

A Fast Handoff Scheme for Wireless Networks

[†]Cheng Lin Tan, ^{*}Stephen Pink
^{†*}Computer Science & Electrical Engineering
Luleå University of Technology
[†]chenglin@acm.org, ^{*}steve@cdt.luth.se

[‡]Kin Mun Lye
^{†‡}Centre for Wireless Communications
National University of Singapore
[‡]elelyekm@leonis.nus.edu.sg

Abstract

In this paper, we propose a fast and efficient handoff scheme to handle the movements of mobile nodes among small wireless cells at the fringes of the Internet. Our scheme adopts a hierarchical mobility management architecture to restrict the handoff processing overheads within the vicinity of the mobile node, and uses multicast as the packet forwarding mechanism to deliver packets to multiple base stations within the vicinity of the mobile node to achieve fast handoff performance. Our scheme is based on the Internet Protocol (IP) and is compatible with Mobile IP and its route optimization option.

We also present simulation results for our simulation using the Network Simulator (ns2). The simulations show that our handoff scheme is fast enough to meet the requirements of an interactive voice communication session. The first packet from the new base station arrives at the mobile node within 10 ms after the mobile node initiates a handoff. Hence our scheme is suitable for roaming mobile nodes which may encounter numerous handoffs while they are in the midst of an interactive voice communication session such as Internet telephony.

1 Introduction

The Internet will be a huge wired internetwork with many small wireless cells at the edges to connect mobile devices to the wired part of the network. Each cell has a base station which acts as the gateway to the wired network for the mobile nodes. The advantages of having a smaller cell size include higher data throughput, better frequency reuse, location information with finer granularity, and the lower power transceivers required at the mobile nodes. Users access the Internet via the base stations using small portable devices, and handoffs between cells in the midst of a data transfer are very common. Hence a fast and efficient handoff scheme is needed for these mobile devices to maintain connectivity, minimize data loss and latency while crossing cell boundaries during data transfers.

The IETF Mobile IP standard [11] was intended to solve

the general problem of host mobility in the Internet. For our scenario of mobile devices crossing cell boundaries frequently, using Mobile-IP will result in frequent location updates, and the latency involved in the re-routing of the packets will not be acceptable for mobile hosts engaging in real-time communication session like Internet telephony. As noted in [8], that while Mobile IP (and IPv6) is designed to address the “macro” mobility management problem, such as supporting host mobility over wide-area networks, it does not address micro-level mobility issues such as packet loss and delay due to handoffs.

In this paper, we propose a fast and efficient handoff scheme that supports handoffs of mobile devices crossing wireless cell boundaries frequently during interactive voice communication sessions. We adopt a domain foreign agent concept to hide mobility of mobile nodes within the foreign domain from the home agent. We achieved fast handoff performance by using multicast as the packet forwarding mechanism from the domain foreign agent to the base stations at the vicinity of the mobile node.

Our base stations are network-layer routers with buffers, and are capable of subscribing to multicast groups. Our scheme aims to reduce the overheads of mobility management while maintaining the quality of a real time interactive session such as Internet telephony during handoffs. It should minimize the protocol overhead on the networks while meeting the stringent requirements of interactive speech places on network jitter, delay and loss.

We have simulated our scheme using the Network Simulator (ns2)¹ and the simulation results show that the first packet from the new base station arrives at the mobile node within 10 ms² after the handoff is initiated by the mobile node. Our result is valid even in a wireless environment with coverage gaps where a mobile node loses connectivity with the previous base station before a handoff is initiated.

The rest of our paper is organized as follows. Section 2 presents related work. Section 3 describes our mobility management architecture and handoff protocol. Section 4 presents the details of our protocol. Section 5 reports the performance results of our simulation. Section 6 highlights some areas for future work, and Section 7 concludes the paper.

¹Network Simulator is a network simulation tool developed by LNBL, USC/ISI, Xerox PARC and UCB.

²For a simulated wireless network of 1.2 Mbps bandwidth and 4 ms link delay.

2 Related Work

The problem of excessive mobility management traffic has been addressed numerous times in the literature. Pollini et al. [14] has shown that cellular telephone networks have to carry much more signaling traffic than wired telephone networks due to mobility management traffic. Various work related to hierarchical mobility management have been proposed to reduce the problem of excessive mobility management traffic in cellular networks. (EIA/TIA) Interim Standard 41(IS-41) [4] and the Global System for Mobile Communications (GSM) mobile application part (MAP) [9] are two standards for location management for cellular networks and both use the Home Location Registers (HLR) and Visitor Location Registers (VLR). Jain and Lin [7] use a previously visited VLR and a series of forwarding pointers to reduce the location update traffic back to the HLR. Ho and Akyildiz [6] choose a VLR close to the mobile terminal to be its local anchor, which is made known to the HLR. Location updates are reported only to the local anchor.

For connection-less network such as the Internet, Cáceres and Padmanabhan [1] propose a hierarchical mobility management scheme where a domain foreign agent is used to hide the mobility from the home agent of the mobile node within the foreign administrative or security domain. The disadvantage is that the domain foreign agent has to maintain per-mobile node routing entries and update them whenever a mobile node moves.

Various proposals for fast handoff schemes can also be found in the literature. For handoff in a connection-oriented pico-cellular network, Ghai and Singh [5] propose a scheme in which picocells are formed as a group and each group is assigned a different multicast address. Messages addressed to a mobile user are multicast from a supervisor host to the mobile user using the multicast address associated with the group that the mobile node is presently located. When a mobile node moves, it moves into a different group and messages are forwarded to the mobile node using the new group multicast address. The disadvantages are that the supervisor host has to inform all the base stations involved (in both the old and new groups) of the movement of the mobile node, and the update and search mechanisms used by the system are not scalable.

For handoff in a connection-less network between adjacent base stations located in the same subnet, Cáceres and Padmanabhan [1] use a non-multicast minimal handoff protocol between base stations and mobile nodes to achieve low latency. The retransmission buffer size is tuned to the number of expected packet losses during a handoff, and the complete buffer is retransmitted from the old to the new base stations after every handoff to reduce packet loss. This option trades off additional handoff overheads over the wired networks for reduced handoff packet loss. The other disadvantage is that if the base stations are not on the same subnet, the method of forwarding packets from the old to the new base stations after a handoff may not meet the requirement of low latency during handoff.

Seshan et al. [15] propose a scheme in which each mobile node is pre-assigned a temporary IP multicast address by its home agent. The home agent encapsulates packets destined for the mobile node and forwards them to its associated multicast group. While only one base station actively forwards packets to the mobile node, the other base stations that are identified as likely handoff targets are asked by the mobile node to join the multicast group. These potential handoff target base stations buffer the recent few packets and can

quickly forward them to the mobile node should a handoff happen. The use of multicast relieves the home agent of the task of keeping track of the mobile node's location, but the scheme has one drawback. The scheme has to handle the IP multicast address management issues across the wide-area networks. Potential multicast address conflicts may happen if the multicast stream passes through an area whereby the same multicast address is used for another communication.

Mysore and Bharghavan [10] propose to use IP multicast as the sole mechanism for addressing and routing packets to mobile nodes. Each mobile node is identified by a unique multicast IP address, and packets destined for the mobile node are sent as multicast packets. This scheme has the advantage of integration of architectures supporting multicast and mobility, as well as seamless mobility during handoffs. However, it suffers from various compatibility issues with existing IP implementations such as ARP reply processing and TCP support over IP multicasting, making wide deployment of this scheme unlikely.

3 Mobility Management

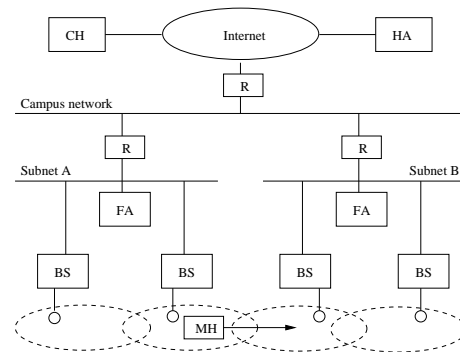


Figure 1: *Mobile IP entities in a wireless internetwork.*

A typical campus network with wireless extensions is shown in Figure 1. The various network segments are interconnected by routers (R), and mobile hosts (MH) access the Internet via the base stations (BSs) over the wireless links. As each MH moves, it leaves the wireless coverage of one cell and enters into another, resulting in a handoff between the BSs. For our scenario of small wireless cells at the fringes of the Internet, such handoffs will be frequent as wireless cells may be of the size of a few meters.

In Mobile IP, when the MH moves from one cell to another, causing a change of foreign agent (FA), the MH is required to send location updates to inform the home agent (HA) of its latest care-of address. This location update traffic has to traverse the wide-area network (if the MH is far away from the HA), and the handoff can incur long delay. Meanwhile packets are forwarded to the wrong FA before the location update message reaches the HA. For the route optimization option [13], packets from the correspondent host (CH) are sent to the wrong FA until the CH is notified of the MH's new care-of address. Hence we can see that Mobile IP alone is not suited to handle frequent handoffs of mobile devices in small wireless cells. Handoff latency is long and the control messages generated due to the frequent handoffs also add on to the load of the wide-area network. Our approach avoids these drawbacks while providing the desired handoff performance.

3.1 Hierarchical Structure

We propose to adopt the domain foreign agent (DFA) concept to shield all mobility within the foreign domain to the domain network itself. As long as the mobility of the mobile node is within the domain of the DFA, no location update traffic is generated. This eliminates any location update traffic going across the wide-area network due to the mobility of MHs within the foreign domain.

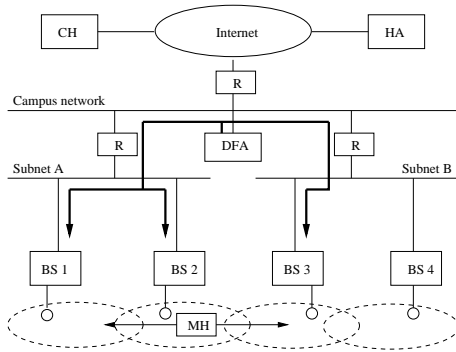


Figure 2: Hierarchical mobility management approach using DFA, and the use of multicast as the packet forwarding mechanism from the DFA to physically adjacent BSs.

For the purpose of exposition, we will use the scenario of a campus environment. As shown in Figure 2, a campus administrative domain may have a domain foreign agent which is responsible for all foreign mobile nodes within the campus. At the subnet level, agent advertisement messages containing the IP address of the DFA are broadcast periodically. Note that this functionality of broadcasting agent advertisement may be subsumed by the BSs for our case. When a mobile node hears a beacon and decides to attach to the wired network, it registers with the DFA and sends the IP address of the DFA to its HA as its care-of address. When the MH moves from one cell to another, resulting in a change of serving BS (but still within the campus), no location update is sent to the HA. As far as the HA is concerned, the MH is still within the campus in the DFA's domain, and all packets meant for the MH are encapsulated and tunneled to the DFA. Since we do not expect mobile nodes to move frequently from one campus to another campus in a very short span of time, minimum location update traffic is generated across the wide-area network. For roaming between domains, which is rather infrequent and slow, Mobile IP alone is adequate to handle such cases.

For route optimization, the CH is informed of the DFA's IP address as the MH's care-of address, and packets meant for the MH are sent directly to the DFA.

3.2 Handoff Protocol

To achieve fast handoff performance, we use multicast as the packet forwarding mechanism from the DFA to the base stations surrounding the MH. When a MH roams into a wireless cell within the campus and registers with the DFA, the DFA assigns a multicast address unique within its domain to the MH. The MH informs the serving BS (the BS serving the cell where the MH presently resides) to subscribe to this multicast group, as the DFA will forward packets meant for the MH as a multicast stream using this multicast address. The serving BS in turn informs all physically adjacent BSs

to subscribe to the same multicast group. While only the serving BS actively forwards packets (as a unicast) to the MH, the other adjacent BSs buffer the recent few packets so that they can forward these packets to the MH should a handoff occur.

The thick arrows from the DFA to the various BSs in Figure 2 show the paths of the multicast stream from the DFA to the physically adjacent BSs. If the MH is to move to either the left or right cell, both BS1 and 3 already have the recent few packets for the MH in their buffers. No forwarding of buffers from the old to the new BSs is needed. Handoff is fast and packet loss due to handoff is minimized. The multicast forwarding algorithm used between the DFA and the BSs can be any of the well-established ones for the wired networks, such as Distance Vector Multicast Routing Protocol (DVMRP) [3], or Protocol Independent Multicast-Dense Mode (PIM-DM) [2].

The advantages of using multicast as the packet forwarding mechanism from the DFA to the BSs are three fold. First, it alleviates the DFA from the task of keeping track of the exact location of the MH. As long as the MH remains within its domain, the DFA can be sure that the MH will receive packets encapsulated in the multicast stream using the assigned multicast address. The choice of keeping track of the exact location of the MH becomes optional, and is not necessary to ensure correct forwarding of packets to the MH. This approach is different from Ghai and Singh [5] whereby the supervisor host and the base stations together track the MHs as they move in order to ensure correct forwarding of packets. In our scheme the DFA only needs to manage the allocation of multicast addresses to the MHs within its domain during registration. This task is much easier than to maintain and update the BSs that the different MHs are attached to. Moreover, the assignment of multicast address to a MH can remain unchanged for the period of time (hours or even days) that the MH is operating within the DFA's domain, regardless of its exact location. This is again different from Ghai and Singh [5] whereby the multicast address used by the supervisor host to forward packets to the mobile node is totally dependent on the location of the mobile node. In our scheme, these multicast streams are restricted within the network in the domain itself and do not traverse the wide-area network. Hence the chances of a multicast address conflict like the case of Seshan et al. [15] are much lower. To eradicate multicast address conflict totally, mechanism of allocating multicast address globally, such as allocating a range of multicast address to each domain, must be used.

Second, no location update message back to the DFA is even needed for the correct forwarding of packets to the MH. When a MH moves from one cell to another, the new serving BS informs its neighboring BSs to subscribe to the multicast group that is assigned to the newly arrived MH. Since adjacent BSs are most likely to reside in the same network segment, we are using multicast for such communications as well to reduce network load. If the DFA is not concerned about the exact location of the MH, no location update message is sent back to the DFA, and minimum location update overhead is generated when MH moves from one cell to another. Hence the use of multicast as the packet forwarding mechanism eliminates the generation of location update traffic back to the DFA, but the correct forwarding of packets to the MH is still ensured.

Third, since physically adjacent cells are most likely to reside on the same network segment, the use of multicast as the packet forwarding mechanism to these adjacent BSs

provides an advantage. This is especially so for the case of shared-medium network such as Ethernet. Comparing the case of unicasting a packet to a single BS, to that of multicasting the same packet to a few BSs residing on the same network segment, the extra network load generated from the use of multicast is negligible. Hence the use of multicast in our scheme enables us to forward packets efficiently to physically adjacent BSs to give our scheme good handoff performance.

We believe that our hierarchical mobility management architecture coupled to the use of multicast as the packet forwarding mechanism, is a scalable and efficient solution to handle frequent handoffs of mobile nodes among small wireless cells.

4 Protocol

We simulated our handoff scheme using the Network Simulator tool (ns2) version 2.1b3. All code is written and simulated on a Sun Sparc 5 workstation running on Solaris 2.5.1 software platform. We highlight the important components of our protocol in the next few sections.

4.1 Registration

For a MH to know that it has roamed into the coverage area of a wireless cell, it has to listen constantly for beacons from the BSs. BSs carry the IP address of the DFA in their beacons as a form of agent advertisement message, as well as their own address. When a MH hears a beacon and decides to attach to the wired network via the BS, the following exchange of messages take place.

- The MH sends a *registration request* message to the DFA.
- The DFA processes the *registration request* and then relays it to the MH's HA. The DFA address is used as the care-of address in the registration request.
- The HA sends a *registration reply* to the DFA to grant the MH's request.
- The DFA relays the *registration reply* together with a IP multicast address to the MH. This IP multicast address is unique for that MH within the DFA's domain.

Note that besides the inclusion of the IP multicast address in the last message from the DFA to the MH, the other message exchanges are exactly the same as those stated in the Mobile IP specification [12]. The inclusion of the IP multicast address should pose no serious compatibility problem as it can be easily added to the list of extensions already present in the registration reply message.

4.2 Dynamic Virtual Macro-cells

In our scheme, we organized the base stations logically into Dynamic Virtual Macro-cells (DVMs). DVMs are formed by clusters of base stations adjacent to each other, and these DVMs overlap each other as shown in Figure 3. A BS may belong to more than one DVM, but each BS is a core of only one DVM. Only the core can transmit information while the other member BS in the same DVM should only listen. Two BSs in the same DVM need not necessarily be in the same subnet. As an illustration, in Figure 3, BS2 is the core of DVM A and also a member of DVM B. Similarly BS 3 is the

core of DVM B as well as a member of DVM A. A handoff can only happen between the core and any of its member BSs in the same DVM. For example, a MH can only handoff from BS2 to BS1 or 3.

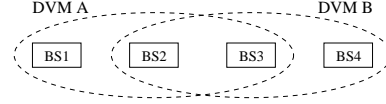


Figure 3: *Dynamic Virtual Macro-cell.*

In Figure 3, assuming that a MH roams into BS2's cell and registers successfully with the DFA. The MH informs BS2 to subscribe to the multicast group assigned to the MH by the DFA. BS2 will inform all other member BSs (BS1 and 3) in DVM A to subscribe to this multicast group as well. Each DVM core sends control information to the other member BSs using multicast and each DVM has its own multicast address. This reduces significantly the network load due to the control information flow between the members of the DVM when a MH crosses cell boundaries.

4.3 Handoff

Referring to the scenario shown in Figure 2, the MH is in BS2's cell and BS1 and 3 are buffering the recent few packets meant for the MH. Assumed that the MH moves to the right, hears the beacon from BS 3, and decides to do a handoff to BS3. The following are the message exchanges.

- After hearing the *beacon* from BS3, the MH decides to switch from BS2 to BS3 based on factors such as radio signal strength and quality of connection.
- MH sends a *greet* message to BS3, indicating its intention to switch over to BS3's cell. The *greet* message contains the IP address of the old serving BS (BS2), the IP multicast address assigned to the MH, and the IP ID of the last forwarded packet received by the MH from the old BS.
- BS3 sends back a *greet ack* message to the MH to confirm the handoff. BS3 also sends a *notify* message to BS2 to inform the later about the handoff. At the same time, BS3 starts forwarding packets to the MH, beginning with packet in its buffer received immediately after the packet whose IP ID is indicated in the *greet* message received from the MH. The use of IP ID is to minimize missing or duplicate packets being forwarded to the MH after the handoff. The protocol critical to a correct handoff can be considered as finished after the exchange of the above messages.
- BS3 multicasts a control message to inform the other member BSs (BS2 and 4) in its DVM to subscribe to the multicast group associated with the newly arrived MH. After receiving this message, BS4 subscribes to the multicast group associated with the MH and starts buffering the recent few packets meant for the MH. BS2 does nothing since it is already subscribed to the multicast group.
- After receiving the *notify* message from BS3, BS2 sends back a *notify ack* message to BS3 and stops forwarding packets to the MH.

- BS 2 multicasts a control message to inform the other member BSs (BS1 and 3) in its DVM that MH has left its cell, and these member BSs can prune themselves off the multicast tree associated with the MH if necessary. In this case, since the serving BS is no longer adjacent to BS1 after the handoff, BS1 can unsubscribe from the multicast group associated with the MH.

The control information packet sent out by BS2 and 3 (as multicast) due to the handoff contains minimum information and has a very short packet size. Hence the extra network load generated by these control packets in our scheme is negligible as compared to other schemes which forwards data packets from the old BS to the new BS, or schemes that require location update packets to traverse the wide-area network when a handoff occurs.

4.4 Beacon Period and Buffer Size

Every BS transmits a beacon periodically. Besides serving as an agent advertisement message, this signal is important to a mobile node as an aid to detect its own movement. A mobile node can listen to transmissions from BSs to other mobile nodes in the same region to identify its own location. If such transmissions are absent, the mobile node can realize that it has roamed out of the coverage area of its serving BS either by detecting a missing beacon from its serving BS, or hearing a beacon from another BS. In our scheme, a mobile node sends out a BS solicitation message once it detects a missing beacon. BSs who heard this solicitation message must send out a beacon, and if the mobile node hears multiple beacons, it can decide which BS to handoff to by looking at criterion such as received signal strength of the beacon.

First consider the general case where wireless cells are overlapping with no coverage gap, and the BSs involved in the handoff are of the same network hierarchy (meaning that a packet from the DFA will arrive at both BSs at about the same time). For the ideal case of eliminating packet loss due to a handoff, the amount of buffers needed at the BSs should be equivalent to the maximum possible amount of packet loss due to the handoff. We define the rendezvous time as the time taken for a mobile node to hear a beacon from a new BS after roaming out of the old BS's cell. Hence the rendezvous time determines how soon a mobile node can detect its movement out of a wireless cell and initiate a handoff. In a wireless environment with approximately-synchronous beacon system (all BSs send out beacons approximately at the same time, with a small time offset just enough to prevent collisions of beacons between adjacent base stations), the worst case rendezvous time is equal to the beacon period. Equation (1) shows the relationship between the rendezvous time, the packet inter-arrival time, and the maximum possible number of packet loss per connection during a handoff (without any buffering scheme). The amount of buffer needed at the base station to support handoff for multiple connections can be extrapolated easily.

$$\text{maximum number of packet loss during a handoff} = \lceil (\text{rendezvous time}) / (\text{packet inter-arrival time}) \rceil + 1 \quad (1)$$

The worst case scenario can be a situation where a packet is transmitted immediately after the beacon from BS X, and the mobile node leaves the wireless cell of BS X before receiving the complete packet. While the mobile node is outside

the coverage area of BS X, all packets for the mobile node which arrive at BS X before any handoff is initiated are lost (assuming no buffering is done), and the mobile node can only initiate a handoff after hearing the beacon from a new BS. The amount of packet loss can be reduced if the mobile node can detect its movement out of a wireless cell sooner and initiate a handoff earlier to a nearer BS. Hence a shorter rendezvous time will help to reduce the amount of buffers needed at the BSs to eliminate packet loss during handoff.

For the case of a small wireless cell environment with coverage gap, the maximum possible number of packet loss is totally dependent on the mobile node's mobility pattern. Infinite amount of buffers are needed if the mobile node decides to stay put at the coverage gap indefinitely.

In our scheme, each BS buffers the recent few packets meant for mobile nodes residing at physically adjacent cells and First-in-First-out (FIFO) buffers are used. The fast handoff is achieved by a BS forwarding these packets from its buffers should the mobile node moves into its cell.

A short beacon period consumes more wireless bandwidth, increases processing overheads, but reduces the amount of buffers required at the BSs because the number of expected lost packets is lower. On the other hand, a longer beacon period increases the number of buffers needed at the BSs but consumes less wireless bandwidth and reduces processing overheads.

5 Performance Simulation

Our aim is to determine whether our handoff scheme can meet the requirements to support real-time interactive communications like Internet telephony during handoff. We also want to find out how our scheme helps in the throughput of a reliable transport protocol transfer during handoff.

5.1 Simulation Scenario

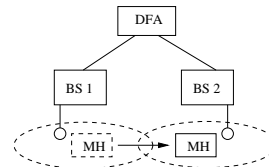


Figure 4: *Simulation Scenario.*

Our simulation scenario is shown in Figure 4. In our ns simulations, the DFA and the BSs are connected by wired links, and the MH is attached to the BSs through wireless links. The wired and wireless networks are simulated using 10 Mbps duplex links and 1.2 Mbps duplex links respectively. No loss of signaling or data packets are simulated. For our packet audio source, we chose the pulse code modulation (PCM) as the Internet telephony audio coding format to simulate the most network resource demanding case. Using Mbone applications like vat as the yardstick, the shortest packet inter-arrival time for PCM format is 20 ms, and the average packet size is 200 bytes. We use these parameters in our simulations of an Internet telephony source.

5.2 Handoff Performance

The objective of this simulation is to find out the time needed for our scheme to complete a handoff. The time to complete a handoff has two components: the rendezvous

time and the handoff latency. The handoff latency is defined as the difference in time between the arrival of the first new packet from the new BS's buffer and the time at which the MH sends a handoff request to the new BS. In our case, this includes the exchange of *greet*, *greet ack* messages and the arrival of the first packet from the new BS at the mobile node.

Our experiment involved the DFA sending out UDP multicast packets of 200 bytes each at an interval of 20 ms, to simulate the forwarding of encapsulated packets from the DFA to the MH. We measured the handoff latency when the MH does a handoff from BS1 to 2 while receiving this multicast stream. We also performed our simulations using wireless networks of different bandwidths and link delays. The networks have no other traffic other than the UDP packets and the protocol overheads. The results are plotted in Figure 5. For a typical wireless network of 1.2 Mbps bandwidth and 4 ms link delay, the handoff latency for our scheme is 10 ms. Hence the total handoff time for our scheme in such a wireless network is equal to the rendezvous time plus 10 ms.

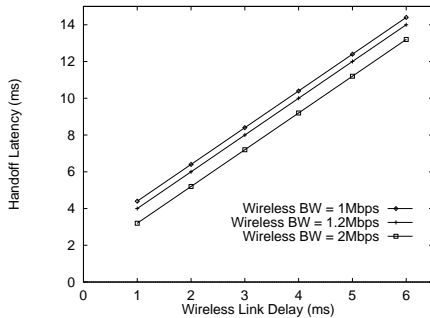


Figure 5: *The handoff latency of our scheme for wireless networks of different bandwidth and link delay parameters. (BW = bandwidth)*

For all the subsequent simulations, the wireless networks simulated have a bandwidth of 1.2 Mbps and 4 ms link delay.

5.3 Beacon Period

The goal of this simulation is to find the minimum beacon period to give a short rendezvous time, and yet one that will not affect the overall throughput performance of the system. This simulation involved the sending of 4 Mbytes of data from the DFA to the MH, with a receiver advertised TCP window size of 15 Kbytes and 1024 bytes segment. Each beacon is 50 bytes long. We measured the throughput of this transfer for the different beacon periods.

It was found that the throughput is above 99% of the maximum for beacon periods of 100 ms or longer, and it drops to 97% of the maximum when the beacon is 13 ms. Hence we conclude that the minimum beacon period we can choose to give a low rendezvous time and involved minimum overhead in our simulations is approximately 100 ms.

5.4 Packet Audio Performance

The objective of this simulation is to find out whether our scheme can minimize the effects of a handoff while the MH is in the midst of an Internet telephony session during handoff. In our simulation, the DFA multicast a stream of packets of

200 bytes every 20 ms to simulate the forwarding of a real-time Internet telephony stream to the MH, while the MH moves from BS 1 to 2. Various human factors studies have shown that the maximum tolerable delay for an interactive conversation is approximately 200 ms. This helps to set the maximum tolerable rendezvous time and beacon period the system can allow.

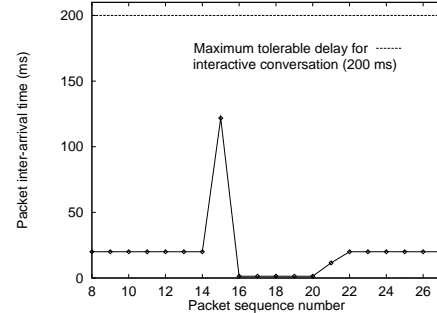


Figure 6: *The packet inter-arrival time before, during and after a handoff.*

Figure 6 shows the packet inter-arrival time before, during and after a handoff for a wireless system with approximately-synchronous beacon period of 100 ms and buffer size of 6 packets at the BSs. The handoff scenario simulated involved the worst rendezvous time of 100 ms. There are several points to highlight regarding these results. First, the jitter introduced by our handoff scheme is below the maximum tolerable delay for interactive conversation. The longest packet inter-arrival time is 122 ms, of which 100 ms is contributed by the rendezvous time. Packets with sequence number from 16 to 20 are transmitted back-to-back from the new BS and have very short packet inter-arrival time between them.

Second, we can see that no packets are lost, duplicated or have arrived in the wrong sequence. The provision of enough buffers at the BSs ensures that packets not received by the mobile node during handoff are stored, and the use of IP IDs in our scheme ensures that these stored packets are forwarded to the mobile node from the new BS in the correct sequence to avoid extra processing at the application level. If the number of buffers at the BSs is smaller than required to eliminate packet loss due to handoff, our scheme is still better than other handoff scheme which involve buffer forwarding like Cáceres and Padmanabhan [1]. Ours has a shorter packet inter-arrival time for those packets forwarded from the buffers of the new BS.

5.5 Reliable Transport Protocol Performance

Besides real time applications, a fast and reliable handoff algorithm should help the performance of applications that use reliable protocols such as TCP. The goal of the simulations in this section is to measure the performance of our handoff scheme when handoff happens while the mobile node is in the midst of a TCP connection, such as a File Transfer Protocol (FTP) transfer.

In these simulations, the ftp source is at DFA and the ftp sink is at the MH. The receiver advertised TCP window size is 15 Kbytes and the segment size is 1024 bytes. We measured the TCP throughput of a 4-Mbyte transfer, with various number of handoffs equally spaced during the lifetime of the file transfer. We did the simulations for rendezvous time of 50 and 200 ms, with and without buffers.

For simulations with buffers, each BS has a buffer size of 15 packets (so that the total buffer size matches the TCP window size of 15 Kbytes). This number of buffers ensures that no packet is lost due to a handoff.

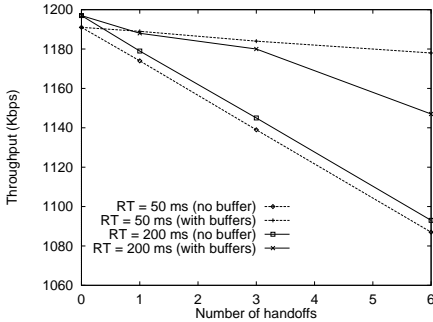


Figure 7: The TCP throughput of a FTP transfer using our handoff scheme (with and without buffers) to handle handoffs during the transfer. RT = rendezvous time.

From Figure 7, we can see that the throughput using our scheme with buffers (equivalent to the TCP window size) is higher than that without buffers for all cases of rendezvous time. For example, in the case where the rendezvous time is 200 ms with 6 handoffs during the lifetime of the file transfer, the TCP throughput with buffers is 5% higher than that without buffers. The cross-over in Figure 7 for cases with buffers (RT=50ms and 200ms) between 0 and 1 handoff can be explained by the fact that the gain in TCP throughput from initiating a handoff earlier more than compensates the loss in throughput due to more frequent beacons in the case of 50ms rendezvous time. Note that in the case where the BSs do not use any buffer, the new route to the MH has already been set up (using our scheme of multicast). No time is needed to compute the new route to the MH after the handoff. Hence our scheme with buffers is expected to perform much better than other schemes which have no buffer and have to update either the home agent or the correspondent host of the new route to the MH after the handoff.

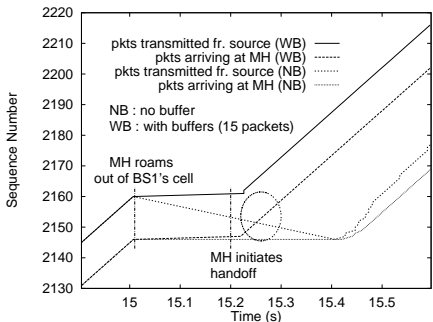


Figure 8: The TCP traces before, during and after the handoff, for scenarios with and without buffer. Rendezvous time = 200 ms.

Figure 8 captures the TCP traces before, during and after the handoff when the MH moves from BS1 to 2. Looking at the TCP trace where the BS has a 15 packet buffer, those packets transmitted by BS1 after the MH has left BS1's cell did not reach the MH. However, these packets were forwarded to the MH from BS2's buffer after the MH initiated a

handoff to BS2. These forwarded packets from BS2's buffer arriving at the MH are shown in the circle in Figure 8. The TCP source did not detect any packet loss after the handoff and the transfer continued from the last packet forwarded from BS1.

For the case where BS did not have any buffer, those packets that did not reach MH are lost. Immediately after the handoff, the source was not able to send anymore packets as the sender's TCP window was zero. This prevented the fast retransmit algorithm from being invoked (which would have shortened the delay for the retransmission) as no duplicate acknowledgments were generated by the MH after the handoff. As a result, the source had to wait for the TCP retransmission timer to expire before retransmitting the last unacknowledged packet to the MH. After the retransmitted packet reached the MH, the TCP transfer continued from there. This explains why the TCP throughput is higher for the scheme where BSs have 15-packet buffers, as compared to one where no buffer is used.

We also performed simulations to find out the effect on TCP throughput when the BSs do not have enough buffers to eliminate all packet loss due to a handoff. This happens when buffer size at the BSs are smaller than the receiver advertised TCP window size at the MH. We measured the TCP throughput of a file transfer for different buffer sizes at the BSs. To simulate multiple handoffs, the MH moves between BS1 and BS2. The results are shown in Figure 9.

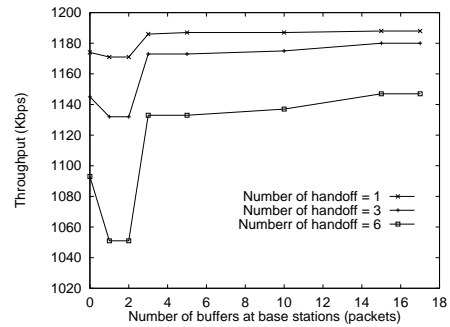


Figure 9: The effect of the number of buffers at the BSs on the throughput of a TCP transfer with handoffs.

From Figure 9, we can see that for buffer size of 3 packets and above, the TCP throughput increases when there is an increase in the number of buffers at the BSs, until the total buffer size reaches the receiver advertised TCP window size (15 Kbytes). Further increase in the number of buffers does not increase the TCP throughput. This is because the number of packets forwarded from the new BS's buffer will not exceed the receiver advertised TCP window size divided by the segment size.

We observe that when the buffer size is 1 or 2 packets, the TCP throughput falls dramatically, even worse than the case when the BSs do not use any buffer! We captured the TCP traces for these cases and identified the cause for the low throughput as the absence of enough duplicate acknowledgment packets from the MH after the handoff to trigger a fast retransmission from the TCP source. Most implementations of TCP assume that a packet is lost and trigger the fast retransmission only if the source receives 3 consecutive duplicate acknowledgment packets from the receiver. In the case of BS having a buffer size of 2 packets, the TCP source only received 2 duplicate acknowledgment packets after the handoff is initiated, and therefore fast retransmission

was not invoked. As the sender TCP window size was zero, the source had to wait for its retransmission timer to expire (starting from the time the last packet was transmitted after the handoff), before it could retransmit the packet with the right sequence number. The retransmission timer for the case where BSs have no buffer expired much earlier since it starts from the time when the last packet was transmitted by the source just before the MH roams out of old BS's cell. Hence the overall TCP throughput for the case where BSs have no buffer is higher than the case where BSs have a buffer size of 1 or 2 packets.

This problem happens when the sender TCP window size is zero just before the handoff. This is a very common situation when the source of a TCP transfer is at the wired network with a higher bandwidth, and the receiver is a mobile node connected to the network via a lower bandwidth wireless link. Hence this is a generic problem for handoff schemes that use buffer forwarding technique, if a buffer size of 1 or 2 packets is used.

From our simulations, we conclude that our handoff scheme works well for TCP if the buffer size chosen per TCP connection is anywhere between 3 and the receiver advertised TCP window size.

6 Future Work

The work presented in this paper can be extended in numerous ways. An intelligent neighborhood discovery protocol can be incorporated to eliminate tedious manual configuration of dynamic virtual macro-cell membership information at the base stations. Authentication mechanism needs to be used for multicast communications among base stations in the same dynamic virtual macro-cell to prevent denial of service attack. Encryption mechanism is also required to prevent eavesdropping of data packets being forwarded from the domain foreign agent to the mobile node using multicast. It is also interesting to extend our handoff scheme to handle multiple traffic types with different requirements for network bandwidth, delay and jitter. Finally, more complex scenarios involving multiple roaming mobile nodes can be studied to determine the best buffer management mechanisms at the base stations for our handoff scheme.

7 Conclusions

In this paper, we have presented a handoff scheme with a hierarchical mobility management architecture coupled to the use of multicast as the packet forwarding mechanism, to handle the frequent handoffs of mobile nodes in an environment with small wireless cells. From simulation results, we have shown that our scheme meets all delay requirement related to interactive voice communications, as well as providing good quality of service to packet audio applications during handoff. Our scheme is compatible with Mobile IP and its route optimization option. We have also shown that our handoff scheme works well for TCP if the buffer size chosen per TCP connection is anywhere between 3 and the receiver advertised TCP window size.

The popularity of Internet telephony and video conferencing applications has demonstrated the demand and usefulness for such real-time interactive applications. The handoff scheme presented in this paper enables mobile users to enjoy the benefits of such applications with the same quality of service without the bound of a tether.

Acknowledgments

We would like to thank Malin Flodin for her help in the initial stage of using ns2, Björn Nordgren for his help in setting up the programming environment, and the reviewers for their comments on the earlier drafts of the paper.

References

- [1] R. Cáceres and V. N. Padmanabhan. Fast and scalable handoffs for wireless internetworks. In *Proc. of ACM/IEEE MobiCom*, pages 56–66, New York, USA, Nov. 1996.
- [2] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, A. Helmy, and L. Wei. Protocol independent multicast version 2, dense mode specification. Internet Draft, Internet Engineering Task Force, May 1998. Work in progress.
- [3] S. Deering and C. Partridge. Distance vector multicast routing protocol. RFC, RFC 1075, Nov. 1988.
- [4] EIA/TIA. *Cellular Radio-Telecommunications Inter-system Operations*. Tech. Rep. IS-1 Revision C, 1995.
- [5] R. Ghai and S. Singh. An architecture and communication protocol for picocellular networks. *IEEE Personal Communications Magazine*, 1(3):36–46, 1994.
- [6] J. S. M. Ho and I. F. Akyildiz. Local anchor scheme for reducing location tracking costs in PCNS. *IEEE/ACM Trans. Networking*, 4(5):709–726, Oct. 1996.
- [7] R. Jain and Y. B. Lin. An auxiliary user location strategy employing forwarding pointers to reduce network impact of pcs. *ACM-Baltzer J. Wireless Networks*, 1(2):197–210, July 1995.
- [8] D. Johnson and C. Perkins. Mobility support in IPv6. Internet Draft, Internet Engineering Task Force, Mar. 1998. Work in progress.
- [9] M. Mouly and M.-B. Pautet. *The GSM System for Mobile Communications*. published by the authors, France, 1992.
- [10] J. Mysore and V. Bharghavan. A new multicasting-based architecture for internet host mobility. In *Proc. of ACM/IEEE MobiCom*, pages 161–172, Budapest, Hungary, Sept. 1997.
- [11] C. Perkins. IP mobility support. RFC, RFC 2002, Oct. 1996.
- [12] C. Perkins. IP mobility support version 2. Internet Draft, Internet Engineering Task Force, Nov. 1997. Work in progress.
- [13] C. Perkins and D. Johnson. Route optimization in mobile IP. Internet Draft, Internet Engineering Task Force, Dec. 1997. Work in progress.
- [14] G. P. Pollini, K. S. Meier-Hellstern, and D. J. Goodman. Signaling traffic volume generated by mobile and personal communications. *IEEE Communications Magazine*, 33(6):60–65, June 1995.
- [15] S. Seshan, H. Balakrishnan, and R. H. Katz. Handoffs in cellular wireless networks: The daedalus implementation and experience. *Kluwer International Journal on Wireless Personal Communications*, Jan. 1997.