A Protocol for a Wireless Network of Mobile Devices

by

Xiaolan Qian

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of

Master of Engineering

and

Bachelor of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2000

Author . . . .
Department of Electrical Engineering and Computer Science
May 19, 2000

Certified by . . . .
John T. Wroclawski
Reseach Scientist
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

A Protocol for a Wireless Network of Mobile Devices

by

Xiaolan Qian

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2000, in partial fulfillment of the
requirements for the degrees of
Master of Engineering
and
Bachelor of Science in Computer Science and Engineering

Abstract

Recent trends in wireless network development have focused on higher bandwidth and robust performance. However, a network of small, cost effective single frequency radios is advantagous when used by a class of applications that do not require the high performance that more sophisticated radios can provide. These applications are simple, ubiquitous and numerous and are targetted to an indoor environment. Low power, small form factor, and cost efficiency are critical to these applications. Existing protocols do not addequately accomodate for the lack of robustness in such radios as they are designed for ones capable of spread-spectrum modulation.

This thesis presents the design, simulation, and analysis of wireless mobility solution for simple wireless applications using small radios in a picocellular environment with a robust MAC protocol to improve performance. The TDMA-based MAC protocol implements an adaptive multi-tiered handoff algorithm to manage mobile connectivity while nodes roam the wireless network. This algorithm provides sufficient responsiveness to mobile movement as well as provisions to optimize network performance. A simulation of the MAC protocol is built and the performance of a wireless network is anaylized.

Thesis Supervisor: John T. Wroclawski
Title: Reseach Scientist

Acknowledgments

I would like to thank my advisor, John Wroclawski for his guidance and support during the course of this project. Without his knowledge and advice, this project would not have been possible. I would also like to thank Dorothy Curtis, John Ankron, Tim Shepard, Xiaowei Yang, Dina Katabi, and Jinyang Li for answering my questions about anything and everything.

My time at MIT has been blessed with the love and support of wonderful friends. I would like to show my graditude to Sandia Ren, Duncan Bryce and Charatpong Chotigavanich, and all my other friends for their understanding, support, and above all, constant friendship that has made these years at MIT most memorable. I thank Cary Lai for helping me get through these few years. I would also like to thank Long Phan for all his patience and encouragement that gave me strength when I needed it the most.

Most importantly, I would like to express my sincerest gratitude to my parents for their love, encouragement, and guidance in every momment of my life.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

The goal of wireless networking is to replace the cables in the network topology, thus allowing untethered mobility. Providing a connection that is just as perpectual and a bandwidth that is just as large as a cable connection is the challenge in building wireless networks. Currently wireless networks have been widely deployed in the form of wireless LANs (WLAN) using spread-spectrum radios that provide bandwidthes of 1-2Mbps, sufficient for most PC-based applications. As radio technology advances at an ever increasing pace, the number of possible applications become unlimited. Some applications may require more bandwidth than the current wireless LAN technology can provide. These applications include ones that deliver multimedia content to the roaming user. Other applications require much less bandwith. They tend to be much smaller than the laptop and have more specific tasks. These include wireless pens, PDA devices, and various environmental controller through out the house. Instead of high bandwidth, this simpler class of applications require radios that have small form factor, lower power consumption, and cost effectivness. The full potential of such applications cannot be realized unless these requirements are met.

Currently, there are no standard wireless solutions that satisfies the above metioned criteria. Existing solutions to wireless mobility managment include the IEEE802.11 Wireless Lan Standard, Bluetooth, and HomeRF. Although these solutions target different environments and offer varying services, they all provide wireless data communications and mobility management. They are also based on more sophisticated radios with spread-spectrum technologies increasing the cost and size of the solution. MobileIP and CellularIP are two solutions that manage mobility on an IP level abstracting away the physical level. These solutions although flexible and scalable, does not react quickly enough to the movements of the mobiles. There are currently no standard wireless solution that satisfies the above mentioned criteria.

By using a network of small simple radios coupled with more robust protocols, a simple class of wireless applications can realize their full benefit. These radios, provide bandwidths that are sufficent to simple applications. They are thin, and are no larger than a centimeter in length and width, making it easy to embed in small objects. Small batteries are sufficient to power the radios for prolonged periods of time since they transmit at less than 1mWatt.

1.2 The Problem

By using a network of small low power narrow band radios coupled with more robust protocols, a simple class of wireless applications can realize their full bennefit. These radios, provide bandwidths that are sufficent to simple aplications. They are thin, and are no larger than a centermeter in length and width, making it easy to embed in small objects. Small batteries are sufficient to power the radios for prolonged periods of time since they transmit at less than 1mWatt.

However, simplicity in a radio is a tradeoff for robustness. The unpredictable behavior of simple radios renders this network inviable without coupling them with protocols that make specific privisions to compensate for the lack of robustness. The signal quality of these radios are much more sensitive to changes in the physical environment than spread-spectrum radios. As distance between two transmitters increase, signal degradation can increase very quickly. Even while stationary, the proximity of a human can cause significant degeneration in signal quality. The protocols above the physical layer must be able to react quickly to changing conditions on the wireless channel and be able to optimize the network to conserve bandwidth while providing seamless communication to the roaming mobile hosts. These are the two issues that must be addressed in this thesis.

1.3 Approach

In order to provide seamless communication between the mobile hosts and the wired network, an infrastructure network topology is adopted. Infrastructure networks resemble classical cellular networks in that they consist of wired networks of access points(AP), each with a limited wireless coverage area that overlaps with that of another, to deliver service to the wireless nodes. Each node is bound to some AP and receives all incoming packets from it. Roaming between overlapping coverage areas requires a rebinding of the mobile to the AP of the new cell. This is handled by handoff algorithms that detect when a handoff is neccessary and execute the rebinding. Maintaining the connection between the wireless node and the wired network is the primary goal of these networks. If the network cannot effectively keep the mobile connected while it roams, the network has failed. Thus the handoff algorithm of this network is crucial to the success of the design.

The challenge to building such a network using simple single-frequency radios is that there is only one available channel. In existing infrastructure networks, radios can usually transmit in distinct channels. Neighboring cells do not transmit on the same channel thus, keeping interference between the cells relatively low. In the context of a network built with single-frequency radios, intra-cell interference is at 100% in the regions where cells overlap. Essentially if there is a mobile in the overlap of the two cells, it forces the two cells to share the same channel. No other mobile can transmit to either basestation if that mobile is transmitting. This dramatically reduces the throughput in both cells. The problem is exacerbated by the fact that each cell has multiple overlapping neighbors. How a mobile located between two cells is assigned to an AP can affect the resulting throughput of other cells in the area. Thus a handoff algorithm has the potential not only to lend reactivity to the simple radio's unpredictable wireless channel, but also to optimize the network for performance.

Existing handoff algorithms try to deal with these two issues in a simple approach. Based on the comparison of the current channel parameters to a calculated threshhold and channel parameters of other basestations, the decision to handoff is made. While this locally maximal approach may be sufficient for existing wireless LAN technology, the network under consideration needs to look at the problem of handoff globally as well. By dividing the handoff algorithm into two sub-algorithms, one local, the other global, we will enable the network to support the target applications. These algorithms will be embodied into a MAC level protocol.

1.4 Scope of Thesis

There are many aspects of the network protocol stack that need to be modified to ensure the optimal performance of a network of simple radios. Enabling the network to mantain optimal connection with the mobile hosts is the first step. It establishes a basis for future development of other networks of this type. Designing a MAC protocol that performs optimal handoff in this network presents insights into what kind of performance is possible. Future work in the areas of channel coding, forward error correction, and extending the complexity of the AP wired network topology will allow real-life applications to make use of this technology.

1.5 Contributions

This thesis will be making the following contributions :

- Present a two-tiered adaptive handoff algorithm.
- Present the design of a MAC protocol that embodies this adaptive handoff algorithm for a network of simple radios.
- A simulation and analysis of a wireless network using this protocol.
- Provide a bases for further development of the ideas and other aspects of the network protocol stack

1.6 Road Map

The following chapter introduces existing wireless network topologies. It also addresses the current wireless network solutions. Chapter 3 discusses the design of the handoff algorithm. Chapter 4 describes the details of a network including the MAC protocol that implements the handoff algorithm. Chapter 5 describes the simulation and the analysis of the network. Finally Chapter 6 concludes with discussion of conclusions and future work.

Chapter 2

Related Work

2.1 Infrastructure Networks

2.1.1 Topology

Infrastructure networks delivers connectivity to wireless nodes using a picocellular architecture based on traditional cellular networks. Traditionally, cellular networks have been the solution to providing wireless communication to users. The overwhelmingly popular application of the the wireless phone has deeply penetrated everyday life. As the user density increased, the original macro-cellular solution, comprised of cells on the scale of kilometers, can no longer support the resulting traffic levels. Shrinking the cell size increases the amount of bandwidth available to each user by decreasing the number of users in each cell.

Since indoor wireless networks typically have a high concentration of devices in a small confined space, many employ picocells as the way to deliver information to mobiles. Picocellular networks have cells in the scale of 10 meters, and thus can serve locations with very densely populated devices. The existing alternative topology is an ad hoc network. These networks are self-configuring without any preconfiguration. However, because all the components of an ad hoc network exist and move independently, it is hard to guarantee the service needed by many applications.

The components of a picocellular infrastructure network are :

- mobile hosts
- Acess Points or Base stations

- Wired network connecting the APs

- A gateway to the rest of the internet

Mobile hosts forward all outgoing packets to an access point and receives all incoming packets from the AP. The dependency on the AP emphasizes the importance to maintain adequate communication channels between the AP and the mobile at all times. How robust the channel is depends on the physical layer as well as the protocols used.

The access points each have a coverage area in which radios in can communicate with the access point. Usually, the coverage areas partially overlap to provide contiguous coverage within a physical volume. The coverage area, often referred to as a cell, is by no means symetrical as is often depicted in diagrams. Depending on the dimensions of the environment the cell may be extremely irregular in shape. As the cell size gets smaller, this phenomenon is more exaggerated.

The wired network links all the APs of the network and thus connect all the individual cells. The wired network maintains the state of mobile associations and delivers packets bound for nodes outside of each cell. If a packet destination is a mobile node inside the network, the packet is delivered to the AP to which the destination node is associated. If the packet is destined for a node outside of the current network, it is routed to a gateway.

The gateway is connected to other nodes on the internet. It routes packets bounded for its network into the network and routes packets bounded for nodes outside of the network out into the internet.

2.1.2 Mobility Management

Mobility management provides the network support to allow stations in wireless networks to roam without losing connectivity to the wired network. It has two central goals. One, the network must keep track of the current location of the mobile in order to correctly deliver the mobile's packets. Two, the network must change the mobile's association at the correct time when the mobile moves from the cell to another. The first goal can be achieved in either a central or a distributed manner. A central Regional Server can be designated to keeps track of each mobile's location and routes packets to the correct Access Point. An alternative is to use a distributed routing algorithm in which AP's discover mobiles and update their routing tables with the changing state of the network. The second goal is achieved with handoff algorithms which changes the association of the mobile from AP to AP.

The goal of an effective handoff scheme is to provide continually acceptable link quality to the mobile host (MH) during the transition between neighboring cells. There are a number of challenges in achieving this goal. One such challenge is minimizing the delay of completing a handover. If the process takes too long, the MH may experience a significant drop in link quality. Another challenge is minimizing the frequency of handoffs while still remaining responsive enough to guarantee link quality. Since each handoff requires extra signaling overhead, too many unnecessary handoffs will cause unwanted network overhead.

There are two general classes of handoff algorithms : intracellular and intercellular. Intracellular handoffs change the channel that the mobile uses to communicate within its current cell to improve the link quality. Intercellular handoffs change the actual Access Point association of the mobile to maintain link quality when the mobile has moved out of the cell. While intercellular handoffs are performed in all cell-based communications, intracellular handoffs are not. The discussion foceses on the former.

Types of intercellular handoffs

Traditionally, there are 3 ways to implement intercellular handoffs :

- Network controlled handoff (NCHO)
- Mobile Assisted handoff (MAHO)
- Mobile Controlled handoff (MCHO)

In NCHO, the MH is completely passive. The mobile switching center (MSC) requests alternative measurements from surrounding basestations and orders intercell handovers. Intracell, or frequency handovers are not possible in this scheme. Handoff delay is on the scale of many seconds.

In MAHO, the MH makes measurements for the alternative APs and transmits them to the current AP periodically. The final decision to initiate handoff is still the decision of the fixed network. Handoff delay is on the scale of one second.

In MCHO, the MH makes its own measurements as well as receive measurements from the BS. The decision to handoff is made by the mobile. The handoff delay is on the scale of hundreds of milliseconds[2].

Handoff Execution

There are also two general mechanisms for handoff executions. A hard handoff is one which the connection with a previous AP is completely broken before the MH connects with the target AP. This method requires less coordination between the APs and the mobile, but may also lose packets if the new AP does not begin forwarding packets to the mobile after the connection with the old AP is broken. In a soft handoff, the MH remains connected to the source as well as the target BS while handoff is performed. This method usually avoids any loss of packets during handoff, but requires more coordination, and duplicate packets may be sent to the mobile during the transition.

Evaluating Handoff

The performance of a handoff algorithm can be measured with the following criteria:

- Probability of unneccessary handff
- Rate of handoff
- Duration of handoff
- Cell Drag

Probability of an unnecessary handoff refers to the probablility that a handoff is executed, even though channel conditions are still sufficient for communication.

The *rate of handoff* refers to the number of times handoff is executed in a set time. The higher the frequency, the more unstable the system is. An unstable system is undesirable since the cost of bandwidth overhead can cut into the bandwidth of both the wirelss and the wired network.

Duration of handoff is the time between handoff initiation and handoff completion. Initiation refers to when the network decides a handoff is needed. Handoff completion refers to when the new AP begins to deliver packets. A large handoff duration decreases the responsiveness of the newtwork to changes in the wireless channels.

Cell drag refers to how far into another cell a mobile enters before an handoff is initialized. High cell drag is an indication that the algorithm is not responding to changes in the channel in due time. The channel will then be allowed to degrade to a point where it may no longer be able to maintain acceptable link quality.

Handoff Algorithms

Basic handoff algorithms attempt to assess the condition of the channel through various metrics of the wireless channel and makes a decison based on its assessment. Channel metrics usually include signal strength, Bit Error Rates, and Packet Error Rate with signal strength being the most widely used. There are multiple approaches to interpreting the measurements to determine whether a handoff is required.

The simplest strategy is *relative signal strength*. This approach compares an averaged measurement of the signal in the home cell with that of a neighboring cell. Handover is initiated when the measurements of the neighboring cell exceeds that of the home cell. Because wireless channel conditions may vacillate quickly, this approach causes many unnecessary handoffs.

Relative Signal Strength with threshold decreases the number of unnecessary handoffs by preventing handoff from occuring unless the the home cell channel is below a threshold TH and the neighboring channel signal is better.

Hysteresis margins require that the neighbor cell signal must be stronger than the home cell signal by a hysterisis margin H before handoff is initiated. The margin prevents unnecessary handoffs from occuring when the channel conditions in the two cells are similar but vary by small quantities with time. This method has also been combined with the use of thresholds to further decrease handoff frequency. Depending on the network, this combination of strategies may reduce handoff frequency to the point where handoff delay increases to undesirable levels.

Predictive techniques have been shown in simulation to reduce unnecessary handoffs by making decisions based on the expected channel conditions derived from past history and signal strength patterns. This method works well in environments where the environment exhibits regular patterns such as an inner city street with regular perpendicular intersections. However, in less predictable indoor invironments the approach may work just as well as one that does not perform any predictions.

Other techniques employ alternative technologies such as neural nets, fuzzy logic, hypothesis testing and dynamic programming. Although these approaches have exhibitted some limitted success, they are not widely deployed partially because the burden on computing resources is too heavy for a realistic implemenatation. Never the less, they provide interesting insights on the problem handoff control.

All of the existing handoff strategies approach handoff as a local problem. The problem they solve involves two overlapping cells with one mobile located between the two. Because most of these strategies assume that each cell transmits at a different frequency, handoff can be approached this way. However in a network where cells transmit at the same frequency, the effects of a handoff decision at one end of a cell can affect the performance of mobiles at seemingly unrelated areas of the network due to having only one shared channel throughout the network. Thus, a handoff algorithm using a global approach is more appropriate for such a network.

2.2 Wireless Links

2.2.1 IR

Infrared communications provide low cost, short range wireless data communication through light in the infrared frequency band. This technology can be found in laptops, PDAs, cell phones and pagers, to name a few. The current IR standard, IRDA, can achieve bandwidths from 9600bps to 4Mbps. Since the technology is based on light waves, IR cannot pennetrate through solid objects as radio frequencies can. Thus a direct line of sight is a requirement for data exchange. IR also has a very limited angle of operation, making connections possible only if two nodes are pointing at each other. Diffuse IR, an alternative IR technology, diffuses the light throughout a space eliminating direct line of sight as a criteria for connection. IR technology is ideal for short range peer-to-peer communications such as connecting a lap top to the printer at a cost as low as $1-$2. However, the inability to penetrate through solid objects such as walls limits its applicability to a wide range of applications.

2.2.2 Narrow Band RF

Narrow band radios transmit data on a single frequency. The radio frequencies can penetrate through walls and other solid objects with the exception of heavy walls such as steel or concrete. Narrow band's range is higher than that of IR and is omni-directional. The exact range of the transmission is inversely proportional to the frequency and proportional to the power of the transmitted signal. At 900MHz, a low power radio transmitting at 1mW can achieve a range of ten meters in dense cubicle environments. Narrow band communications is a more flexible wireless link solution than IR.

However, narrow band wireless channels are very susceptable to interference. If any other signal at that frequency is combined with the data, the additive property of RF will corrupt the data trasmission. Narrow-band also suffers from multi-path fading, a destructive interference caused by the geometry of the environment. Multiple copies of the same signal can take different pathes to reach the same receiver at which the superposition of these out of phase copies can cancel each other out completely. This is a highly directional and environmentally dependent phenomenon and usually affects one direction of a link.

Low power narrow-band wireless solutions are low cost and offer adequate bandwidth and range for indoor data communications. Although it is susceptable to interference, protocols in the higher layer may be able to compensate for the lack of robustness.

2.2.3 Spread Spectrum RF

Unlike narrow band which concentrates energy around a single frequency, spread spectrum is a modulation technique that distributes the signal over a frequency band. It decreases the average engergy at any frequency, but increases redundancy to create a more robust channel. At the receiver, the signal is despread with the same technique as it was spread, to extract the original baseband signal. Many existing wireless standards use spread spectrum radios in the 2.4GHz ISM frequency band which does not need to be licensed.

There are two primary types of spread spectrum techniques, Direct Sequence(DHSS) and Frequency Hopping(FHSS). Direct sequence modulates the narrow band signal with a psuedo-noise digital signal, effectively spreading the original signal throughout the entire frequency band. At any given frequency, this signal appears as low powered noise. At the receiver, the same psuedo-noise digital signal is modulated with noise-like signal to recover the original baseband. Frequency Hopping radios employs a psudo-noise code to control the periodic shift of the frequency at which it is transmitting. By decreasing the amount of time spent transmitting at any frequency, the probability and the impact of interference are greatly reduced. The receiver shifts its receiving frequency determined by the same psuedo-noise code.

Both types of spread spectrum techniques have been shown to be robust in the presence of transient and even constant noise. Multipath fading, the common problem that plagues narrow band transmissions is frequency dependent. Since spread spectrum does not depend on any single frequency, the problem is a non-issue. Jamming is also a non-issue for the same reason. However, due to the inherent complexity of spread spectrum radios, the cost of these radios are currently mostly over $100. They are usually available in a PCMCIA card form factor.

2.3 Current Wireless Network Solutions

Currently there are many efforts to standardize the "air interface" of wireless networks. Some of the most prominent ones have been the IEEE 802.11, Bluetooth, and HomeRF. These networks base their protocols on spread spectrum radios that are usually costly, more power consuming and larger in size than non spread spectrum.

2.3.1 IEEE 802.11

Although IEEE 802.11 compliant solutions are reasonable for heavy duty applications such as the lap top, they do not presently meet the price and form factor criteria for simple applications. IEEE 802.11 is a 2.4GHz wireless LAN standard finalized since 1997. The specification supports either Ad Hoc or Infrastructure networks. Infrastructure networks are picocellular networks that have been discussed in detail in the previous section. The Ad Hoc network architecture supports spontaneous peer-to-peer communications in the absence of structured network support.

The standard addresses the physical layer and the media access control layer. The physical layer supports 3 types of physical links : diffused infrared, direct sequence spread spectrum and frequency hopping spread spectrum. Diffused IR operates in the baseband while the two RF based links operate at the 2.4GHz or ISM frequency band.

The MAC layer provides a number of services include data transfer, association, reassociation, authentication, privacy and power management. Data transfer provides access to the shared medium by using Channel Sense Multiple Access with Collision Avoidance (CSMA/CA). Radios can enter a power-saving mode when it is not actively transmitting and receiving data.

2.3.2 HomeRF's SWAP Specification

HomeRF's Shared Wireless Access Protocol (SWAP) is a home networking specification that provides a standard method of connecting all the devices of the home with wireless voice and data links. Wireless voice links are supported by the TDMA service via a central Control Point. Data link service is provided by an ad hoc network with an internet gateway. Media access is controlled by the CSMA/CA algorithm.

All data travels over frequency hopping spread spectrum radios that operates at the 2.4GHz ISM band with a 50 meter transmission range. Data rates can reach up to 2Mbps depending on the method of modulation. To support the transmission distance which usually covers the typical home and yard, the transmitter consumes 100-milliWatt of power. Another drain on power is the use of CSMA/CA for data transmission. It is difficult for simple applications to sustain such power consumptions for long periods of time yielding this wireless solution impractical for very low-cost or minitiature devices as well.

2.3.3 Bluetooth

Bluetooth is a wireless standard that aims to use a low-cost, short range radio built into a 9mmX9mm microchip. Bluetooth has three primary goals. First it replaces the many proprietary cables that connect devices with one standardized wireless link. The wide range of applications include laptops, printers, PDA's, fax machines, keyboards and mice. Second, it provides a way to bridge existing data networks such as a wireless and a wired network. Finally, it allows devices to form ad hoc networks when structured networks are not available. Bluetooth also privides provisions for security in the link level supporting both authentication and encription.

The radio link operates at the 2.4GHz using a Frequency Hopping Spread Spectrum radio. Bluetooth also specifies three different type of error correction schemes : 1/3 rate FEC, 2/3 rate FEC, and automatic repeat request (ARQ). The combination of FEC and fequency hopping makes the link robust in a noisy environment. The radio can transmit at a gross data rate of up to 1Mbp/s to a distance of between 10 centimeters to 10 meters. First generation Bluetooth radios cost around $20.

2.4 Mobility in the IP Layer

MobileIP and CellularIP are solutions that manage mobility in the network layer independent of the physical layer. However, these protocols lack the speed needed to accomodate the quickly changing conditions of the mobiles due the inherent delay involved in routing.

2.4.1 MobileIP

Mobile IP (MIP) provides mobility management at the IP layer[4]. Hosts can roam in local and wide area networks in an IP environment. The advantage of Mobile IP is that roaming between different types of networks is transparent to the application. The IP address is static throughout migration, and packets are redirected from the mobile's home network to its current network through MIP's tunneling mechanism. This solution, although flexible and scalable for macro movements between networks, is not suitable for "mirco" mobility, the movements within a network. This is true especially in an picocellular network. Roaming between picocells require fast handoffs which are hard to achieve when using tunnelling each time a mobile changes an AP. Movement within picocells in the same network can be handled by the lower layer protocols much more efficiently than by Mobile IP.

2.4.2 CellularIP

CellularIP combines the flexibility of MobileIP with the efficient location management of cellular networks.[5] The user can roam between networks as well as within a network wirelessly. Inter-network roaming is enabled using MobileIP tunneling. Intra-network roaming is enabled using a network of routing Access Point nodes. Each node acts as a router which maintains routing information for mobiles who's packets travel through the node on its way to the network's gateway. When the network executes a mobile handoff, the routes in the network of APs must be changed in order for the mobile's packets to reach its new Access Point. The handoff duration, defined to be the time between the handoff to the time the mobile receives its first packet from its new AP is bounded by the round trip time between the mobile and the gateway of the AP network. The delay is still inappropriately high for picocellular handoffs.

2.5 Summary

Through this discussion of related work in the area of wireless mobility management, it is apparent that an addequate solution does not exist for simple applications at this point. Currently, existing wireless solutions either lack the correct physical characteristics or lack the speed necessary to carry out fast handoffs. In proposing a network of small low power radios, an adaptive handoff algorithm necessary to make this possible.

Chapter 3

Adaptive Handoff

Adaptability is the key to enabling robust wireless communication. In this chapter, we present a handoff algorithm that adaptively makes decisions based on the condition of the network. The algorithm approaches handoff not only as a localized problem, but as a global problem as well. In the following sections we first present the general idea of a two-tiered adaptive handoff. Then, the algorithmic details of each tier will be discussed in detail.

3.1 Algorithm Structure

In general, handoff is used to improve the channel quality of each mobile using local knowledge. Although the greedy approach used in many of today's cellular systems provides good performance, there are situations where the network can be further optimized given information about other areas of the network that is not available locally. This global approach to handoff may generate handoffs that will not necessarily result in an immediate performance improvement, but by doing so, performance for the overall network is improved. Whether the performance of individual mobiles should be constrained in the optimization process is a QoS issue. If the system does need to provide some minimal service guarantees, these constraints can be applied to the optimizer, and handoffs are made within these constraints.

This new approach to handoff management is implemented with a two-tiered handoff algorithm. The lower tier performs connectivity handoffs. It maintains sufficient connectivity to each mobile, moving mobiles to other cells when the channel metrics fall below a minimal threshold. The upper tier performs optimization handoffs. It optimizes a specific aspect of network performance. Whether this is aggregate throughput, average throughput, error rate, or delay is dependent on the individual network. The algorithm modularizes the optimization requirement such that the optimization goal can be easily changed.

The goal of each tier is fundamentally different; thus each tier bases its decisions on a different time scale. Connectivity handoff is initialized when the link quality begins to degrade. In indoor picocellular environments, link degradation occurs very quickly due to the small size of the cells and the complexity of the environment. Therefore, the connectivity handoff must be very fast. Thus, it operates at a very short time scale, looking at the performance of the link over a short amount of time. Optimizing handoffs are initialized when the network is sure that by moving the mobile from one cell to another, there will be an improvement in overall performance. Using long-term averaged measurements provides the network with a more solid picture of link quality than short-term measurements. Short term measurements are much more sensitive to the transient changes in the channel that does not reflect the overall quality of the link. Therefore, the higher tier of the algorithm runs on a slower time scale than the lower tier.

The adaptability of this algorithm is in its ability to optmize when the network is stable and to maintain connectivity when the mobile is moving. While connectivity handoffs are enabled at all times, optimizing handoffs should not be enabled when the network is unstable. Only when a stability trigger detects that most of the network is stable (most of the mobiles are not migrating between cells) will it enable optimization handoffs. Since the link conditions are constantly changing when the mobiles are migrating between cells, any optimization attempts will result in wasted resources. Thus overall, when the network is stable, optimization handoffs are performed. When the network is unstable, or becomes unstable, optimization handoffs are diabled while connectivity handoffs handles all migrations.

3.2 Connectivity

The goal of this algorithm is to maintain minimal connectivity between the mobile and its AP with a good level of reactivity while not sacreficing stability. Minimal connectivity is the minimum link quality that is sufficient to sustain a wireless application. How sufficient link quality should be defined is specific to the network and to the applications that are on it. The link quality can be defined using various metrics available to the hardware and the protocol stack including signal stregnth, bit error rate, packet error rate (PER), throughput, and traffic load. With some applications, channel performance considered sufficient to sustain application function varies. For now, we assume a uniform link quality requirement, however, this algorithm is applicable for specific Qualities of Service (Qos).

Reactivity and stability are two characteristics of handoff algorithms that are tradeoffs of each other. Reactivity refers to how quickly the network detects that handoff is needed. Low reactivity causes handoff delay which allows the mobile to move into the neighboring cell while still bounded to the home AP. Too much reactivity causes instability in the system. Thus maintaining relative stability eliminates unnecessary handoffs. The connectivity handoff balances these two characterstics.

The criteria for handoff is charactrized by the following:

IF ($M_{x,i}$ < TH)
    IF ($M_{x,j} > M_{x,i}$ + H)
    INITIATE_HANDOFF

The algorithm uses a channel metric, M and applies it to a criterion that combines the ideas of relative strength, absolute threshold, and hysteresis. $M_{x,i}$ is a short term time averaged measurement of the uplink channel from a mobile x to an AP i. Channel metrics from other APs are not considered unless M at the home AP falls below some minimum level TH. The target AP must have a channel with better metrics than the current AP, but it must also be better than the current channel by an improvement of H. Only then will the handoff take place. Using an absolute threshold TH ensures that the mobile is always bound to an AP that can offer minimal channel conditions greater than TH. Short-term measurements have very little history. This causes the algorithm to react quickly to recent changes in the channel. Since absolute threshold and short-term measurements tend to cause many unecessary handoffs when a mobile is between two cells that have similar but still varying channel qualities, adding a hysterisis margin curbs the number of uncessary handoffs.

3.3 Optimization

The Optimization algorithm uses handoff to maximize aspects of the network's performance. These include agreggate throughput, mobile throughput, packet error rates, average delay, etc. This algorithm deals with optimizations that cannot be achieved without a global perspective of the network.

Optimizing networks with a global perspective is a complex problem. To illustrate this, consider the task of optimizing the network for aggregate throughput. In the context of the proposed single frequency picocellular network, optimal aggregate througput is achieved when at any momment, all access points are transfering data to or from mobiles in the network. The single frequency restriction makes this a challenging goal. If there is a mobile M situated on the overlap between two cells, no other source in either cell can transmit when M transmits or when M's AP is transmitting to it. Therefore, for each mobile situated on cell overlaps, the throughput of the two cell can be dramatically reduced.

However, if the handover algorithm can look at the network as a whole, which cell the mobile is partitioned to can increase number of APs that can communicate in parallel in the network. Take a simple example of a network with three overlapping cells, A, B, C, in a line. Assume a mobile is located at each overlap (m_1, m_2). If m_1 is assigned to cell A and m_2 is assigned to cell B, then when m_1 transmits, m_2 can't transmit and vice versa. However, if m_1 is assigned to A and m_2 is assigned to C, then the two mobiles can communicate with their respective APs in parallel, which increased the throughput of the network by two. In larger networks, this simple problem is expanded into a complex optimization problem.

3.3.1 Tree Search

A tree search can solve the handoff optimization problem by searching all the possible combinations of handoff decisions that can be made throughout the network. Optimization problems can also be solved with dynamic programming with better performance. However, dynamic programming is not appropriate for this problem because it solves problems that display optimial subproblem structure. An optimal configuration may not be optimal if

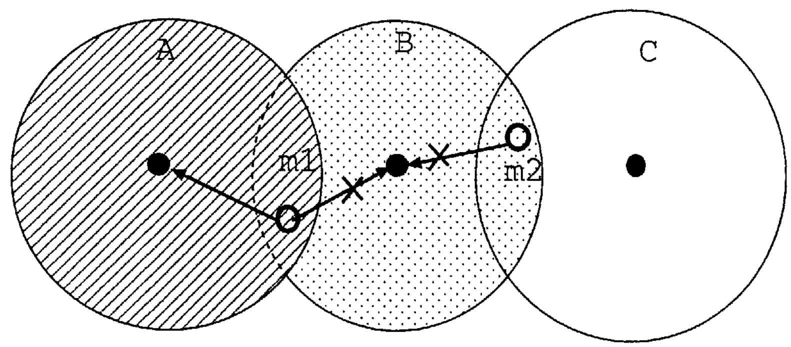

Figure 3-1: Inefficient mobile partition

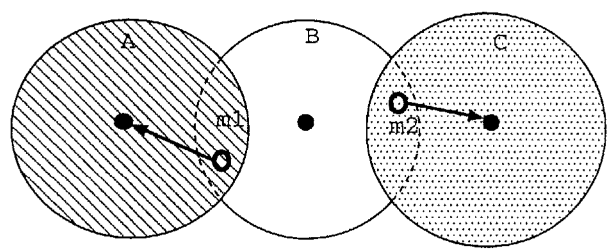

Figure 3-2: Efficient partition with increased agregate throughput

an extra mobile is added to the system. The addition of the new mobile may cause the previously partitioned mobiles to perform poorly. Thus the new optimal partition requires the optimal subproblem to be reconfigured.

It is important to note that tree search algorithms are difficult to implement as distributed algorithm. Although searching a given tree can be done collaboratively among several processes, building a tree and searching it distributedly proves to be difficult. Handoff algorithms are usually discussed in a distributed context. However, to simplify our problem, this discussion assumes a centralized algorithm which resides on a node in the wired backbone of a wireless network.

In order to build the tree the algorithm first creates a database that represents the current state of the network. It does so by using the periodic updates that each AP sends to it. These updates contain :

- Mobiles that the AP can hear
- Which mobiles are bound to the AP
- The AP's assessment of link quality with each mobile that it hears

- Whether the AP has been involved in a handoff since the last update

From this all relative information about the network can be derived. Mobiles that lie between cells can be identified by finding the intersection of mobiles that each cell can hear. Each mobile's current AP is directly reported in the update. Each mobile's alternative links and their qualities are also compiled. The resulting database of network statistics is time averaged until the algorithm is triggered. If in the process of collecting data and waiting for the stability trigger, the network becomes unstable by way of an increased number of moving mobiles, the database is cleared. This way, optimization only uses data collected while the network is in its current configuration.

The tree's representation is as follows. Each level of the tree expands the branches by nodes that represent the possible partion choices of one mobile. Mobiles that are not located on a cell overlap, are still included in the tree, but only add a single branch to each existing one.

The tree building process is constrained by additional requirements that the network enforces on each partition. Using constraints prevents the optimizer from wasting network resources on executing a handoff that significantly degrades a link metric that is important to the network. Therefore, before adding each node to a branch, the partition is passed to a constraint module which decides if making this partition violates any constrains of the network. For example, while the algorithm is optimizing aggregate throughput, it can also add constraints on error rate to prevent the optimizer from passing the node into a cell that will only cause a conectivity handoff back to the original cell. After a constraint violation is detected, that branch is killed and no further branches will expand from it. This adds extra dimensions to the optimization process which can consequently support QoS in the network. It can also potentially reduce the number of total partitions that need to be evaluated at the end.

A partition evaluation module evaluates each path of the tree and gives it a reward value. This value depends on the predicted performance from all the partitions. The number of handoffs that need to be made discounts this reward value since handovers require the use of extra network resources.

When the algorithm is enabled by the stability trigger, it builds a tree based on the data in the database, collected from the nework. The following is the psudocode for the tree building algorithm :

```
OPTIMIZE-PARTITION(database){
    tree
    if (stable (database))
       cell = pickRandom(database) //start by picking a appropriate cell
     TREEBUILD(cell, 0) // tree build depth first
}

TREEBUILD (cell, level)
    mark(cell) // notes that the algorithm has inspected the cell
    foreach inter ∈ cell.INTERSECTIONS
        foreach mobile ∈ inter.MOBILES
            if unmarked(mobile)
            foreach partition ∈ unmarked(mobile.PARTITIONS)
                if pass_constraint(partition)
                    tree.add(level, partition);
                    if partition.target ≠ mobile.AP
                        partition.penalty = partition.parentPenalty + HANDOFF_PENALTY

    if hasMoreNeighbors(cell)
        TREEBUILD (getUnmarkedNeighbor(cell), level + 1)
    else if hasMoreCells()
        TREEBUILD (getUnmarkedCell(cell), level + 1)
    else
        foreach leaf ∈ leaves(tree)
            leaf_value = evaluate_performance(leaf);
            if leaf_value > curmax
                curmax = leaf_value
                curPartition = leaf

        return curPartition
```

The algorithm begins by selecting an intersection between two cells. For each of the nodes that lie in this intersection, the tree is expanded by a level, with each node representing a possible partition. After all nodes in the intersection are added, the algorithm continues the same process with all other intersections in the cell. Once all intersections in the cell have been processed, then the algorithm repeats the same proces for the current cell's neighbor. If all neighbors have been examined, then a random cell will be selected. Before each partition is added to the tree, a constraint subroutine is consulted. If the partition causes the current network configuration to violate the constraints, then nodes that represent this partition decision will not be added. The reason that neighbors of cells are looked at first before random cells is because if there is any combination of local partitions that do not satisfy the system constraints, it will be quickly discovered. The related branch will be terminated quickly. Randomly processing cells will allow these branches to expand exponentially for a long time before it is discovered that the resulting configuration does not work.

The stability trigger enables optimization handoffs only if the network is stable. In a network where a majority of the mobiles are moving, connectivity handoffs will dominate, quickly cancelling any optimization handoffs that may have occured. To measure the stability of the system, the stability metric is defined to be the standard deviation of the channel metrics. The larger the data set is, the more representative the standard deviation is of the true conditions of the channel. Therefore, the stabilty trigger should only be activated when the stability metric is smaller than a threshold STh and has a set of data spanning a time greater than $STime$.

3.4 Summary

The handoff algorithm discussed in this chapter adaptively maintains sufficient link quality and optimizes the configuration of the network. To prevent the link quality from falling below an acceptable level, connectivity handoffs quickly switch APs for a mobile who's short-term averaged channel measurements fall below a given threshold. When the network is determined to be stable using the standard deviation of channel quality, optimization handoffs adjust the network configuration to optmize for secondary network parameters such as aggregate throughput. By adding extra flexibility in the handoff algorithm, the overall network performance is improved.

Chapter 4

Network Design

Chapter 3 discussed an adaptive dual-tier handoff algorithm. In order to demonstrate the viability of such an algorithm, a network using low power narrow band radios is designed. This chapter discusses the design of such a network.

4.1 Network Topology

The network assumes an infrastructure topology as discussed in Chapter 2. It is designed to operate in an indoor environment where there are small offices along large corridors. The access points are strategically placed in this environment such that each cell overlaps with other cells to allow continuous connection while moving between cells. The APs are joined by an Ethernet LAN to form the wired backbone. A mobile node forwards all outgoing packets to and receives all incoming packets from its home AP. Much of the functionality of mobility management is placed in the AP's to decrease the computation load on the MH. This is motivated by the fact that mobile applications are limited by size, power, and cost, so the processing power available to them is much more limited than that of the APs.

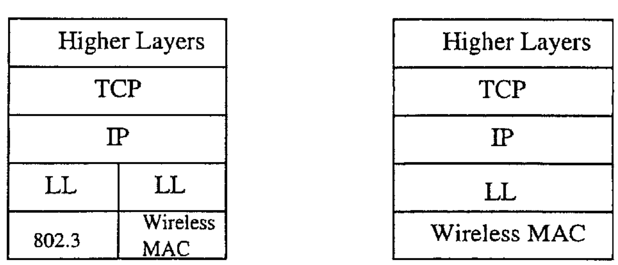

Figure 4-1: Protocol Stack of AP and Mobile

The nodes of the network have a layered protocol structure. The application lies on top of the TCP/IP protocol suite. The universality of TCP/IP allows almost any type of networked application to run over this network. Below the TCP and IP levels, is the link layer which manages the pipe between the mobile and the AP. Finally, the Medium Access Control (MAC) layer manages the sharing of the common transmission medium. The AP has two network interfaces, one for the wireless network and one for the Ethernet LAN.

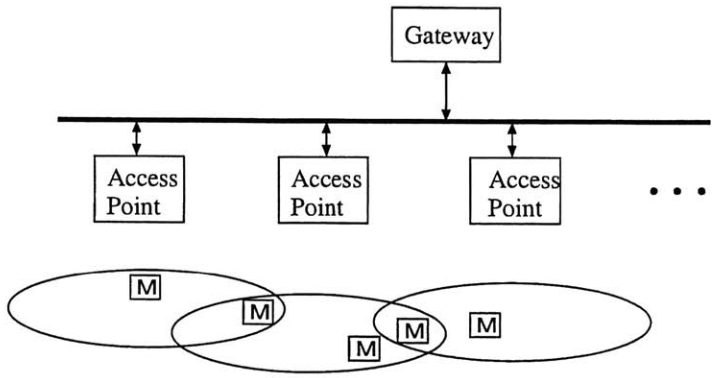

Figure 4-2: Picocellular network with AP Ethernet LAN

The focus of this design is in the implementation of the handoff algorithm in this network. Placing most of the functionality at the MAC layer allows the fastest reaction to changing conditions in the network. Therefore, the discussion that follows will be focussed on the MAC layer of the protocol stack.

4.2 The Wireless Link

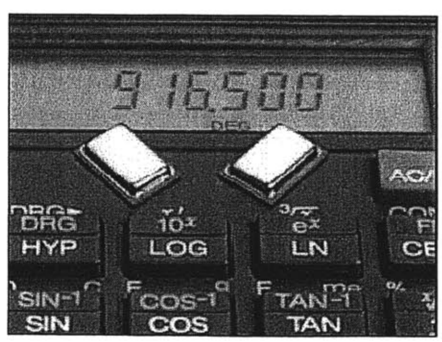

Figure 4-3: RFM ASH Trasceiver

This network employs an off the shelf, SAW-based hybrid transceiver produced by RF

Monolithics, Inc (Figure 4-3)[1] . This low cost, low power, surface-acoustic-wave (SAW) based hybrid radio transceiver has a transmission range of 10 meters while transmitting at 1 mWatt on the 916.5 MHz frequency. The receiver architecture (Figure 4-4)[2] provides stability and great out-of-band rejection. It offers a maximum bandwidth of 115kb/sec.

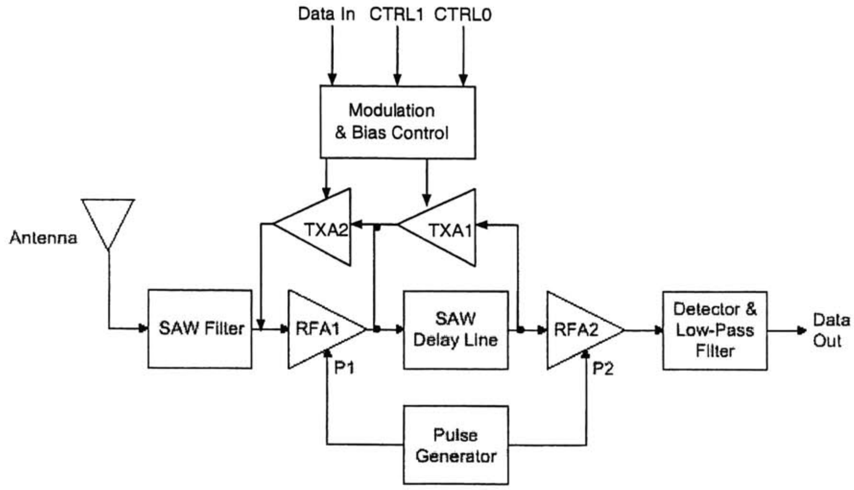

Figure 4-4: RFM ASH Trasceiver Block Diagram

4.3 MAC protocol

The MAC protocol controls the access to the shared transmision medium of the wireless network. Without a MAC protocol to mediate the channel, multiple transmissions will collide causing the information to be lost. Time Division Multiple Access (TDMA) and Channel Sensing Multiple Access with collision avoidance (CSMA/CA) are the two predominant methods for medium access control. TDMA divides a single channel into several time slots. A transmitter can transmit only in the slot it is assigned, thus effectively preventing collisions. Nodes in a CSMA/CA network listens to the channel and only transmit when the channel is idle. They may avoid collisions by using an exchange of Request To Send (RTS) and Clear To Send (CTS) packets before each transmission. If in the case that there is a collision, retransmission time is determined with exponential backoff. While TDMA does have a larger latency than CSMA/CA, it delegates control of the medium to a centralized source minimizing the computation expended in the resource limitted mobiles. For this

[1]Courtesy of RF Monolithics, Inc.
[2]Courtesy of RF Monolithics, Inc.

reason and the fact that it is a simpler solution to implement, the proposed network will employ a MAC protocol based on the concept of TDMA.

The MAC protocol on the AP and the one on the mobile have different tasks to fulfill. On the mobiles, the MAC protocol listens for the downstream control messages, or beacons, from the AP and then transmits packets during its assigned slots. The AP assigns the available slots to the mobiles that are bound to it, synchronizes time with other APs and manages handoff. In order to offer these services, the MAC protocol in the wireless interface communicates with other AP MAC entities through the Ethernet interface.

4.3.1 Frame Structure

The channel is divided into frames each of length L_f, and each frame is then divided into n_s equal length slots. In the beginning of each frame, the first slot is used by the AP to broadcast signaling information to its mobiles. This signaling information constrains a schedule of slot assignments in this frame and also serves as an AP beacon. The last slot of the frame is a contention based slot where any mobile can send control messages to the AP on a contention basis. The slots in between are data slots that only specific mobiles can try to access as dictated by the schedule. The first half of each data slot is used for uplink and the second half is used for the downlink. At the beginning of each slot, a small percentage of the total slot length, t_c is used for contention resolution. It is divided into three periods, the $Control\_Idle$, $Data\_Idle_1$ $Data\_Idle_2$ respectively.

4.3.2 Contention Resolution

Contention resolution prevents scheduling conflicts from disrupting the operation of the network. Scheduling conflicts occur when two transmitting mobiles are assigned the same slot thus jamming each other's signal. In this proposed single channel picocellular network, the downlink and uplink control messages each occupy a slot and therefore are susceptable to scheduling conflicts. Contention resolution ensures that control slots will always be available to control messages.

Scheduling conflict can occur in the beacon if all the frames in the network are synchronized. Each AP will transmit a beacon at the same time causing beacon collisions at the cell overlap regions. Any mobile in that region will not be able to receive the frame schedule and thus cannot transmit. To remedy this, the start of a frame at each AP is offset by a random number of slots to eliminate frame synchronization. Although there is a chance that two neighbors will still have the same slots, these rare occasions can be corrected by messaging through the uplink control slot.

Scheduling conflicts can also occur between the control slots and the data slots. If a mobile on a cell overlap is assigned the same slot as its neighbor's beacon slot, then the mobile will jam the beacon transmission at every frame. If another mobile is assigned the uplink control slot, any attempt for mobiles to correct the beacon collision will itself be jammed. It is necessary to establish the priority of control transmissions over data transmissions. This is achieved by introducing assymetry within each time slot.

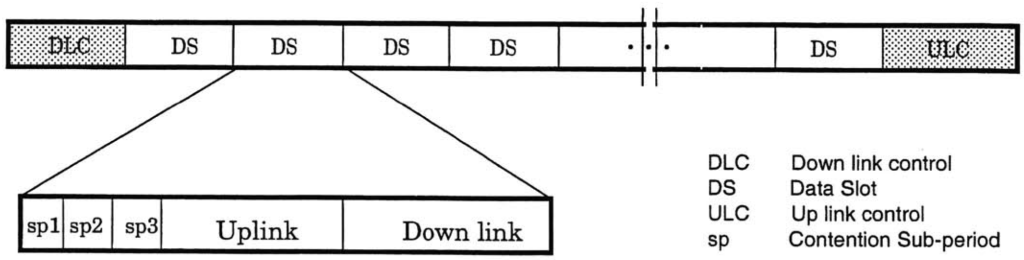

Figure 4-5: Frame and slot structure

In the beginning of each slot, there is a small period used for contention resolution. This period is divided into three sub-periods. To send a beacon packet, the AP waits for the channel to be idle for the duration of the first sub-period and if it does not detect any transmissions, it sends the beacon. Mobiles transmitting uplink control messages also follow the same procedure before transmitting in the uplink control slot. Other mobiles wait either two or three sub-periods before sending data packets. A randomizer in the mobile decides which length the mobile should wait at the begining of frame. If a mobile M waited two sub-periods, and some source still transmitted before the M did, the this source must be transmitting a beacon or control slot. The mobile notifies its AP with this information to be rescheduled since the it will never have a chance to transmit in this slot.

However, this scheme does not prevent collisions caused by asymetric channels and the hidden node problem. Asymetric channels occur due to multipath fading such that a mobile cannot hear the transmitter, but the transmitter can hear it. The hidden node problem occurs when two mobiles who are out of range of each other transmit to a node in between them. Both these problems are solvable by using RTS and CTS packets. However, this approach would add too much delay on top of a scheme that already has moderate delay. Instead, we let retransmission recover these collisions. ACKS embedded in the beacon indicate whether the last packet was received correctly and the mobile will know to retransmit if necessary. An additional benefit of this scheme is that when a mobile moves from one cell to another, even though it has the same slot assignment as a mobile in the new cell, both mobiles can still transmit correctly if they are with in hearing range of each other. It also allows the AP to assign multiple mobiles to the same slot when the cell is loaded with more mobiles than slots.

4.3.3 Slot Assignment

If the number of mobiles in a cell A, N_A, is less than n_s, then some mobiles may be assigned more than one slot, according to past performance history. This will ensure that the bandwidth is used up almost 100% of the time. If N_A is greater than n_s then each slot may have more than one mobile assigned to it. The contention resolution mechanism in each slot should allow each mobile to continue transmitting, but at a much higher delay. Therefore, the scheduler pairs mobiles who's history indicates low traffic load and adjusts pairings as traffic patterns change.

4.3.4 Synchronization

Since correct operation of a TDMA network is heavily dependent on timing, synchronization is a necessity. Synchronization can be divided into two scopes. One is the synchronization within a cell, between the AP and its mobiles. The other is the synchronization throughout the network, between all the APs. Within the cell, synchronization can be established through the downstream control beacon. Each mobile compares the time stamp on the beacon to its own local time and if they are different by greater than the latency from AP to mobile, it adjusts itself to the beacon time.

To keep time synchronized between the AP's, a similar approach is also used. Every $t_{backend}$ seconds, each AP sends a message to the backend, *Backend_Beacon* containing its current time. To prevent all all the APs from sending beacon at the same time, each transmission time is offset according to the MAC address of the AP. If an AP receives a backend beacon with a time stamp higher than its own, it sets the local time forward. Since backend beacons are sent often, the amount of clock skew that could happen between each backend beacons, is very low, so the amount adjusted should be very small. When a mobile enters another cell, it should still be transmitting within the confines of a slot. If for some reason, there is a large difference, the current frame will be abandoned and the next frame will start according to the new adjusted time.

Initializing the network requires extra provisions for synchronization. When the protocol is first initialized in the AP, it listens for backend beacons for two frame periods. If this is the first AP in the network to be initiated, it will not hear other backend beacons. Once time is up, it sets its own time to be zero and sends out a backend beacon. Else, it hears a backend beacon and sets its time according to the timestamp. If because of delay and loss, the AP just failed to hear other AP's in the network, then the provisions for resynchronization as discussed above is used.

4.4 Mobility Management

Mobility management allows mobiles to move throughout the coverage of the network and still be connected to the network. There are two aspects of mobility management. First the mobile must be allowed to roam the network. This is enabled by handoff. Secondly, once the mobile moves from one cell to another, the network must know to deliver packets for the mobile to the new AP. This is managed by routing.

4.4.1 Routing

The IP layer of the AP keeps track of the location of a mobile in the network with its routing table. The IP addressing topology is flat in this network so all mobiles and AP's have the same network address. Assigning subnet addresses does not add any advantage since the AP's cannot use hiearchical routing. Whether the packet is addressed to a mobile that is bound to the AP is the only factor that determines which network interface the packet should be forwarded to.

The routing table keeps a one-to-one mapping of each mobile that is bound to the AP and maps them to the wireless interface. For packets from the wireless interface addressed to any other address not in the routing table, the table maps it to the *IP_BROADCAST* address on the wired interface. This packet will be broadcast to all APs and the gateway in the network. The AP who has an entry for the IP address forwards it to its wireless interface while all other APs drop the packet. This is an acceptable solution because the bandwidth of the Ethernet is much greater than that of the radio.

Routing table entries can expire or be deleted by a handoff. Every packet from a mobile bound to the AP will update its routing table entry. However, the entry will expire if the entry is not updated before the expiration time. This is either due to the removal of the mobile from the network, failure of the mobile, or there is a very poor link between the mobile and the AP. In all cases, the entry is removed.

Once the mobile is no longer in the global state of the network, it will not hear any beacons with its address in the schedule. If the mobile is still active, or if it becomes active again, it will discover this and reset itself to an uninitialized state. It then sends an upstream control message requesting to register with one of the AP's whose beacon the mobile heard.

4.4.2 Handoff

When the network finds it necessary, control of a mobile is transfered from one AP to another via a soft handoff. An active mobile will never be "lost" in the sense that its address is always in the routing table of some AP in the network. Therefore, no packets are ever lost due to routing. The special instances where this is not true was discussed in the previous section on routing. Since this MAC protocol implements the adaptive handoff algorithm discussed in the previous chapter, this section will discuss how the MAC protocol implements Connectivity and Optimization handoffs.

Connectivity

Connectivity handoffs react quickly when the mobile-AP link degrades to a point where it is no longer considered acceptable. There are many dimensions of the link that can be used for assesment of the link quality. In the case of this network, error rates, specifically packet error rates (PER) are used. PER is easy to measure from either the MAC or the link layer. The PER performance of the RFM ASH transceiver has been studied and has shown to correlate with distance, number of walls, and other geometric features between transmitter and receiver.

In the description of the connectivity handoff sub-algorithm in Section 3.2, we indicate that a short-term averaged measurement of a link parameter should be used to ensure the reactivity of the algorithm. Here, this is defined to be $1 - PER$. PER is calculated every

t_PER seconds which in itself is a short-term averaged measurement of a channel metric, packet loss. Therefore, in the normal case, no averaging is necessary.

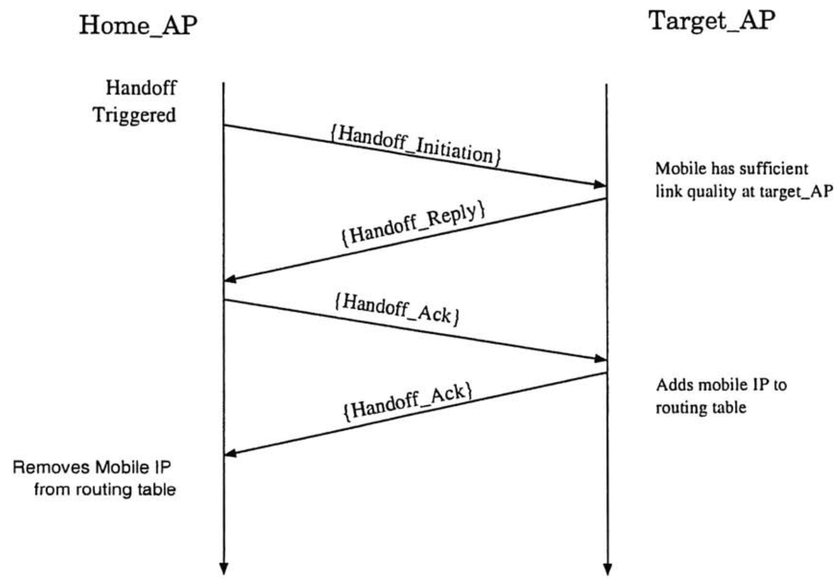

Figure 4-6: Message Exchange of the Connectivity Handoff

Once the connectivity algorithm determines that a handoff is necessary, the the following steps are executed by the network. First, the home AP of the mobile broadcasts a $HANDOFF\_INITIATION$ packet who's payload includes the mobiles IP address, the current channel metric, and a time stamp. Other APs keep link information for all mobiles it hears. If another AP can hear the mobile, and its link information satisfies the criteria specified by the algorithm, then it sends a $HANDOFF\_REPLY$ packet containing the time stamp of the original $HANDOFF\_INITIATION$ message, the mobile's IP address. The home AP replies to the first $HANDOFF\_REPLY$ it receives for the mobile with a $HANDOFF\_ACK$. Once the target AP receives the ACK, it updates its routing table, sends a $HANDOFF\_ACK$ back to the original AP, and begins to act as the mobile's home AP. The original AP removes the mobile's address from its routing table once it receives the $HANDOFF\_ACK$. If the mobile can still hear the original beacon, it will discover that it is no longer in the schedule, and will then begin to examin other beacons. If it cannot hear the beacon anymore, the mobile starts to search for a new beacon. Once it finds a beacon with its MAC address in the schedule, it will begin to transmit and receive with the new AP. If a mobile is without an AP for a maximum time $T_{restart}$ and begins in dormant state.

Optimization

In the context of the TDMA-based contention MAC protocol, the simple throughput optimization discussed in section 3.3 is still valid. In a simple network where there are three linearly arranged overlapping cells A, B, C, and a mobile located in each of the two overlapping borders (see Figure 3-1), if cell A assigns M_1 to slot S_1 then, it will jam the channel at AP_B forcing it to lose a slot. However, if M_2 is assigned to C then it can transmit in the same slot as M_1 and thus adds to the aggregate throughput.

The algorithm receives $OPTMIZE\_UPDATE$ packets every t_{opt} timing synchronization beacon period. These packets contain all the information that the algorithm needs to build its database of total network state. The packet includes the following information:

- M_i is the IP address of a mobile that is bound to the sender.
- num_slots_i is the number of slots that are assigned to M_i.
- OM_i is the IP address of an outside mobile that is not bound to the AP but the AP has successfully received packets from it.
- CM_i is the uplink channel metric of OM_i

The optimizing algorithm runs on a node in the Ethernet LAN. The optimizer component that implements the algorithm lies just above the Ethernet MAC layer. Timing wise, we assume that the algorithm runs fast enough such that between the sending of the first $OPTIMIZER\_UPDATE$ packet to the time to the time it sends out the $OPTIMIZER\_NOTIFICATION$ packet notifying which mobiles should be handed off, there are very few if any connectivity handoffs. The primary data structures used consists of a search tree and a database that contains all the information on the state of the network.

Besides the explicit information included in the update packets, the following information can be inferred by the database in the optimizer. Whether a mobile lies on an overlap of multiple cells is determined by whether more than one AP reports hearing the mobile. Two cells are neighbors if they hear the same mobile. From this, the database can satisfy the queries made by the tree search algorithm. Additionally, the network is considered stable if more than eighty percent of the mobiles are stable.

The constraint used in this network is the minimal requirement applied to the link error rate by the connectivity handoff algorithm. This constraint avoids handing off a mobile

mobile over to a cell that will cause a connectivity handoff back to the original cell again. This is the minimal constraint that should be applied.

The reward calculation process used at the end of the tree search algorithm evaluates how much parallel transmission can occur in the network. It assigns as many mobiles located on overlaps in parallel as possible and selects the partition that can be assigned with the least number of slots plus the penalty of all the handoffs that must be executed to form this partition from the orginal network partition. First it tries to assign as many mobiles on overlaps as possible in the first slot. If there is a conflict within a cell, defined to be when one mobile on the cell's overlap area is already assigned to the slot, a second round will try to assign the new mobile to a second slot. This will continue until all mobiles on overlaps are scheduled.

After the optimal partition with the highest reward value has been determined, the optimizer node sends out an *OPTIMIZER_NOTIFICATION* message containing all the mobiles that need to be handed off and to which AP. The AP's will then execute the handoff in a similar fashion as the connectivity handoff. The home AP sends to the target AP a *HANDOFF_REQUEST* message for a mobile M, the target updates its routing table, sends back an *HANDOFF_ACK*, and begins to act as the new mobiles home AP. Once the original AP receives the ACK, it removes the mobile from its routing table.

Chapter 5

Simulation and Analysis

The protocol discussed in the previous chapter implements an adaptive handoff algorithm that maintains link quality and reconfigures the mobile configuration to optimize network performance. To investigate the validity of these ideas, a simulation module extending the existing *NS* simulator is built. Then, using the extended simulator, scenarios are created to investigate whether the performance of the algorithm matches that of its design criteria.

5.1 NS Simulation

The NS simulator is a discrete event driven simulator targeted for networking research. It is the product of the Virtual InterNetwork Testbed project. NS has a composable simulation framework which allows the integration of independent simulation efforts into one simulator, thus allowing great flexibility to explore the complexities of scale and interaction between network components.

In NS, simulation is executed as a series of events dispatched by a central scheduling module. As the simulation progresses, events are dequeued and executed according to their specified time. A global queue of events is maintained and events are executed in series. This simulator is appropriate for small to medium sized networks, as large high-speed simulations are extremely slow due to the nature of the event handling module.

NS is ideal for this project due to its composable simulation framework. Currently, NS has support for numerous network architectures and protocols. These include wired networks, Local area networks (LAN), satellite networks, and mobile networks with wireless support. These network topologies are themselves composed of specific protocol and physical layers. Various protocols in the application, transport, network, link and MAC layers are supported by the simulator. On an even lower level, the simulator provides timing mechanisms that are crucial to building timing-sensitive protocols such as TDMA.

To build the topology of the proposed network, first nodes and all their protocol layers must be built, then, they must be linked according to the specified connection medium to form the network topology. Simulation of this proposed network cannot be readily built with the existing components of NS. First, there is no support for any TDMA based protocol. Second, a new routing protocol must be added since there is no support for the mixed wired-wireless topology of the picocellular network. Finally, a different network interface and propagation model need to be developed to better simulate the wireless link.

5.2 Node Architecture

5.2.1 Packet Path

The mobile node and the access point node have two different architectures. The mobile node architecture illustrated in Figure 5-1[1] has only one interface. Packets are generated in the source (Src) and passed to an entry point (entry_) where they are filtered through an address demux. All packets addressed to other IP addresses are passed to the routing agent which sets the next hop value of the packet. Packets for the node's IP address are passed to the port demux which delivers the packet to the port specified in the IP header. The link layer (LL) manages a virtual Address Resolution Protocol (VARP) table. Unlike an actual ARP protocol, the VARP is statically calculated in the begining of the simulation. The link layer resolves the MAC address of the next hop destination with VARP, inserts it into the MAC header of the packet and passes it on to the interface queue (IFq). The interface queue buffers packets so that the timing dependent MAC layer receives a packet from the higher layers only when it is done servicing a packet. The network interface (NetIF) simulates the hardware that access the wireless channel. It is suceptable to collisions and interference depending on the propagation model used. The radio propagation model contains specific information about transmitting characteristics of the radio so that the network interface can decide whether a packet should be received successfully.

[1] *ns* Notes and Documentation by Kevin Fall

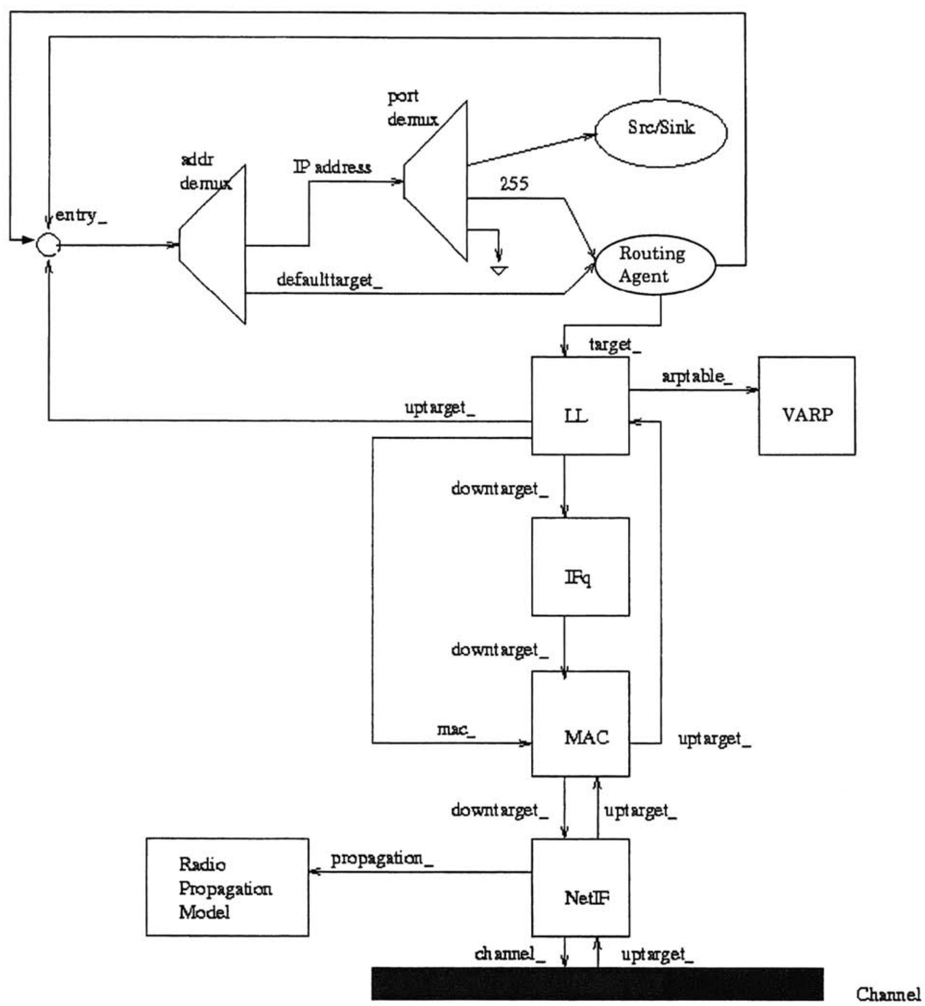

Figure 5-1: Mobile Simulation Object

The access point has a dual wired and wirelss architecture. It has the wireless architecture of the wireless node, as well as a wired LAN architecture as shown in Figure 5-2 both joined by a common address classifier at the head of the classifier chain. Packets from higher layers such as the application, transport, and network layers are passed to the node's entry point. It is then passed through a series of classifiers depending on the address architecture, and multiplexed to the Queue of the link layer. The link layer sets the MAC destination address in the MAC header of the packet. This involves first finding the next-hop destination IP and then resolving it to the MAC address using a VARP module. After the MAC address is set, the packet is first passed to an interface queue, which prevents the MAC layer from being flooded with packets and consequently dropping most of them. The MAC layer implements the IEEE 802.3 Ethernet specification which implements channel sensing and collision detection. At the physical layer, the Channel simulates the actual transmission of the packet in a shared medium. Finally, the MAC classifier is responsible for delivering MAC packets to the right MAC layer in the LAN. It also simulates a broadcast medium by replicating packets to deliver to every MAC layer in the LAN for MAC broadcast messages. On its way up to the higher layers, the MAC packet is passed from the Link Layer to the entry point of the node to be filtered throught the classifier chain.

5.2.2 Routing

For routing of packets between wired nodes, the simulation uses static unicast routing which precalculates the routes at the beginning of simulation using Dijkstra's all-pairs SPF algorithm. It assumes that the wired topology does not change and doesn't account for link failures. The LAN is regarded as a single hop with the architecture as illustrated in Figure 5-3. Each node in the LAN shares a single LanRouter which is also precalculated with the static unicast routing.

The Access Point has two routing mechanisms. The wireless interface passes packets up to the entry point of the node where these are passed through the classifiers. Packets not destined for the access point itself or for a wired node in the simulation, will be forwarded to the access point routing agent BaseRtg. BaseRtg maintains entries in its routing table that maps IP's of mobiles bound to this AP to their own addresses, and for all other IPs it maps to the IP broadcast address (IP_BROADCAST). For all packets for a bound mobile from either the wired or wireless interface, BaseRtg sets the next hop address in the IP

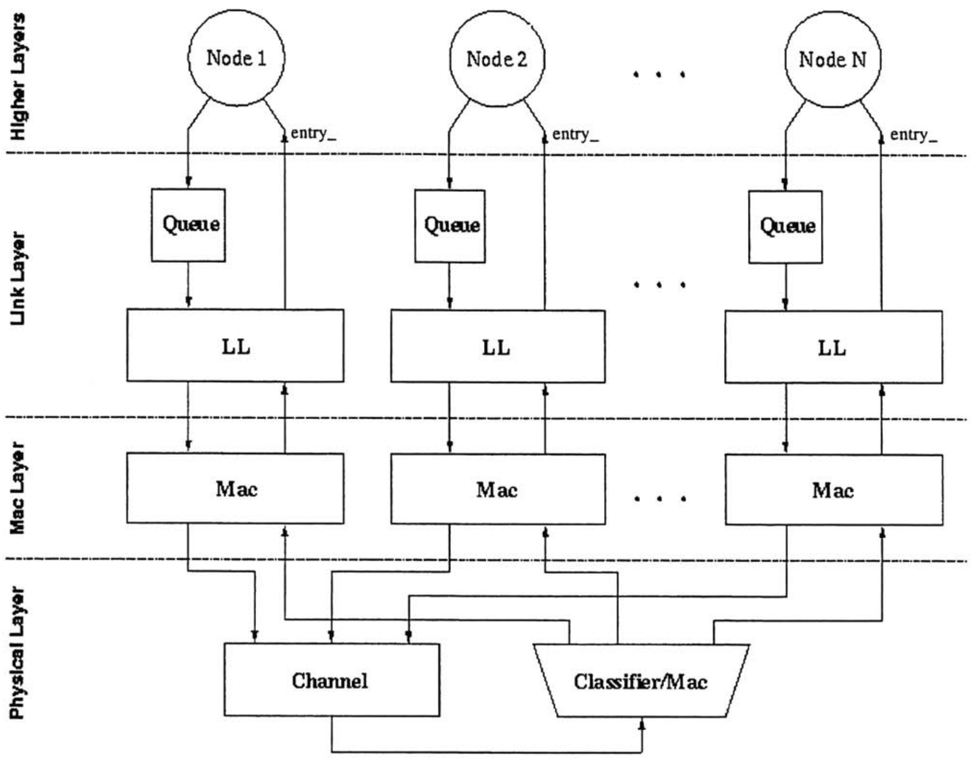

Figure 5-2: LAN Simulation Object

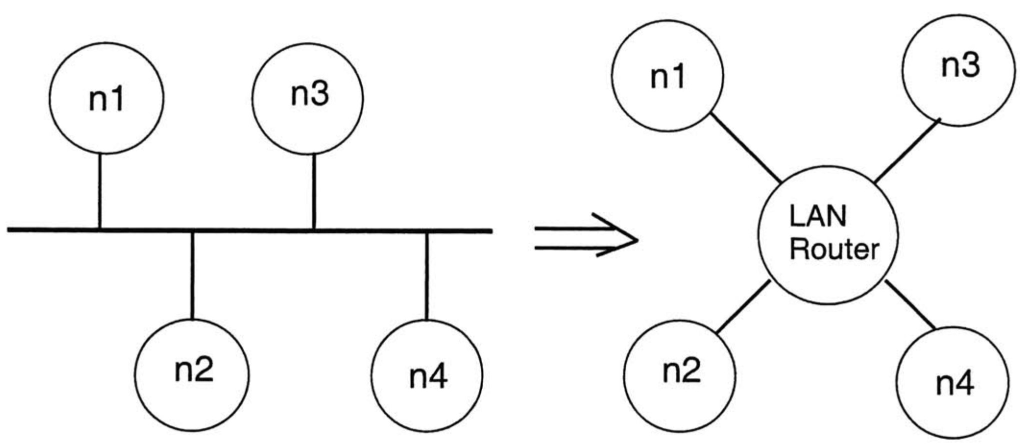

Figure 5-3: LAN routing in the NS simulator
2

header to the destination address and sends the packet down to the wireless interface. If the packet is for some mobile not bound to this AP from the wireless interface, BaseRtg sets the next hop address to IP_BROADCAST and passes it to the wired interface. Once the data packet reaches the BaseRtg at other mobiles, the AP that has the mobile forwards the packet to the wireless interface while all other APs drop the packet.

5.2.3 Network Interface

The network interface simulates the hardware that accesses the shared medium, in the case of a wireless channel this is the radio. NS has an implementation that approximately models the Lucent WaveLan Direct Sequence Spread Spectrum radio (DSSS). However, DSSS radio's spread spectrum modulation scheme is much more robust than the amplitude modulation scheme that the radio in this network employs. This among other RF factors make the current implementation inappropriate for the purposes of this simulation. Using the existing wireless propagation model, and creating a new network interface for the class of radios such as the one in this network makes this simulation more realistic. This involves changing the energy model of the module which, given the receiver's sensitivity and out-of-band rejection, determines whether the receiver has the bility to receive this packet.

5.2.4 Handoff Simulation

Two separate approaches were used in implementing handoff in the network. Both the decision making and the execution of connectivity handoffs are implemented in a distributed manner in the simulation as described in the previous chapter. The optimizing handoff is simulated in a centralized module that does not simulate the message passing in the back-end. Instead, the optimizing module maintains a list of all the access points in the system and directly collects the needed information from each AP. Since the entire algorithm is centralized, adding the message passing does not add any significant insight to the simulation. The execution of the optimization handoff, however, is implemented in a distributed manner.

5.3 Evaluation of Performance

The evaluation of network performance focusses on the performance of the handoff algorithm. In general, the throughput of the entire network is evaluated to determine whether it can sufficiently support simple applications. More specifically, the reactivity of connectivity handoffs to changing channel conditions and the optimization handoff's ability to improve existing network throughput are to be examined. Additionally, the interactions between the two handoff layers are examined to guage whether they compliment each other as they are designed to.

5.3.1 General Performance

The aggregate throughput and average throughput of the network can be studied under varying network conditions. The number of mobiles and the amount of aggregate movement in the network are both interesting factors that affect throughput. Mobile movement in this experiment is limited such that most of the network is relatively stable to ensure that the optimizing handoff is triggered. The resulting improvement is observed. Throughput is measured for each run with increasing numbers of mobiles interdispersed through out the network. Each mobile will be generating simulated FTP traffic for the entire duration of the simulation. Open space assumptions are used for radio signal propagation.

Average throughput is the number of packets that each mobile node transmits and receives in the duration of the simulation divided by the total simulation time averaged over all the mobiles in the system. This measurement reflects the performance that each mobile expects to see in this network. We expect the function of average througput versus number of nodes in the network to decrease semi-linearly and then converge to zero. This curve has three stages. The first stage is when there are more slots in each cell than mobiles. The slot scheduler will assign more than one slot to a mobile in this situation. This is where throughput is the highest since most slots are used and the chances of collision is small. No mobiles share cells slots and thus are highly guaranteed to tranmit and receive successfully. However, as the number of nodes increase, slots begin to be shared and the throughput of each sharing mobile is further reduced along with their probablility for collisions. This state of the curve has a markedly sharper negative slope. Finally, the average throughput of the system converges to zero as the number of mobiles increases further.

The aggregate througput of the system is the sum of all the packets transmitted and received by mobiles averaged by the length of the simulation. It shows how well the system's bandwidth resources are managed under various network conditions. The expected curve of aggregate throughput as a function of number of mobiles in the network begins at zero with zero mobiles and increases almost linearly to close to the maximum aggragate throughput of the system (number of cells X maximum radio bandwidth). This is when there is at least one mobile using the bandwidth of each cell for most of the simulation. Good resource management can keep the throughput at this maximum value for a higher range of mobile densities. The curve will decrease at a saturation point until eventually it converges on to zero.

5.3.2 Effect of Optimizating Handoffs

By comparing the results of the previous experiment with and without optimizing handoffs, we expect a significant improvement in performance in the network as a result of the optimizating handoff. Average throughput and aggregate throughput are measured. Without optimizing handoffs network aggregate throughput is expected to remain at its peak value for a shorter amount of time, Since optimizing handoff tries to optimize throughput, it has little or no affect on the throughput when the network is sparse. But as mobile density increases, the probability of a network configuration that can be optimized through handoff increases. Therefore, the two curves described in the pervious section are very similar to the ones that will be produced here at small mobile densities. As the number of mobiles increases, the ability of the network to share its resources is not as effective without the optimizer, and thus, it will remain in the highest throughput point very briefly before dropping quickly to zero. With the optimizer, a larger range of node densities will be able to sustain better throughput resulting in a curve that drops off much slower as mobile density increases.

5.3.3 Reactivity and Stability of Connectivity Handoff

Reactivity without instability is one of the main design criteria of the handoff algorithm. To characterize this aspect of handoff, a different simulation is needed. We move a mobile moves back and forth between two over lapping cells in a straight line at a constant speed. The location of the mobile when it hands off is recorded. A histogram of the number of handoffs verses the location at which each is executed can be plotted with reference to the approximate cell boundaries. The plot of a reactive algorithm is one which has a high rate of handoff at the boundary of the cells. For each location, the amount of time between consecutive handoffs is calculated and plotted. Ideally only one handoff is executed at the cell border. However, while the mobile is in the overlapping area, the link conditions of both sides may be similar, causing a number of quick handoffs back and forth between the two cells. If the average time between handoffs is much lower in the cell border area, then this would be an indication of instability in the algrithm.

Chapter 6

Conclusions

The future of computing will be one of high mobility and high specificity of applications. Wireless communications provide the underlying network support for such a vision. The physical unpredicability of the wireless channel is a fundamental characterisitc of wireless communications and proves to be the challenge to overcome in order to build a successful network. In infrastructure networks, mobility is managed with handoff. Since the existence of cellular phone networks, handoff has been studied extensively. These efforts, however, remained within the localized context of a single mobile and two cells.

6.1 Summary

In this thesis, we presented an adaptive handoff algorithm that can solve the handoff problem both in a localized context of one mobile and two cells, and in a global context of aggregate preformance optimization. This handoff algorithm utilzes a two-tiered approach that views the network with both a small and large time scale. Using a short time scale, the first tier enforces a minimum link quality while a large time scale provides the second tier with the confidence to optimize the mobile configuration among the cells.

A MAC layer protocol along with relevant network design was presented to give the algorithm a network context in which to study. The TDMA-based MAC protocol manages the channel with dynamic slot assignement and minimizes collisions as a result of scheduling conflicts with a contention resolution scheme.

The resulting network design was built in the VINT NS simulator. A series of tests using this simulation is presented to investigate various aspects of the handoff algorithm.

Investigation of aggregate throughput, average throughput, handoff delay and stability were discussed.

6.2 Future Work

So far the discussion of the applications to be used on this proposed network with the adaptive handoff algorithm has been purely theoretical. We make certain assumptions about their service needs from the underlying network. We also simulated the network, abstracting aspects of the physical network such as the wireless channel and the radio transceiver. Developing applications on top of an actual implementation of the given network will be a physical demonstration of the viability of the handoff algorithm and the protocol.

A two tiered adaptive handoff algorithm has an applicability beyond the specific network that was presented in this thesis. In a IEEE 802.11 network of FHSS or DSSS wireless radios, intercell interference is low, thus the method of optimization used in the single frequency network is not applicable. However, the benefits of a handoff algorithm that allows the network not only to quickly handoff mobiles when they are moving between cells, but also to optimize the mobile configuration in a larger time scale, is worth further investigation and development.

Bibliography

[1] Andreas Fasbender and Frank Reichert *Any Network, Any Terminal, Anywhere* IEEE Personal Communications, April 1999

[2] Rohit Ghai and Surresh Singh. *Protocol for Seamless Communication in a Cellular Network* Proc. IEEE International Conference on Communications, May 1994.

[3] Eunyong Ha and Yanghee Choi *New Pre-Handoff Scheme for Picolellular Networks* Proc. IEEE International Conference on Communications, February 1996.

[4] C. Perkins *RFC 2002: IP Mobility Support* http://andrew2.andrew.cmu.edu/rfc/rfc2002.html.

[5] A. G. Valko *Cellular IP - A New Approach to Internet Host Mobility* ACM Computer Communication Review, January 1999.

[6] G. P. Pollini *Trends in Handover Design* IEEE Communications, Vol 34, No. 3, March 1993.

[7] Nishith D. Tripathi, Jeffrey H. Reed, and Hugh F. VanLandingham *Handoff in Cellular Systems* IEEE Personal Communications, December 1998.

[8] *Bluetooth Special Interest Group* August 1999.

[9] IEEE *Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications* July 1998.

[10] L. Harte, A. Smith, and C. Jacobs *IS-136 TDMA Technology, Economics, and Services* 1998, Artech House Publishers.

[11] L. Peterson and B. Davie *Computer Networks, A Systems Approach* 1996, Morgan Kauffmann Publishers.