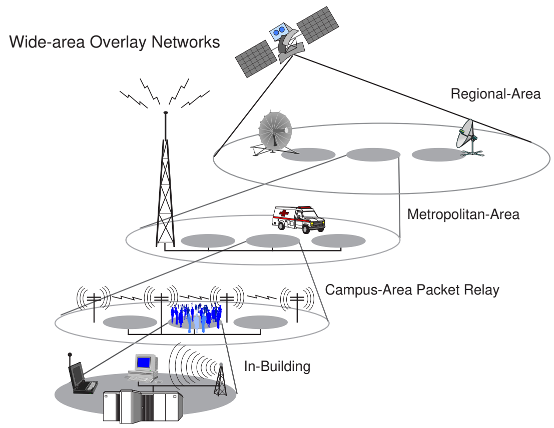

Figure 1. Wireless Overlay Network Structure

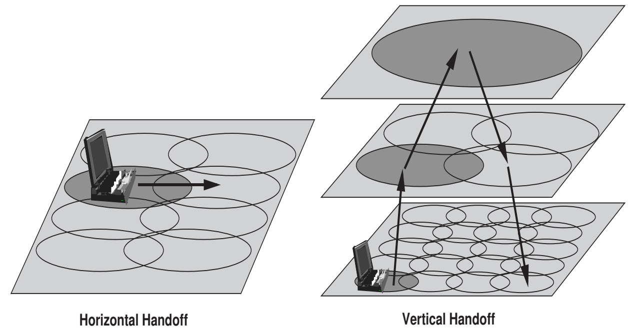

Figure 2. Horizontal vs. Vertical Handoffs

latency, higher/lower packet loss rate, etc.).

Our implementation delivers on the promise of seamless coverage: the typical handoff latency between networks is a few hundred milliseconds with minimal bandwidth and power overheads.

The rest of this paper is organized as follows: in Section 2, we describe in more detail the concept of wireless overlay networks and the technical challenges to be addressed in our handoff scheme. Section 3 describes our implementation of vertical handoffs. Section 4 presents the metrics used to quantify the performance and cost of our system. In Section 5 we present our experimental wireless testbed and performance results for the base handoff system, showing that handoff latency is dominated by the *discovery time*, the amount of time before a mobile discovers that it has moved into or out of a new wireless overlay. In Section 6, we present several enhancements that can be employed to decrease discovery time for applications that are sensitive to disruption. Section 7 discusses related work in low-latency handoff, overlay networks, and the use of multiple network interfaces. In Section 8, we conclude and Section 9 describes some ongoing and future projects in our system.

2. Wireless Overlays and Vertical Handoffs

In this section, we describe the wireless overlay network concept, why wireless overlay networks present new challenges compared to existing handoff systems, and the specific challenges to be met in our approach.

2.1 The Wireless Overlay Network Structure

Figure 1 shows an example of a wireless overlay network. Lower levels are comprised of high bandwidth wireless cells that cover a relatively small area. Higher levels in the hierarchy provide a lower bandwidth per unit area connection over a larger geographic area. In our system, we have three overlay levels. The lowest level comprises a collection of disjoint room-size high bandwidth networks, which provide the highest bandwidth per-unit-area: 1 Mbit/sec or more per room. The second level consists of building-size high bandwidth networks that provide approximately the same bandwidth as the room-size networks, but cover a larger area (for example, a single floor of a building). The final level is a wide-area data network, which provides a much lower bandwidth connection (tens of kilobits) over a much wider geographic area.

2.2 Horizontal versus Vertical Handoffs

We define a *horizontal handoff* as a handoff between base stations that are using the same type of wireless network interface. This is the traditional definition of handoff for homogeneous cellular systems such as cellular telephony systems, wide-area data systems, and wireless local area networks. We also define a new type of handoff, a *vertical handoff,* between base stations that are using different wireless network technologies. The terms horizontal and vertical follow from the overlay network structure that has networks with increasing cell sizes at higher levels in the hierarchy (Figure 2).

We divide vertical handoffs into two categories: an *upward vertical handoff* is a handoff to a wireless overlay with a larger cell size (and lower bandwidth per unit area), and a *downward vertical handoff* is a handoff to a wireless overlay with a smaller cell size (and higher bandwidth per unit area). A vertical handoff may be to an immediately higher or lower overlay, or the mobile host may "skip" an overlay. For example, a mobile may hand off from an in-room network directly to a wide-area network, or vice versa.

There are some important differences between the horizontal handoff problem and the vertical handoff problem that affect our strategy for implementing vertical handoffs:

- In horizontal handoff systems, a mobile host performs a handoff from cell A to cell B while moving out of the coverage area of cell A into the coverage area of cell B. In our system, this is not necessarily the case. For example, when a user performs an upward vertical handoff from an in-room cell A to an in-building cell B, the user is moving out of the coverage of cell A. However, when a user performs a downward vertical handoff from cell B to cell A, the user is not moving out of the coverage of cell B. This implies that downward vertical handoffs are less time-critical, because a mobile can always stay connected to a upper overlay while handing off to a lower overlay.

- Many network interfaces have an inherent diversity that arises because they operate at different frequencies. For example, a room-size overlay may use infrared frequencies, a building-size overlay network may use one set of radio frequencies, and a wide-area data system may use another set of radio frequencies. Another way in which diversity exists is in the spread spectrum techniques of different devices. Some devices use Direct Sequence Spread Spectrum, (DSSS), while others use Frequency Hopping Spread Spectrum (FHSS). The enhancements described in Section 6 that reduce the discovery time take advantage of this diversity between network interfaces
- In a network of homogeneous base stations, the choice of "best" base station is usually obvious: the mobile chooses the base station with the highest signal strength after incorporating some thresholding and hysteresis. In a multiple-overlay network, the choice of the "best" network cannot usually be determined by channel-specific factors such as signal strength because different overlay levels may have widely varying characteristics. For example, an in-building RF network with a low signal strength may yield better performance than a wide-area data network with a high signal strength. There are also considerations of monetary cost (some networks charge per minute or byte) that do not arise in a homogeneous handoff system.

2.3 Primary Objectives and Challenges

Unlike previous work that has studied aggregate metrics from a large population of mobile users [26] [9], our work focuses on the performance of an individual user roaming in a Wireless Overlay Network environment. In this work, the primary objective is to minimize the vertical handoff latency for an individual user while keeping bandwidth and power overheads as low as possible. These trade-offs are described in more detail below.

The primary technical objectives in the design of a seamless vertical handoff system are:

- *Low Latency Handoff: make the switch between networks as seamless as possible for disruption-intolerant applications and with as little data loss as possible.*

Our goal is to enable a typical user to use fully-interactive multimedia communication applications across many networks. As a user roams from areas of good connectivity to areas of poor connectivity, the only user-visible change should be due to the limitations of the specific wireless interfaces. For example, lower overlay levels may support full-motion video and high-quality audio, while higher overlay levels may support only audio. Our goal is to reduce handoff disruption as much as possible, reducing any user-visible changes to those inherent in the wireless technologies.

- *Power Savings: minimize the power drain due to multiple simultaneously active network interfaces.*

The simplest approach to managing multiple wireless network interfaces (NIs) is to keep all of them on all the time. However, measurements of commercially available wireless network interfaces [27] show that an IBM Infrared and WaveLAN [28] RF interface together consume approximately 1.5 watts *even when not sending or receiving packets.* This is approximately 20% of the total power drain of a typical laptop computer [11]. At these levels of power consumption, effective management of network interfaces is crucial.

- *Bandwidth Overhead: minimize the amount of additional network traffic used to implement handoffs.*

Implementing vertical handoffs in wireless overlays consumes bandwidth in the form of beacon packets and handoff messages that is necessary to provide service to roaming users, and we want to minimize these costs while also providing good performance.

There are many inherent trade-offs in meeting these objectives, and we must avoid situations that realize one goal at the expense of others. For example, reducing power consumption by keeping network interfaces off when not in use increases handoff latency. Similarly, zero-latency handoff could be achieved by simply sending and receiving data across all network interfaces simultaneously at all times, but this results in an inordinate waste of bandwidth and power. Our goal is to balance low latency handoffs with the unavoidable costs that arise from implementing them.

Challenges in realizing these objectives include:

- *Discovering the right time to perform handoffs in a wireless channel whose behavior can be difficult to predict and characterize.*

Ideally, a user should stay connected to the lowest overlay network (where the bandwidth per unit area is largest) for as long as possible until it is absolutely necessary to move to a higher overlay. While a user is roaming within an overlay, our system should behave exactly like a homogeneous cellular system. The primary trigger for a vertical handoff is that the currently active overlay network is no longer reachable because the mobile host has moved out of coverage of that overlay. For specific RF systems that transmit by modulating a fixed carrier frequency (e.g. GSM, DECT, AMPS), much work has been done in modeling channel quality and predicting bit error rate (BER) from channel-specific measurements [7] [3]. Although it may be possible to reapply these techniques to use channel characteristics to predict when a dis-

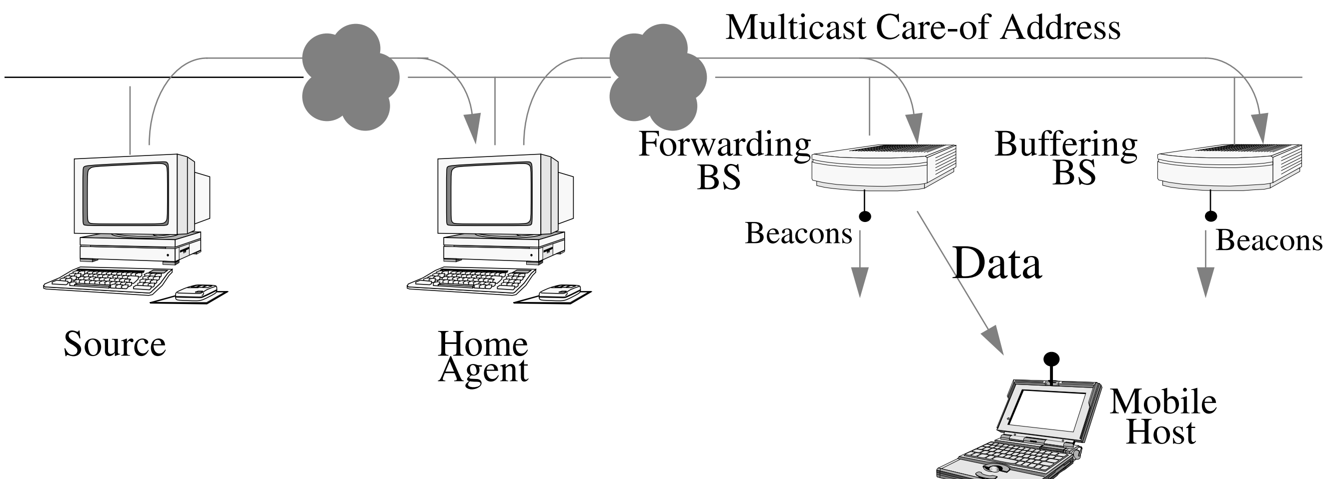

Figure 3. Overview of the Handoff System

connection is likely, we are interested in supporting a wide variety of wireless devices across different frequencies and physical layers. In addition, there is no requirement that all overlays must be wireless. An additional overlay could consist of a wired rather than wireless network, where channel measurements are meaningless and the transition from a connected state to a disconnected state is instantaneous. Rather than incorporate new network-specific policies for each new wireless or wired network interface, we chose to depend only on the presence or absence of packets to trigger a vertical handoff. Basing our scheme on the presence or absence of packets results in a system with an acceptable handoff latency while avoiding network-specific dependencies. This results in a more robust system with a dramatically reduced complexity where new network technologies can be added easily.

- *Interoperation with commercially available services and technologies that we cannot modify.*

We must depend on existing networking technologies and wireless data providers to have a full range of wireless networks at our disposal. Although we assume that we can modify some components of these systems (e.g. base station software), this may not be true for some overlay networks. In our system, for example, we can modify and experiment with the base stations for the room-size and building-size overlays, but the wide-area data overlay is owned and administered by a third party. As a result, we cannot directly control the overlay's infrastructure. This is an important consideration because it limits the modifications we can make to support vertical handoffs.

3. The Basic Handoff System

In this section we describe our wireless testbed and the basic system used to implement vertical handoffs.

Handoffs are built on top of the mobile routing capabilities of Mobile IP (Figure 3). The infrastructure we use is similar to the one described in [24] and the Mobile IP specification [22]. Mobile Hosts (MHs) connect to a wired infrastructure via Base Stations (BSs) which act as Foreign Agents (FAs). A Home Agent (HA) performs the same functions as in Mobile IP, encapsulating packets from the source and forwarding them to the FAs. One important difference is that the care-of address is a multicast rather than unicast address. A small group of BSs are selected by the mobile to listen to this multicast address for packets encapsulated and sent by the HA. One of the BSs is selected by the MH to be a *forwarding* BS; it decapsulates the packets sent by the HA and forwards those packets to the MH. The other BSs are *buffering* BSs; they hold a small number of packets from the HA in a circular buffer. When the mobile initiates a handoff, it instructs the old BS to move from forwarding to buffering mode, and the new BS to move from buffering to forwarding mode. The new BS forwards the buffered packets that the mobile has not yet received. For networks in which the BS infrastructure is not under our control, the Home Agent acts as the BS to the Mobile Host; the FA functionality with respect to that wireless network is incorporated at the HA machine instead of being incorporated at the gateway between the wired and wireless network.

BSs send out periodic beacons similar to Mobile IP foreign agent advertisements. The MH listens to these packets and determines which BS should forward packets for the mobile, which BSs should buffer packets in anticipation of a handoff, and which BSs should belong to the multicast group assigned for a single mobile.

Figure 4 shows a detailed breakdown of the state and agents that implement the handoff system. The network layer of the Home Agent includes a translation table that maps from a MH's home address to a multicast care-of address. All incoming packets are compared against the entries in the table. Matching packets are encapsulated and forwarded

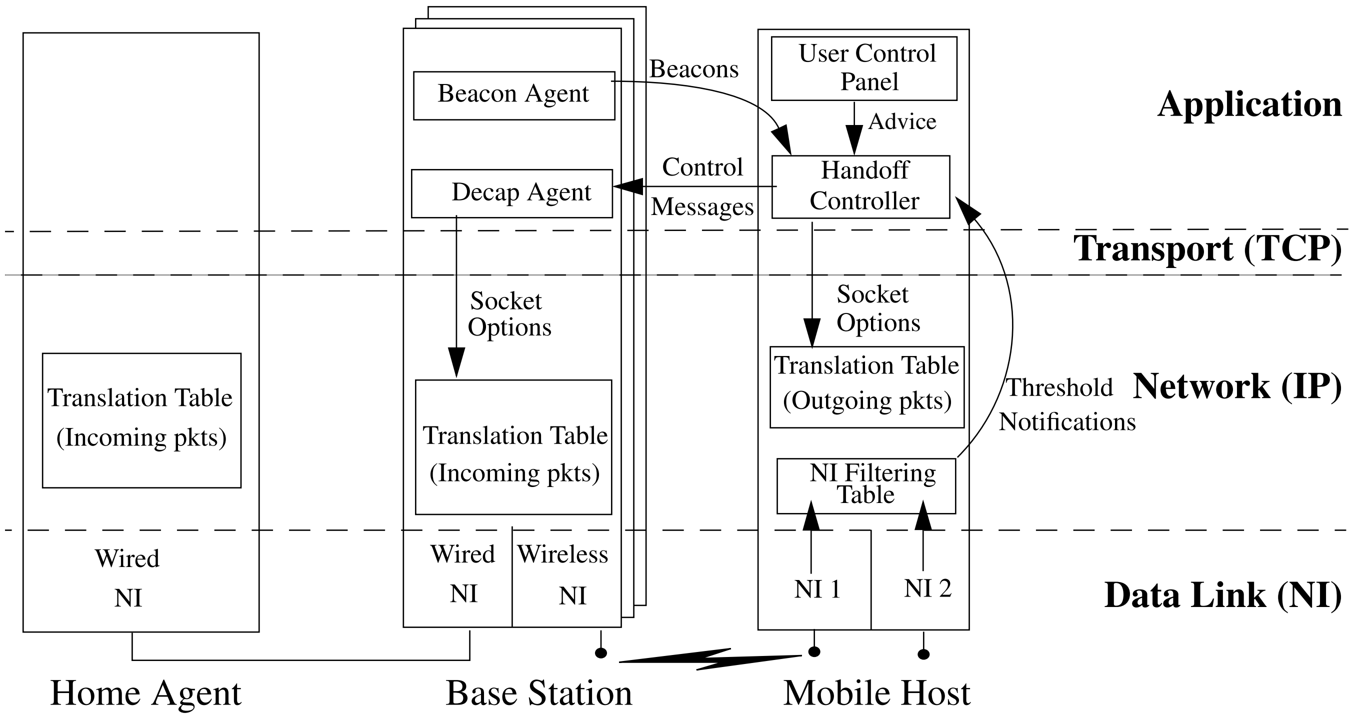

Figure 4. Network Stack for Home Agent, Base Station, and Mobile Host

using the corresponding multicast care-of address. At each BS there is a translation table that maps a MH's multicast care-of address to a local address. The translation table also includes the state of the BS with respect to this MH (e.g. buffering packets, forwarding packets, etc.). All incoming packets are compared against the entries in the table and the operation in the table (forward to mobile, buffer packet for mobile, etc.) is performed for matching packets. There are two user-level agents at the BS: a *beacon agent* that transmits beacon packets, and a *decapsulation agent* that receives control messages from the MH that modify the kernel-level translation table. The decapsulation agent manipulates the translation tables from user-level using socket options. At the mobile host, there is a single packet header translation table that inserts the MH's home address in all outgoing packets. There is also a network interface (NI)-specific table that keeps track of the number of packets that have arrived for the MH over each network interface and filters out duplicate packets that are received over multiple network interfaces. A user-level process can register a callback with the networking stack to be notified when changes occur in this table. When more than a threshold number of packets arrives over a single interface, the user-level process is notified. This table and the associated threshold notification callbacks are used in the doublecasting schemes described in Section 6.2. There are two user-level agents at the mobile host: a *handoff controller* that uses beacons to determine the overlay network and BS to connect to, and a *user control panel* that allows the user to control the choice of network or BS to use via *advice*, described in Section 3.2.

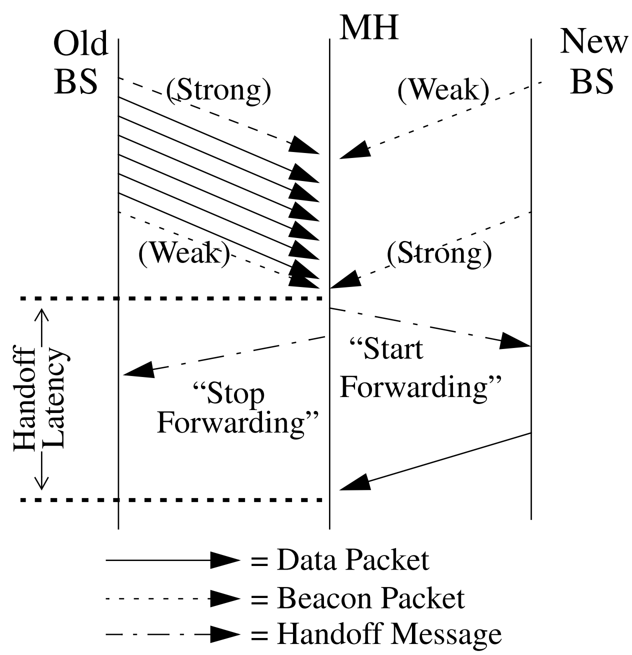

Figure 5. Breakdown of Horizontal Handoff

3.1 Triggering Handoffs

In a network of homogeneous BSs, the relative signal strength of beacons is compared and the BS with the highest is chosen as the forwarding BS. Figure 5 shows in detail the breakdown of a horizontal handoff. The three vertical lines represent the old BS, the MH, and the new BS, respectively, and the arrows represent messages sent from one machine to

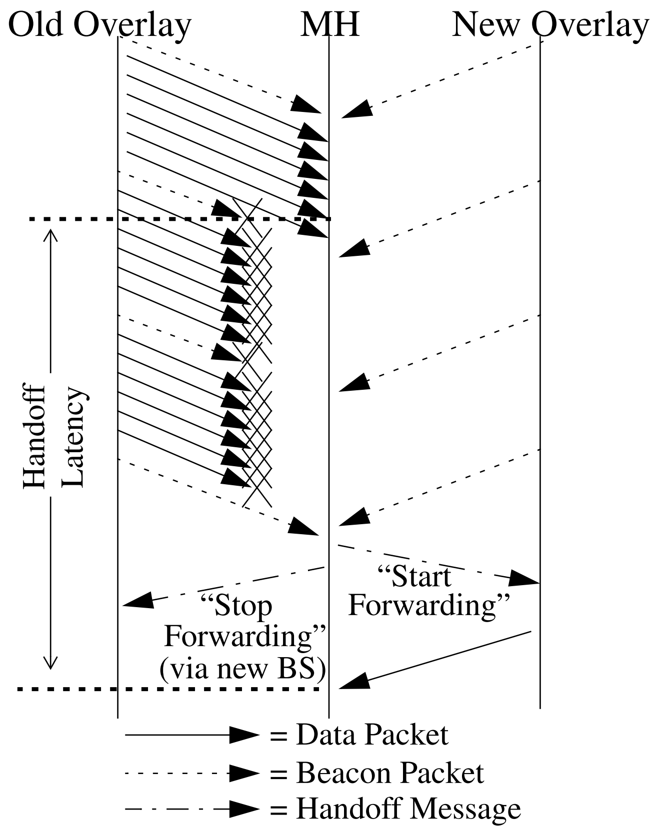

Figure 6. Breakdown of Basic Upward Vertical Handoff

another. The BSs transmit infrequent beacon packets to the broadcast address of the local subnet. Data packets are also forwarded from the old BS. At some point, the signal strength of the new BS is greater than that of the old BS, and the MH initiates a handoff to the new BS. It instructs the new BS to stop buffering packets and start forwarding packets to the MH. The MH also instructs the old BS to stop forwarding packets and start buffering packets. In the homogeneous handoff system, the handoff latency is measured from the time the mobile decides that the new BS has a larger signal strength until the first data packet arrives from it.

In our system, while a MH roams within the cells that comprise a single overlay, handoffs happen just as in the original system. The MH uses a channel-specific metric to compare different BSs and connects to the best one according to that metric. This allows the horizontal handoff system to operate seamlessly underneath the vertical handoff system. For an overlay network that handles mobility directly (for example, CDPD [12] or Metricom's Ricochet [21] network), our system does nothing and lets that network make all mobility decisions.

Figure 6 shows the breakdown of a typical vertical handoff. An upward handoff is initiated when several beacons from the current overlay network are not received. The MH decides that the current network is unreachable and hands over to the next higher network. Even though the MH cannot directly hear the old overlay network, it must still instruct the BS of the old overlay to stop forwarding packets. This request is routed through the new BS. The arrows represent the logical endpoints of a message, not the path that the message takes from source to destination. Downward vertical handoffs are initiated when several beacons are heard from a lower overlay's NI. The MH determines that the mobile is now within range of the lower overlay's NI and switches to the lower overlay. The handoff starts when the lower overlay becomes reachable or unreachable, and ends when the first data packet forwarded from the new overlay network arrives at the MH. As previously mentioned, our system only depends on the presence or absence of packets to make vertical handoff decisions.

3.2 Mechanisms for Customization

In our system, the handoff controller at the MH has primary responsibility for initiating handoffs. There may be situations, however, where the handoff controller cannot make the "best" decision about the choice of network or BS to connect to or the handoff enhancement to use (these enhancements are described in Section 6). To allow for more flexibility, the MH can take *advice* from an external source about the choice of network or BS which to connect as well as the handoff mechanism to use. Possible external sources include:

- A *user-visible control panel* that allows the user to specify specific constraints about which networks to use.
- A *subnet manager* that offers heuristic advice to avoid cell hotspots and increase the utilization of sparsely populated cells. For example, it may be advantageous to switch some users to a higher overlay network if the cell that they are currently using is congested or close to capacity. This increases the average effective bandwidth per user by eliminating bottleneck cells.

This advice could suggest a network or BS to switch to or whether to strive for low-latency handoffs. This mechanism allows for the implementation of policies for load balancing and user- or application-customization.

3.3 Power Management

As previously mentioned in Section 2.3, power management of multiple wireless devices is important. Table 1 shows the steady state power consumption of network interfaces when they are in an idle state. Our system handles power management by the turning off idle network interfaces when not in use. All network interfaces for overlays higher than the current network interface are kept off by default. They are turned on when geographic or other hints indicate that a handoff may be likely. By guessing that a handoff is likely, this reduces the probability that a sleeping network interface must be turned on before a handoff can complete. The NI for the overlay immediately below the current overlay is put into a power saving low duty cycle sleep state where it wakes up every few seconds and listens for beacons on the lower interface for a short time. This may increase the latency for downward vertical handoffs, as a mobile will take longer to discover that it has re-entered a new overlay. However, the mobile will not be disconnected during the discovery time and there will be no application-visible disruption.

4. Description of Metrics

In this section, we describe the parameters and metrics that we use to quantify the performance and overhead of our handoff system.

4.1 Parameters

We use the following variables:

S_H = the size of an IP Header + Link-layer header (in bits).

S_B = size of a beacon packet (in bits).

S_M = size of a mobile-initiated handoff message (in bits).

S_D = size of a user's data packet (in bits).

L_U = latency of the upper network interface (in seconds)

L_L = latency of the lower network interface (in seconds)

B_U = bandwidth of the upper network interface (in bits/sec).

B_L= bandwidth of the lower network interface (in bits/sec).

P_L = power consumption of the lower interface (in mW).

P_U = power consumption of the upper interface (in mW).

N_B= spacing between beacon packets (in seconds).

N_D = spacing between user data packets (in seconds).

T_B= threshold number of beacon packets heard or not heard before initiating a handoff.

T_D = threshold number of data packets heard or not heard on a new interface before initiating a handoff to that new interface.

D = length of power-saving duty cycle for NIs that are in sleep mode (in seconds).

Note that the actual values for each of these variables may differ from network to network. For example, the threshold number of beacons may differ for a WaveLan and Ricochet network. Also note that the packet size S_D may vary from application to application. For the applications we are most interested in, however (Section 6), we assume a fixed packet size.

4.2 Metrics

We define the vertical handoff latency L as the amount of time from when the mobile is disconnected from the old BS to when the mobile receives the first packet from the new BS. Note that this definition of latency implies a handoff due to mobility. If the handoff reason were due to other reasons (such as a user manually switching between interfaces), the definition of handoff latency would be the same as the horizontal handoff system. We break down the latency required to complete a vertical handoff into the following components:

$$L = L_D + L_P + L_N + L_F$$

- L_D is the component of latency during which the mobile discovers that it must hand off to a new wireless overlay. This could be to an upper overlay as a result of moving out of range of the current overlay: for example. moving out of a room or moving out of a building. This could also be to a lower overlay as a result of moving back into coverage of a lower overlay: for example, moving back into a room or building. In the basic system, this is largely a function of the beaconing frequency. A smaller beacon frequency increases L_D, and a larger beacon frequency decreases L_D. As previously noted, for most horizontal and downward vertical handoffs this component of latency is not visible to the user as a disconnection, because the mobile is still connected to the old BS while it discovers that it can hear the new BS.
- L_P is the latency for the mobile to power on the upper or lower network interface, including any network registration time. This component of latency may or may not be visible to the user depending on whether the device was already on at the time the handoff occurred. Ideally, with the mechanisms described in Section 6.1, we can predict when the user is likely to hand off and can hide this latency from the user.
- L_N is the latency for the mobile to inform the new BS to start forwarding data to the mobile. This is usually a function of the network latency between the MH and BS.
- L_F is the latency for the BS to send the first data packet across the new network to the mobile. If there is no outstanding data to send to the MH, then this component of the latency is zero. For the measurements in Section 5 and Section 6, we made sure that there was outstanding data to forward. This component of the latency is a function of the latency and bandwidth between the MH and the BS.

Some of the components of latency may sometimes overlap, while others can not overlap. L_D and L_P can overlap if a mobile "guesses" that an overlay will soon become unreachable and powers on a network interface in advance. Similarly, L_D, L_N and L_F can overlap if a BS is already forwarding packets to a mobile when a handoff occurs (the Packet Doublecasting scheme described in Section 6.3.2). L_P

Type of medium	User-visible Bandwidth	Cell Diameter	Latency	Registration Time (95% Conf Interval)	Power Consumption (mW)
Infrared (IBM Infrared)	800 kb/sec	7 meters	2-5ms	6.7 ms (5.7-7.8 ms)	349.6
In-Building RF (915 Mhz/2.4Ghz WaveLAN)	1.6 Mb/sec	100 meters	2-5 ms	110.4 ms (93.8-127.0 ms)	1148.6 (915) 1318.8 (2.4)
Wide-Area Data (Ricochet)	60 kb/sec	1 km	≈100 ms	7.6 sec (6.3-8.9 sec)	346.9

TABLE 1. Bandwidths, Latencies, and Registration Times for Our Networks

and L_N usually do not overlap, however; a NI must be powered on and registered before it can accept and deliver packets. In addition, some of these components of the latency may be large and not under our control. For example, many wide-area wireless networks such as Ricochet and CDPD have a network registration process that must occur before a device can be connected, increasing the value of L_P. Wide-area wireless networks often have a much larger latency than local-area wireless networks, which increases L_F.

We define the power overhead P as the amount of power from network interfaces that must be consumed by a MH while making handoff decisions. This is a function of the number and type of wireless interfaces that are powered on.

We define the bandwidth overhead B as the number of bits sent per second by the BS that are not actual data packets, such as beacon packets or other control messages that the mobile uses to initiate a handoff.

5. Results for the Base System

In the following sections, we focus on the handoff between two overlays: upward vertical handoffs from a lower overlay to an upper overlay, and downward vertical handoffs from an upper overlay to a lower overlay.

5.1 Measurement Testbed

Our testbed consists of IBM ThinkPads, Gateway 2000 Solo laptops, and Intel-based PCs running a modified version of BSD/OS 2.1 and BSD/OS 3.0. Table 1 shows the specific wireless networks that we use along with typical bandwidths, latencies, and registration times. We use the IBM Infrared Wireless LAN [14] network as our room-size network, the AT&T WaveLAN [28] as our building-size network, and the Metricom Ricochet Network [21] as our wide-area data network. The registration time includes the time to power on the network interface as well as register with the wired infrastructure. The Ricochet network is the only network that must register with a wired infrastructure. The registration times were measured by sending a stream of UDP packets to a mobile host, turning on the network interface, and marking the time between when the network interface was turned on and when the first data packet was received by the mobile.

5.2 Measurement Methodology

We measured the latency of handoffs by sending a continuous stream of 1024 byte UDP packets to the MH. For the Infrared to WaveLAN transitions, this was limited to 500 kilobits/sec. For the WaveLAN to Ricochet transitions, this was limited to 50 kilobits/sec. The handoff was initiated by forcing the MH to turn the lower interface off and on in response to external messages. An observer machine was running tcpdump [19] and the resulting packet trace was post-processed to determine when the external messages triggered the turn-on and turn-off of the interface, when the MH sent the control messages to the BSs, and when the first packets arrived over the new interface to the MH.

5.3 Predicted Performance

Table 2 shows algebraic derivations for L_D, L_N, L_F, P, and B as a function of the variables in Section 4. For upward handoffs in the basic system, the MH must wait for approximately T_B beacons to determine that the current overlay is no longer reachable. The $N_B/2$ term accounts for the fact that a mobile may move out of the coverage of an overlay anywhere between two beacon times: on average this happens midway between two beacons. For downward handoffs, an additional $D/2$ seconds must be spent waiting for the lower interface to come out of its power saving state and hear the lower overlay's beacons. The mobile must then notify the upper BS to start forwarding new packets: this takes $L_U + S_M/B_U$ seconds for the upward handoffs and $L_D + S_M/B_D$ seconds for downward handoffs. Finally, the new BS must forward the first data packet to the mobile: this takes L_U+S_D/B_U seconds for upward handoffs and L_U+S_D/B_U seconds for downward handoffs. The steady state power consumption of this scheme is only P_L or P_U mW, because

Type of Handoff	L_D	L_N	L_F	P	B
Basic Upward	$N_B(T_B\text{-}1)+N_B/2$	L_U+S_M/B_U	L_U+S_D/B_U	P_L	$(1/N_B)S_B$
Basic Downward	$D/2+ N_B(T_B)+N_B/2$	L_L+S_M/B_D	L_L+S_D/B_L	P_U	$(1/N_B)S_B$

TABLE 2. Predicted latency and cost for the basic system

Transition	L_D+L_N (sec)	95% Conf (L_D+L_N)	L_F (ms)	95% Conf (L_F)	Total (sec)	B (bits/sec)
Infrared→ WaveLAN (Measured/Predicted)	2.5 2.50385	1.85 -3.25	8.35 9.0	7.19-9.51	2.508 2.51285	512
WaveLAN→ Infrared (Measured/Predicted)	3.5 3.50422	N/A	20.34 14.5	4.634 -36.05	3.520 3.51872	512
WaveLAN→Ricochet (Measured/Predicted)	2.79 2.6144	2.7-2.99	295.96 320	221.71-370.21	3.086 2.9344	512
Ricochet→WaveLAN (Measured/Predicted)	3.8 3.50386	N/A	8.72 9.0	7.47-9.97	3.80872 3.51286	512

TABLE 3. Breakdown of handoff latency for the basic system

only a single interface needs to be on to trigger a handoff. The steady-state bandwidth overhead is from the beacon messages: this consumes $(1/N_B)*S_B$ bits per second.

5.4 Measured Performance

Table 3 shows the measured and predicted results for the basic system. L_P is not included; we assume that the interface is already turned on. We use a N_B of 1 second, and a T_B of 3; when more than three beacon times pass without hearing any beacons, the MH considers the current overlay unreachable and switches to the next higher overlay. Similarly, when the MH hears three beacons from a previously disconnected overlay, the MH switches back to the old overlay. The choice of three is a heuristic; for heavily loaded networks, beacons may be delayed or lost and a small value of N_B may cause unnecessary handoffs. A value of three for T_B incorporated enough hysteresis to account for lost beacons and eliminate unnecessary handoffs. The predicted values of L_D, L_N, and L_F agree well with the measured values. L_N and L_F take approximately as long of that in the horizontal handoff system [24]. From Figure 7, we see that the handoff latency is dominated by L_D+L_N, from 2.5 to 3.8 seconds. Even for the wide-area data network (Ricochet), which has very different network characteristics, the handoff time is dominated by L_D+L_N. Because we use an observer machine to record when events occur and only the MH decides when an overlay has become (un)reachable, it is impossible to separately measure the L_D and L_N components of latency. To measure them separately would take perfect clock synchronization at the observer machine and MH. We therefore made separate measurements of L_N from the mobile host (not included in this paper) to verify that the actual value of L_N was close to the predicted value. Because L_N is only a function of the NI's latency and bandwidth, the handoff time is dominated by L_D. This illustrates one of the difficulties that arise from using heterogeneous network interfaces. In a system with homogeneous BSs, it is easy to make comparisons about the quality of the connection to each BS by using channel specific measurements. Because we cannot make direct comparisons between BSs using channel-specific measurements, we must wait for an overlay to become reachable or unreachable before determining that a handoff to a higher or lower overlay must occur.

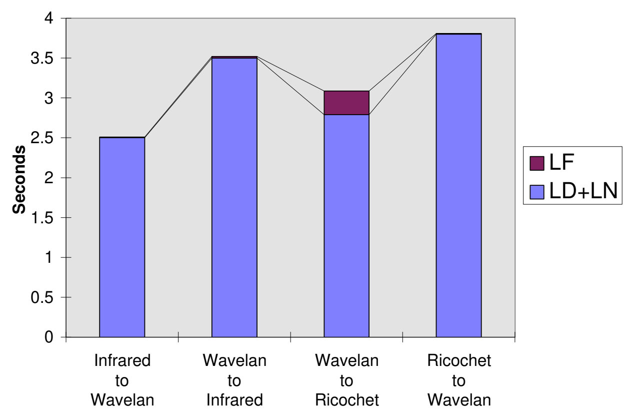

Figure 7. Breakdown of Handoff Latency in Basic System

The main conclusion from the basic system is that vertical handoffs are dominated by the time before the MH discovers that it has moved into/out of coverage of an overlay (L_D), and that any enhancements to the basic system should concentrate on reducing this component of the latency.

6. Enhancements

One of the goals in our handoff system is to support interactive multimedia communication across multiple network interfaces, and for these applications, a latency of approximately three seconds is unacceptable. Even for non-real time applications such as non-interactive file transfers and WWW browsing, a latency of several seconds will lead to a loss of multiple data segments. Previous work has also shown that packet losses during handoff has detrimental effects on reliable transport protocols such as TCP [8]. With this in mind, we examined several enhancements to the base strategy that allow us to reduce the value of L_D during handoff.

6.1 Hints for Enabling Enhancements

The schemes described in this section are used in situations where the application indicates that a low handoff latency (less than 300-500ms) is important, such as real time interactive voice or video. Even when an application indicates that low-latency handoff is important, these enhancements are not used continuously, because of the bandwidth/power overheads. They are used only when the mobile is in a situation where it may hand off soon. Note that this is not the same as determining that a mobile must hand off immediately (i.e., the mobile is now disconnected). Alternative hints can be used to predict that a handoff is likely. These include

- *User input*: The user can instruct the MH to be more aggressive about handoff by using these enhancements. When the user is likely to leave the building, she can put the MH in a mode that uses these enhancements. The user can take the MH out of this state when not moving.
- *Received signal strength*: Although signal strength indicators, when present, may not be a good indicator of imminent handoff, they do well at indicating the distance between a MH and BS. When a MH notices that the signal strength is gradually decreasing it can assume that the user is moving away from a BS, and when the signal strength is increasing a MH can assume that the user is moving toward a BS. When the best BS that a MH can hear has a low signal strength that has been decreasing, a MH can assume that a vertical handoff may be needed and start using some of these enhancements.
- *Geographic hints*: We can use traces to predict which cells are the gateways to a new overlay network. Although the overlapping nature of wireless overlays means that a user can be potentially connected to multiple networks at once, the transitions between networks are a function of the building geography. A vertical handoff is only possible from certain places in the building, and only certain cells cover these locations (e.g. only one in-building RF cell is likely to cover the exit of an office building). The BSs covering these cells could add information in their beacon packets indicating that this cell is near the exit to a building, and that a vertical handoff to a wide-area network is likely.

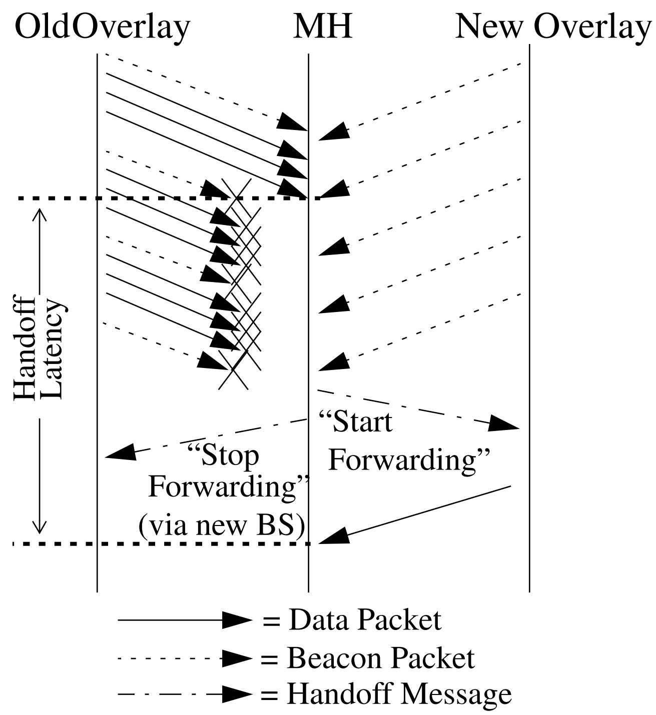

Figure 8. Breakdown of Fast Beaconing Handoff

- *Handoff Frequency*: The MH can also track the frequency of handoffs and use these enhancements when more handoffs are occurring, indicating that movement out of this overlay's coverage is more likely. This approach has been suggested for switching between high-tier and low-tier PCS systems [23].
- *Missing a single beacon*: We mentioned in Section 4.1 that the MH waits for multiple beacon packets before determining that an overlay is (un)reachable and switching to a new overlay. The MH could turn on some of these optimizations after missing a single beacon packet, as an attempt to verify that an overlay is (un)reachable.

6.2 Enhancements

We can make the following enhancements to reduce handoff latency. All of these enhancements have some additional cost in terms of power or overhead bandwidth.

- **Fast Beaconing**: The MH can selectively instruct a subset of the BSs that are listening to its multicast group to transmit beacon packets at a higher frequency than once

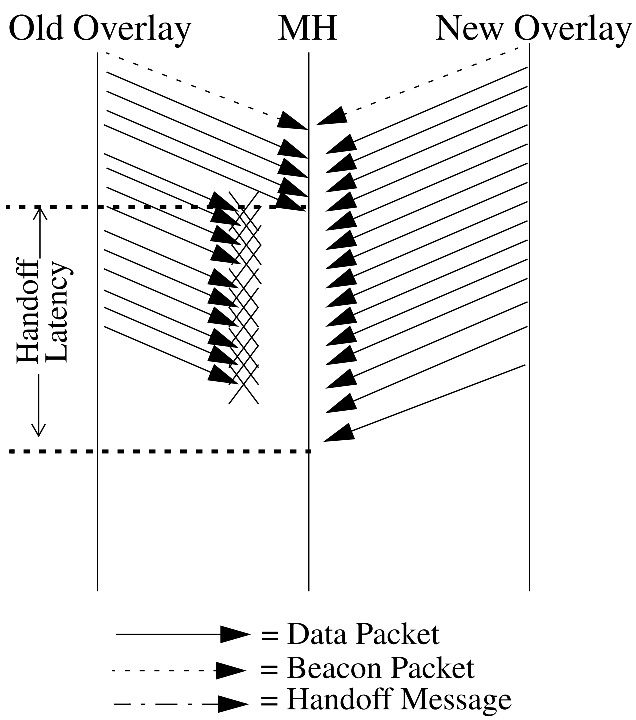

Figure 9. Breakdown of Packet Doublecasting Handoff

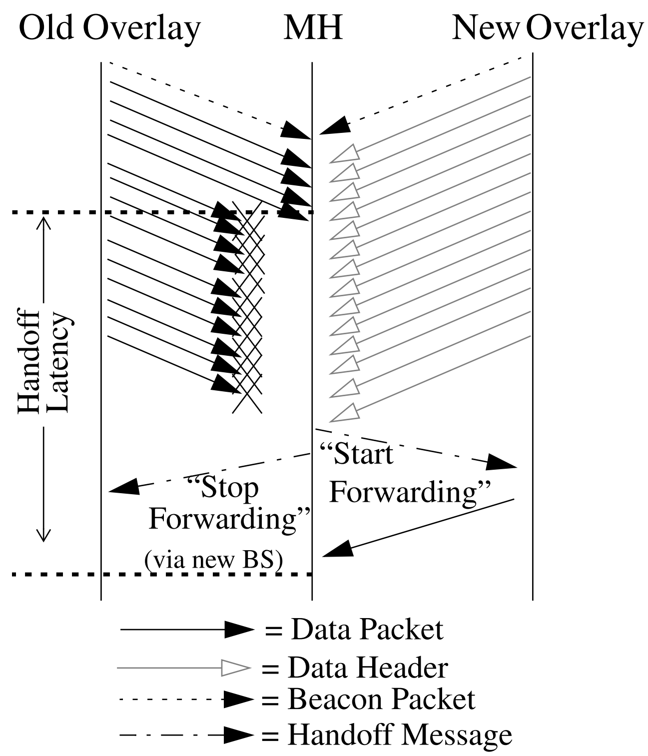

Figure 10. Breakdown of Header Doublecasting

per second. The MH still waits for T_B beacons to be lost before initiating a handoff, but the beacons are transmitted more quickly and L_D is reduced. The breakdown of a handoff is described in Figure 8. The handoff proceeds exactly as in Figure 6 — the beacon packets are simply received more quickly.

- **Packet Doublecasting(Figure 9)**: The MH can place into forwarding mode a subset of the BSs that are listening to the multicast group for the MH. This means that multiple copies of the packet will be transmitted from multiple BSs to the MH. In our scheme, two BSs are placed in forwarding mode simultaneously; the current BS and a BS of the next higher overlay. Duplicate packets are filtered out at the network layer at the MH by keeping a small cache of received IP packets and filtering out received packets whose IP ids are already in the cache. Although not strictly needed, this prevents unnecessary congestion control mechanisms from being invoked at the transport layer. The network layer at the MH also keeps track of packets that have been received by each interface. When more than T_D consecutive packets are received on a new interface with none received on the old interface, the MH decides that the old overlay is unreachable and initiates a handoff to the new interface. A breakdown of the handoff is shown in Figure 9. Two copies of each packet are sent to the MH, one from each BS. After T_D=10 packets are missed from the old overlay, the mobile switches to the new overlay. The packets kept in the network-level cache on the MH are forwarded to higher layers. In cases where no data is currently being sent to the MH, beacons are used to trigger a handoff. By utilizing diversity that arises from multiple network interfaces, this approach does at the network layer what the IS-95 CDMA Cellular phone standard [18] and the ARDIS wide-area data system [2] do at the physical layer. In IS-95, multiple BSs send duplicate copies of the same data using the same CDMA codes. The MH's receiver is already equipped to handle multiple time-shifted copies of the same waveform, and a MH moves into the cell of the new BS seamlessly. In ARDIS, multiple BSs transmit the same data at the physical layer to achieve better in-building penetration.

- **Header Doublecasting (Figure 10)**: This approach takes advantage of the fact that in the Packet Doublecasting approach, duplicate packets on the upper interface are used only as an indicator of handoff. Therefore, full packets do not have to be sent until the actual handoff occurs. In this approach the MH places a BS into a mode where it continues to buffer packets destined for the mobile host. However, the BS also forwards a packet containing the IP header of the buffered packet to the MH. The network layer at the MH keeps track of which packets or packet headers has been received by the mobile. The MH switches to the new BS when more than T_D headers have been received via new BS while no packets have been received via the old BS. The new BS forwards the packets just as in the Basic System. This approach has an advantage over Packet Doublecasting in that less data is sent on the upper overlay.

Type	L_D (up/down)	L_N(up/down)	L_F(up/down)	P	B
Fast Beacons	$N_B(T_B\text{-}1)+N_B/2$ $D/2+ N_B(T_B)+N_B/2$	L_U+S_M/B_U L_L+S_M/B_L	L_U+S_D/B_U L_L+S_D/B_L	P_L	$(1/N_B)S_B$
Packet Doublecasting	$N_D(T_D\text{-}1)+N_D/2$	0 0	0 0	P_L+P_U	$(1/N_D)S_D$
Header Doublecasting	$N_D(T_D\text{-}1)+N_D/2$	L_U+S_M/B_U L_L+S_M/B_L	L_U+S_D/B_U L_L+S_D/B_L	P_L+P_U	$(1/N_D)S_H$

TABLE 4. Algebraic expressions for L and B for the Enhancements

Both doublecasting approaches have an advantage over the beaconing systems in that they use extra resources only when the MH is actively receiving data. When the user is not receiving data, no extra bandwidth is used. Additionally, beacons sent from the base station affect all mobile devices in the wireless cell, and if beacons are sent at very high frequencies, media access affects (such as exponential backoff during link activity) may dramatically reduce the effective bandwidth of mobile hosts in the same cell. Another advantage of the packet doublecasting approaches is that the packets that trigger a handoff are not redundant; they are consumed by actual applications. If fast beaconing were used, then beacons (useless application-level data) would be competing with application-level data for network resources at all times.

A disadvantage of the doublecasting approaches is that both overlays must be able to support the same network load. Packet doublecasting across a high-bandwidth and low-bandwidth network will not work. Another advantage of the beaconing systems is that multiple users in a cell can use the same beacon packets (rather than separate data packets) to make handoff decisions.

6.3 Performance

Table 4 shows the algebraic expressions of L_D, L_N, L_F, P and B as a function of the variables described in Section 4. The formulas are very similar to those in Table 2. The fast beaconing system is identical to the basic system. In the header and packet doublecast systems, the MH must wait for T_D data packets (or T_B beacons, if no data is currently being sent) to arrive over the upper interface before the handoff is triggered. It takes $N_D(T_D\text{-}1)+N_D/2$ seconds for the mobile to determine that the packets have not arrived. In the packet doublecast system, the notification and forwarding latencies are effectively zero: the mobile only has to change the NI-specific filtering table in the kernel and forward the packets buffered at the network layer of the MH to higher layers. The header doublecasting scheme has the same notification and forwarding latencies as the beaconing system. The power consumption P of the doublecast schemes is more than that of the beaconing schemes, as both wireless interfaces must be on for the MH to make the handoff decision. The bandwidth overhead of the packet doublecast scheme is equal to the data rate at which the MH is receiving data ($(1/N_D)*S_D$). In the header doublecast scheme, the bandwidth is proportional to the data rate at which the MH is receiving data, but only a small header of size S_H is sent on the upper overlay for every data packet of size S_D sent on the lower overlay.

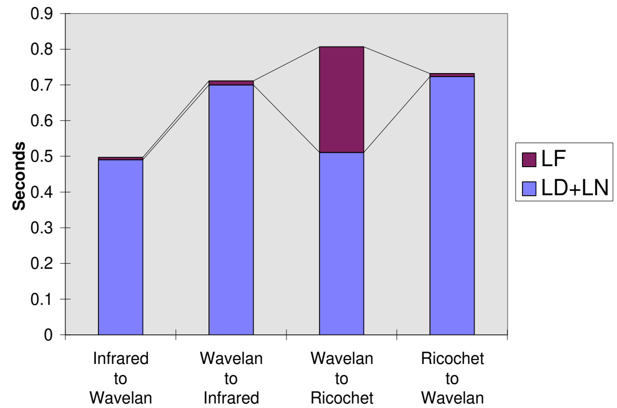

Figure 11. Breakdown of Handoff Latency for the Fast Beaconing System

6.3.1 Fast Beaconing

Table 5 shows the handoff latency and bandwidth overheads for the fast beaconing system, and Figure 11 shows graphically the breakdown of handoff latency for the fast beaconing system. Beacons were sent out every 200 ms instead of every second. As in the basic system, the beacon threshold was set to 3 beacons. The measured values of L_D, L_F and B agree well with the algebraic results. In all cases, the predicted latency is within or very close to the confidence interval for the measured latency. The latency has dropped by a factor of approximately 5 when compared to the basic system with a factor of 5 increase in bandwidth overhead. Because the algebraic values agree with the measured values, we would expect that faster beaconing would lead to further reductions in L_D with increases in B.

Transition	L_D+L_N (ms)	95% Conf L_D+L_N	L_F (ms)	95% conf L_F	Total L (ms)	B (bits/sec)
Infrared→WaveLAN (actual/predicted)	490 503.86	256-723	7.02 9.0	3.6-10.4	497.02 512.86	2480
WaveLAN→Infrared (actual/predicted)	700 704.22	N/A	11.1 14.5	5.87-16.3	711.1 718.72	2480
WaveLAN→Ricochet (actual/predicted)	511 614.4	457-607	Same as Basic	Same as Basic	806.96 934.4	2480
Ricochet→WaveLAN (actual/predicted)	723 703.86	N/A	Same as Basic	Same as Basic	731.72 712.86	2480

TABLE 5. Actual values of L and B for the Fast Beaconing Enhancement

Transition	L_D+L_N (ms)	95% Conf L_D+L_N	L_F	95% Conf L_F	Total L (ms)	B (bits/sec)
Infrared→ WaveLAN (actual/predicted)	202.4 165.892	131.3-243.46	0 0	0	202.4 165.892	520000
WaveLAN→ Infrared (actual/predicted)	200 183.308	N/A	0 0	0	200 183.308	520000
WaveLAN→ Ricochet (actual/predicted)	1599.7 1734.72	1470.4-1729.09	0 0	0	1599.7 1734.72	50000
Ricochet → WaveLAN (actual/predicted)	2396.5 1774.74	2186.1-2606.8	0 0	0	2396.5 1774.74	50000

TABLE 6. Actual values of L, B, and P for the Packet Doublecasting Enhancement

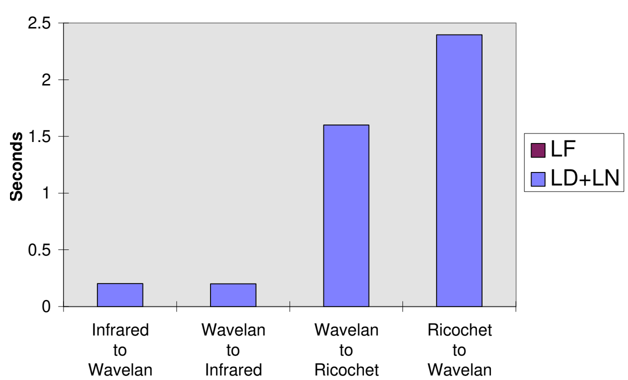

Figure 12. Breakdown of Handoff Latency for the Packet Doublecasting System

6.3.2 Packet Doublecasting

Table 6 shows the handoff latency and bandwidth overheads for the packet doublecasting enhancement, and Figure 12 shows graphically the breakdown of handoff latency for the packet doublecasting enhancement. We used a packet threshold of 10 packets — 10 packets must be received by the mobile over one interface before the MH decides to switch to a new overlay. The choice of 10 is a heuristic: ideally, when packets are being sent over multiple interfaces there is a perfect interleaving of packets from the lower and upper interfaces. In practice, due to the way in which our network interface drivers process packets, the interleaving is rather coarse-grained: a burst of packets arrives over 1 interface, followed by a burst over the other interface. The value of 10 was chosen to be larger than the largest burst of packets that we observed on heavily loaded networks.

In all cases other than L_D+L_N for the Ricochet → WaveLAN transition, the predicted latency is within or very close to the confidence interval for the measured latency. For the Infrared to WaveLAN handoffs, this approach achieves a lower handoff latency than the basic system (approximately 200ms), but at a considerable cost, as full packets must be sent over both network interfaces. For the WaveLAN network, this overhead of 520 kbits/sec is approximately one-third of the network's maximum user-visible bandwidth of

Transition	L_D+L_N (ms)	95% Conf L_D+L_N	L_F (ms)	95% Conf L_F	Total (ms)	B (bits/sec)
Infrared→ WaveLAN	170.8	133.75-208.01	10.9	10.2-11.7	181	16600
WaveLAN→ Infrared	170ms	N/A	11.7	9.1-24.3	181.7	16600
WaveLAN → Ricochet	Same as Packet	Same as Packet	Same as Basic	Same as Basic	1725.69	1660
Ricochet → WaveLAN	Same as Packet	Same as Packet	Same as Basic	Same as Basic	2530.47	1660

TABLE 7. Actual values of L and B for the Header Doublecast Enhancement

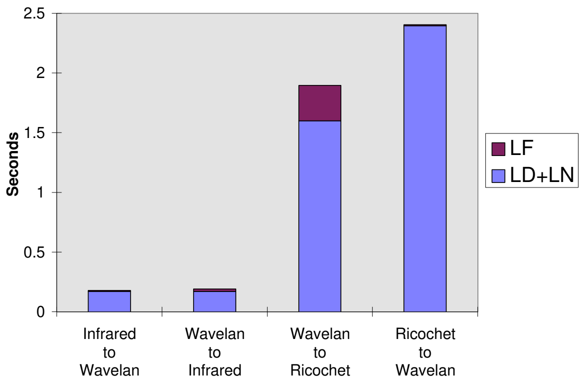

Figure 13. Breakdown of Handoff Latency for the Header Doublecasting System

1.6 Mbits/sec. For the WaveLAN to Ricochet handoffs, the latency is much larger than the approach that uses fast beaconing. The reason for this comes from the way in which vertical handoffs are initiated. In the fast beaconing system, the networks are considered independently: the presence or absence of beacons indicates whether or not to hand off. In the multicast approaches, this independence is lost because the networks are being compared in a relative manner. Packets arrive over multiple network interfaces and must be considered together before a handoff decision can be made. For the Ricochet network, the available bandwidth is sufficiently low that the amount of time it takes for the threshold number of packets to arrive is greater than the time it takes to independently consider the WaveLAN.

6.3.3 Header Doublecasting

Table 7 shows the handoff latency and bandwidth overhead for the header doublecast enhancement, and Figure 13 shows graphically the breakdown of handoff latency for the header doublecasting enhancement. As in the packet doublecasting scheme, we used a header threshold of 10 packet headers. The predicted measurements the same as in the packet doublecasting and beaconing systems.

The table shows that for WaveLAN to Infrared handoffs, the header doublecasting scheme achieves a slightly lower latency than the packet doublecasting scheme (171ms vs. 200ms) with a dramatic decrease in bandwidth resources on the upper network. For the WaveLAN network, this overhead is approximately 1% of the user-visible bandwidth of 1.6 Mb/sec. For the WaveLAN to Ricochet handoff, the bandwidth overhead is dramatically decreased, but the value of L_D has not dropped equally. The reason for this is that the Ricochet network is mainly latency bound: it can transmit approximately the same number of packets per second regardless of their size. We believe that this is because the Ricochet system is a multi-hop packet radio system with hop-by-hop acknowledgments, and the channel turn-around time while sending these acknowledgments decreases the packet throughput. Also, since packets must be forwarded from the Home Agent, the value of L_F has now increased. This implies that the header and packet doublecasting approaches hold little advantage over the Beaconing approaches when used on low-bandwidth, high latency networks such as wide-area data networks.

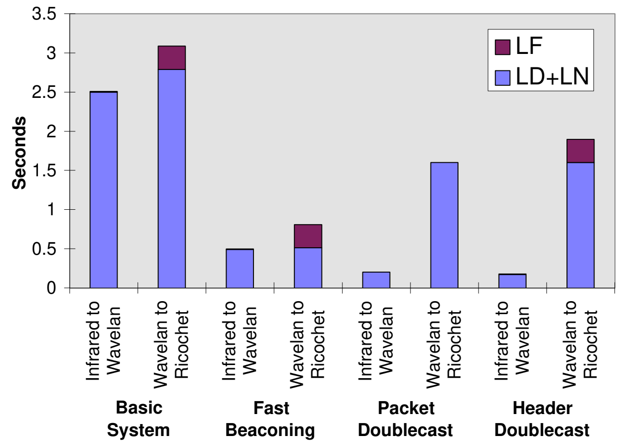

Figure 14. Comparison of Handoff Latency for Basic System and Enhancements

Figure 14 summarizes the performance of the basic handoff system and each of the enhancements for each of the upward vertical handoffs. We have learned the following things about the enhancements proposed to reduce handoff latency:

- Fast beaconing results in a decrease in latency proportional to the bandwidth overhead. This approach consumes bandwidth whether or not data is being sent to the mobile device.

- Packet doublecasting results in a loss-free zero latency handoff, but at a prohibitive cost.
- Header doublecasting results in a latency similar to the packet doublecasting scheme, but with a dramatic decrease in overhead.
- For handoffs between in-building and wide-area overlays, doublecasting approaches have limited effect due to the latency-bound nature of the wide-area network we used.

For the network interfaces in our overlay network structure, header doublecasting performs the best for transitions between in-room and in-building networks, and beaconing works best for transitions between in-building and wide-area networks.

7. Related Work

Related work in this area focuses on three areas: overlay networks in a cellular telephony rather than data-oriented context, improving handoff performance in a homogeneous environment, and the management of multiple network interfaces.

Wireless data overlays have been described in many places. The term "Wireless Overlay Networks" was first introduced in [17]. CDPD [12] can be described as a data overlay on top of the cellular phone system. Data overlays have also been studied in the context of cellular telephony. Other work ([26], [9]) focuses on a cellular system with large macrocells overlaid on a traditional microcellular system. These papers focus on large-scale metrics from a large number of mobile users such as call blocking and dropping probabilities, channel utilization, and spectral efficiency, without describing how handoffs would actually be implemented. Our work differs from theirs in that it shows how to implement a handoff system in the presence of heterogeneous network technologies, focusing on the handoff latency and overhead of a single mobile as it roams in its environment. However, these two approaches are complementary rather than competing. A large-scale view indicates the scaling properties of an overlay network structure for a large number of users, while our work focuses on the ability of the system to provide interruption-free service to individual users with a minimum of overhead.

The concept of overlay networks was also introduced in the context of high-tier and low-tier PCS systems [10]. Our work differs in the way mobile users are assigned to wireless cells. In microcell-macrocell systems, it is assumed that low-speed mobile users would be assigned to microcells while higher-speed mobile users would be assigned to macrocells. Our work takes a more generalized approach and focuses on providing the best possible connectivity to mobile users without depending on knowing *a priori* the speed of the user. Another significant difference between our work and other microcell-macrocell work is that most previous work assumes that all areas are covered by microcells as well as macrocells. We do not make this assumption, and assume that there are some areas of coverage that are only covered by macrocells. This distinction is important; because we assume regions with only macrocell coverage, we are forced to handle cases where a microcell becomes unavailable and even low-speed mobile users must perform vertical handoffs to higher overlays.

There have been numerous papers dealing with handoff across homogeneous cellular [23], ATM [1], and picocellular [13] networks and mobility in IP networks [15] [22] [16]. Seshan et al. [25] [24] [5] implemented a system for low-latency horizontal handoffs. Our work expands upon theirs in that it handles multiple wireless networks and cases where the mobile device cannot use channel characteristics to trigger handoffs.

Recent work has also addressed the problem of integration of multiple network interfaces in a single mobile. The MosquitoNet project at Stanford [4] has mobile devices equipped with Ethernet PCMCIA cards and Ricochet modems. They trigger handoffs from one network to another based on the insertion and removal of Ethernet PCMCIA cards. Bhagwat [6] also deals with the problem of multiple network interfaces, handling the routing aspects of multiple network interfaces as a special case of Mobile IP. Our work differs from theirs in that it focuses on how to switch from one network interface to another in a manner that is completely transparent to the user.

8. Conclusions

We have described additions to a horizontal handoff system to support the simultaneous operation of multiple wireless network interfaces. This vertical handoff system gives mobile devices the ability to roam freely in wireless overlay networks with seamless transitions between networks and with negligible interruption to applications. Implementing handoffs efficiently between multiple network interfaces introduces inherent trade-offs between handoff latency and power and bandwidth overheads. Rather that depending on network-specific channel measurements to predict disconnections, our schemes require no knowledge about specific channel characteristics and depend only on higher-order information such as the presence or absence of beacon and data packets. We present detailed measurements of handoff latencies and their costs in terms of network resources for a variety of different schemes. Results show that a simple scheme leads to a handoff latency that is seconds long and is dominated by the time it takes the mobile to discover that the current overlay is unreachable. Enhancements to this basic scheme can reduce this penalty to as low as 170ms with a 1.5% overhead on network resources. For transitions from room-size to wide-area data networks, the handoff latency from the basic system can be reduced to approximately 800ms as a result of fast beaconing. Other enhancements either have a high cost in terms of bandwidth overhead or do not decrease handoff latency, due largely to the latency-bound nature of the wide-area network being used.

We can make the following generalizations from the specifics of our implementation for future designers of overlay networks:

- Not all transitions between levels in the overlay network hierarchy can be treated identically. In our system, the choice of enhancement that resulted in the best performance was specific to the pair of networks that were chosen. This implies that a fixed policy will not work well for all choices of pairs of network interfaces, and a more flexible (and heuristic) approach will have to be used.
- The diversity that arises from being able to receive packets on multiple network interfaces simultaneously was invaluable in implementing the enhancements of Section 6.
- Depending on the presence or absence of data packets rather than channel measurements allowed us to rapidly add new network interfaces to our hierarchy. For example, adding the Ricochet overlay to the experimental setup took a matter of hours. If we had depended on channel-specific measurements to trigger a handoff, adding the Ricochet overlay to our system would have taken much longer. In addition, by depending only on packet reception, we can handle in an identical way causes for disconnection other than mobility, such as the insertion and removal of PCMCIA network cards. This can be considered the end-to-end approach in determining connectivity — the most meaningful metric is whether or not a MH can receive data via a particular overlay.

9. Future Work

Future directions for research are:

- Our working system does not use geographic hints to limit the use of the enhancements described in Section 6 to predict when a handoff is likely. We plan to add and analyze the effectiveness that simple hints such as cell connectivity have in predicting the likelihood of imminent handoffs.
- This work presents a single policy that drives the choice of enhancement, BS and network to use. By using the advice mechanism described in Section 3.2, we plan to experiment with more sophisticated application- and user-specific policies for choosing enhancements and forcing handoffs to new BSs and networks.
- The header and packet doublecasting enhancements we use depend on the fact that packets are being sent to BSs of different networks. Currently, these data flows are identical. For networks that have vastly different characteristics, this is not an ideal situation for a user who is receiving 500Kb of full-motion audio and video over an in-building RF network and is about to hand off to a wide-area data network. Similar to the approach of layered video dissemination in [20], we are experimenting with the idea of *delivery classes* of traffic specified at the source and routing different subsets of delivery classes to different networks as a function of the network's characteristics.

10. Acknowledgments

Thanks go to Hari Balakrishnan, who helped debug some of the kernel enhancements made for faster handoffs. Hari Balakrishnan, Armando Fox, Yatin Chawathe and Venkat Padmanabhan provided many helpful comments on early drafts of this paper that greatly increased the presentation of the material. This work is supported by DARPA contract DAAB07-95-C-D154 and grants from the California MICRO Program, Hughes Aircraft Corporation, Metricom, Deimler Benz, and PCSI.

References

[1] A. S. Acampora and M. Naghshineh. An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM. *IEEE Journal on Selected Areas in Communications*, 12(8):1365–1375, October 1994.

[2] Ardis web page. http://www.ardis.com, 1996.

[3] M. D. Austin and G. L. Stuber. In-service Signal Quality Estimation for TDMA Cellular Systems. In *Proc Sixth IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '95)*, volume 2, pages 836–840, Sept 1995.

[4] M. Baker, X. Zhao, S. Cheshire, and J. Stone. Supporting Mobility in MosquitoNet. In *Proc. of the 1996 USENIX Conference*, January 1996.

[5] H. Balakrishnan, S. Seshan, and R.H. Katz. Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks. *ACM Wireless Networks*, 1(4), December 1995.

[6] P. Bhagwat. *A framework for integrating Mobile Hosts within the Internet*. PhD thesis, University of Maryland, December 1995.

[7] A. Brandao, L. Lopes, and D. McLernon. Quality assessment for pre-detection diversity switching. In *Proc Sixth IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '95)*, volume 2, pages 577–581, Sept 1995.

[8] R. Caceres and L. Iftode. Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments. *IEEE Journal on Selected Areas in Communications*, 13(5), June 1995.

[9] J. Chen and M. Schwartz. Two-tier Resource Allocation for a Multimedia Micro-cellular Mobile System: Performance Summary. In *Proc Sixth IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC '95)*, volume 3, pages 1067–1072, Sept 1995.

[10] D. C. Cox. Wireless personal communications: What is it? *IEEE Personal Communications*, pages 20–34, April 1995.

[11] George H. Foreman and John Zahojaran. The challenges of mobile computing. *IEEE Computer*, 27(4):38–47, April 1994.

[12] V. Garg and J. Wilkes. *Wireless and Personal Communications Systems*. Prentice-Hall, 1996. Chapter 8.

[13] R. Ghai and S. Singh. An Architecture and Communications Protocol for Picocellular Networks. *IEEE Personal Communications Magazine*, 1(3):36–46, 1994.

[14] IBM Infrared Wireless LAN Adapter Technical Reference. IBM Microelectronics - Toronto Lab, 1995.

[15] J. Ioannidis, D. Duchamp, and G. Q. Maguire. IP-based Protocols for Mobile Internetworking. In *Proc. ACM SIGCOMM '91*, pages 235–245, 1991.

[16] J. Ioannidis and G. Q. Maguire. The Design and Implementation of a Mobile Internetworking Architecture. In *Proc. Winter '93 Usenix Conference*, San Diego, CA, January 1993.

[17] R. H. Katz and E. A. Brewer. The Case for Wireless Overlay Networks. In *Proc. SPIE Multimedia and Networking Conference (MMNC'96)*, San Jose, CA, Jan 1996.

[18] J. s. Lee. Overview of the Technical Basis of Qualcomm's CDMA Cellular Telephone System Design: a View of North American TIA/EIA IS-95. In *Proc ICCS '94*, volume 2, pages 353–358, Nov 1994.

[19] S. McCanne and V. Jacobson. The BSD Packet Filter: A New Architecture for User-Level Packet Capture. In *Proc. Winter '93 USENIX Conference*, San Diego, CA, January 1993.

[20] Steven McCanne, Van Jacobson, and Martin Vetterli. Receiver-driven Layered Multicast. In *Proc ACM SIGCOMM '96*, August 1996.

[21] Metricom web page. http://www.metricom.com, 1996.

[22] C. Perkins. *IP Mobility Support*. RFC 2002, Oct 1996.

[23] Tekinay S. and B. Jabbari. Handover and Channel Assignment in Mobile Cellular Networks. *IEEE Communications Magazine*, 29(11):42–46, November 1991.

[24] S. Seshan. *Low Latency Handoffs in Cellular Data Networks*. PhD thesis, University of California at Berkeley, December 1995.

[25] S. Seshan, H. Balakrishnan, and R. H. Katz. Handoffs in Cellular Wireless Networks: The Daedalus Implementation and Experience. *Kluwer Journal on Wireless Personal Communications*, January 1997.

[26] R. Steele and M. Nofal. Teletraffic Performance of Microcellular Personal Communication Networks. *IEE Proceedings I (Communications, Speech and Vision)*, 139(4):448–461, August 1992.

[27] M. Stemm, P. Gauthier, D. Harada, and R. H Katz. Reducing Power Consumption of Network Interfaces in Hand-Held Devices. In *Proc. Third Workshop on Mobile Multimedia Communications (MoMuC-3)*, Sept 1996.

[28] AT&T WaveLAN: PC/AT Card Installation and Operation. AT&T manual, 1994.

Mark Stemm (ACM S 1995, IEEE S 95) is a Ph.D. candidate in Computer Science at the University of California at Berkeley. His research interests are in the areas of wireless and wide-area computer networks, mobile computing, and operating systems. He received his B. S. degree in Computer Science with University and Departmental honors from Carnegie Mellon University in 1994 and his M.S. degree in Computer Science from the University of California at Berkeley in 1996.

On the WWW, his URL is: http://www.cs.berkeley.edu/~stemm. His e-mail address is: stemm@cs.berkeley.edu.

Professor Randy H. Katz is a leading researcher in computer system design and implementation. His research experience has spanned numerous disciplines. He has written over 120 technical publications on CAD, database management, multiprocessor architectures, high performance storage systems, and video server architectures. He was responsible for developing the concept of Redundant Arrays of Inexpensive Disks (RAID), now a $3 billion industry segment. Katz's recent research has focused on wireless communications and mobile computing applications.

From January 1993 through December 1994, Katz was a program manager and deputy director of the Computing Systems Technology Office of ARPA. He was responsible for "wiring the White House" to the Internet, and also assisted the Clinton Administration in formulating policies related to the National Information Infrastructure and wireless technologies.

He is a fellow of the ACM and the IEEE.