A NOVEL ADAPTIVE METHOD FOR MINIMIZING RECURRENT HANDOFF DELAYS IN MOBILE IPV6 NETWORKS

by

May Siksik

BASc: University of British Columbia 2001

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in

The Faculty of Graduate Studies

(Electrical & Computer Engineering)

The University of British Columbia
April 2006

ABSTRACT

Future wireless networks will be entirely based on the IP protocol. High degrees of portability available in an all-IP network will be possible only if we are able to address the insufficiencies of mobility management. One of the major insufficiencies affecting mobility management is the delay associated with the handoff process. During a handoff, the mobile node can become unreachable causing data destined to the node to be lost. As a result, it is important to minimize the delays associated with the handoff process. In this research, we evaluate various mobility management protocols and show that Mobile IPv6 is the most appropriate solution for micro-mobility management. We study the handoff process using Mobile IPv6 and characterize its phases. This characterization shows that move-detection delay is the bottleneck of the handoff process accounting for almost 60% of the combined Layer-2 and Layer-3 handoff delays. To minimize this delay, the router advertisement interval should be set as short as possible without exceeding its assigned bandwidth. This approach has been studied in previous work but only in the context of a single handoff. We propose an improved mobility analysis that assumes a more realistic model that incorporates recurrent handoffs likely to occur during Eager Cell Switching. We then derive an expression for bandwidth that accounts for the non-zero delay separating consecutive handoffs. Further, we use our analysis to provide a new adaptive mobility method capable of dynamically adjusting the advertisement interval as a function of the recurrent cell switching behavior and the available bandwidth. Finally, we verify this model using analysis and simulations using OPNET® and MATLAB® based simulators. Results show the reduction in move-detection delay to be as much as 65% using the analysis and as much as 56% using simulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my husband Peter for his incredible support throughout this work. I would like to thank my parents and the rest of my family for helping me with all my achievements.

I would like to thank my supervisors Dr. Hussein Alnuweiri and Dr. Saif Zahir for their continued guidance and support throughout this work.

1 INTRODUCTION AND OVERVIEW

1.1 Motivation

The demand for wireless access to the Internet through mobile devices continues to grow. To meet this demand, Internet and wireless networks have merged into a unified network, the Wireless Internet. Enabling the Internet requires that several issues be addressed. In a Wired Internet, all nodes are reachable during normal operation. In a Wireless Internet, however, nodes constantly move and their addresses change in a corresponding manner.

When a mobile node (MN) moves between cells, it must change its point of attachment (or Access Point) to the network. This process is referred to as a handoff. The handoff process becomes critical in the case of micro-mobility where the MN moves within the same domain and is therefore likely to experience frequent handoffs.

An adverse effect of a handoff is that the MN is not reachable throughout the process. As a result, it is important to minimize the delay associated with the handoff process. To help gauge where research effort should be focused, it is important to fully understand what parts of the handoff are responsible for the most delay. Further, it is important that the delay information be provided with as much detail as possible and in a consistent manner. Part of this research work provides such an analysis. Results provided in Chapter 4 show that the most prominent source of delay is Layer-3 move-detection accounting for almost 60% of the combined Layer-2 and Layer-3 handoff delays. This result motivates the need for a more thorough analysis of move-detection in order to identify which parameters have the highest impact on move-detection delay and bandwidth overheads. Once these parameters are identified, a model that manipulates such parameters in order to minimize move-detection delay can be proposed.

The delay of move-detection is highly dependant on the handoff initiation algorithm used. Two main algorithms have been proposed for the initiation of handoff in Mobile IP [1,2] networks, namely Eager Cell Switching (ECS) and Lazy Cell Switching (LCS) [3]. In ECS, the MN switches to a new router through a Layer-3 handoff as soon as it starts to receive Router Advertisement (RA) from the new router. LCS differs in that it depends on the lifetime field within the main body of the ICMP6 portion of the RA. If the MN does not receive an RA within the specified lifetime, it can either wait to receive two more RAs before declaring the current router unreachable or it can probe the router to learn if it is still reachable. Once LCS has determined that the old router is unreachable, it initiates a handoff to a new router as soon as it finds a valid care-of-address in the new RA.

Previous work has found that ECS performs better than LCS with respect to handoff delays [4,5,6]; however, this analysis assumed perfect cell boundaries (i.e. non-overlapping cells), and ignored the possibility of recurrent handoffs between different networks. Recurrent handoffs can be a result of the recurrent detection of RAs from new but unstable networks barely within reach. If this is taken into consideration, it is possible that ECS would not perform well when the MN switches between different routers at a high rate.

Previous work oversimplified the physical nature of a network by only considering a single handoff and by assuming that handoffs involve a "straight-line" motion. This research work proposes an approach that considers the problem of a single user experiencing multiple handoffs as they move along a boundary between two or more cells or areas covered by multiple routers. This problem causes multiple consecutive handoffs to take place causing unwarranted delays and straining of mobility management resources in the associated routers.

1.2 Research Approach and Objectives

1.2.1 Research Objectives

The main goal of this research is to provide a method to minimize the delays associated with the handoff process and consequently to minimize interruption time and the resulting data loss encountered by an MN.

The following subsections provide a summary of the various stages of this research work:

1.2.1.1 Selecting a Mobility Management Protocol

Before proposing any enhancements to the handoff management mechanism, it is important to first select a mobility (and handoff) management protocol based on predetermined evaluation criteria. Although many mobility management protocols have been proposed, the only one that has been standardized by the IETF is Mobile IP [1,2]. In spite of this, it is important to evaluate mobility management using various protocols according to specific criteria in order to determine which protocol is the most appropriate to use for mobility management. Because the handoff process occurs more frequently in the case of Micro-mobility (where the MN roams within the same domain), we chose to specifically evaluate Micro-mobility management using those protocols. The work done in this stage was published in [7].

1.2.1.2 Study of the Handoff Process and Identifying its Bottlenecks

Once a protocol has been selected, the various phases of the handoff process should be characterized using a consistent simulation platform in order to determine where bottlenecks exist. This requires a detailed breakdown of the handoff process in layers-2 and layer-3 and a breakdown of the delays associated with each phase. The work done in this stage of the research was published in [8].

1.2.1.3 Addressing Handoff Bottlenecks

Once handoff bottlenecks have been identified, the next stage is to design a method that addresses these bottlenecks. As will be shown in Chapter 4, move-detection is the prominent source of delay in the handoff process. Because the delay of move-detection is highly dependant on the handoff initiation algorithm used and the RA interval, it is important to consider different handoff initiation algorithms and provide an analysis of the RA bandwidth and RA interval limit as a function of the total available bandwidth. Such an analysis should take into account the cell switching rate and provide a network model capable of dynamically adjusting the advertisement interval as a function of the cell switching behavior and available bandwidth with the goal of minimizing move-detection delay.

1.2.1.4 Verification of the Proposed Method

In this stage, the proposed method is verified by studying the trends exhibited by the model and then comparing them against simulation results.

1.2.2 Required Tools

Two simulators were built to implement the second and fourth stage experiments. We used OPNET® as the basic platform and added the required features to form the final simulator needed for simulating a layer-2 handoff using IEEE 802.11 and layer-3 handoff using Mobile IPv6. This simulator provided the delay profiles for the handoff process. To verify the method proposed in the fourth stage, we built a MATLAB® based simulator that imports the necessary values from the simulator built for stage two.

1.3 Previous Work and Contributions

1.3.1 Analysis of the Handoff Process and Associated Delays

Previous research in [9] and [10] provided a detailed analysis of layer-2 handoff delays in IEEE 802.11 networks but did not consider layer-3. Velayos et al. [9] showed that move-detection is responsible for the most delay in layer-2. Mishra et al. [10] did not consider the move-detection phase but instead divided the handoff process into discovery (i.e. to search for an AP) and re-authentication (i.e. to switch to a new AP). The work in [10] showed that the discovery phase contributes the most delay. The work in [11] provided a comparison of the overall delay due to layer-2 handoff to the overall delay due to layer-3 handoff. However, [11] did not consider the break down of the total delay to show the contribution of each phase and did not specify relevant factors such as the beacon period, RA interval, and the number of RAs after that mark a move-detection.

One of the contributions of this research is the detailed analysis of the handoff process in layer-2 and layer-3. This is an important contribution in that it provides delay information for all relevant phases using a consistent simulation platform. This helps to better understand what parts of the handoff are responsible for the most delay and consequently gauge where research effort should be focused in order to minimize these delays.

1.3.2 Addressing Handoff Delays

The handoff analysis performed in this research showed that the main bottleneck in handoff delays is move-detection which is the time it takes an MN to detect that it has moved. Previous mobility research that addressed this bottleneck such as [12,13,14,15] have focused on commercial cellular networks. The models developed by this research are inadequate for mobile networks covered with WLAN APs based on the IEEE 802.11 standard. In WLANs, the coverage and density of APs are relatively ad-hoc and sensitive to many factors including signal strength, fading, interference from neighboring WLAN devices, etc. Signal and move-detection is more variable under these conditions and must be analyzed as a function of the bandwidth capacity of the system.

The authors in [16] analyzed move-detection delay based on a fixed RA interval. They assumed that the handoff execution time is equivalent to the registration time and is small enough to be disregarded. Further, the work in [16] did not provide a model for move-detection and RA delays, nor did it consider the impact of the MN cell switching rate.

In [17,18], the move-detection delay was analyzed based on a uniform distribution advertisement interval. The authors in [17,18] define C as the time when an MN enters a new cell, R as the time when an MN receives the first RA from a router in the new cell, and 0 as the time when the router sends an advertisement prior to R. The authors rely on the intuition that the probability that C occurs in an interval is proportional to the interval.

In addition, it is assumed in this work that there are perfect cell boundaries implying that entering the range of a new network coincides with leaving the current network. It is not sufficient to analyze the move-detection delay based on such simplifications since "real" cell boundaries invariably have overlap.

In [19], move-detection was reanalyzed without the intuition relied upon in [17,18], and an expression was derived for the expected handoff delay that was incorporated into a bandwidth calculation. The main idea of this work was to dynamically adjust the RA interval as a function of bandwidth. Although providing substantial improvement over previous models, the work in [19] still assumed perfect cell boundaries and ignored the effect of recurrent handoffs indicative of ECS. This simplification erroneously increases the bandwidth requirements per handoff by ignoring the random delay between consecutive handoffs. This simplification is not just an issue for [19] but for all single handoff models.

In [14], the performance of "back-and-forth" movement was measured for various move-detection algorithms and compared against "straight line" movement. This work does not relate the "back-and-forth" movement to bandwidth and the RA interval and does not model these relationships analytically.

The work in this thesis assumes an ECS handoff policy to help minimize the handoff delay. Because of the recurrent cell-switching behavior associated with ECS, our approach accounts for the consecutive handoffs that could occur for an MN moving amongst several overlapping cells. To accomplish this, a simple single-handoff model could be used, but would not scale well because it ignores the random delay that exists between consecutive handoffs. Figure 1 shows the delay breakdown of $i+1$ consecutive handoffs. This figure illustrates that each cell-switch instance is comprised of the delay needed to perform a handoff (*HP*) plus a random delay (δ) before the next cell-switch occurs. *HP* includes layer-2 handoff delay + layer-3 move-detection delay + layer-3 binding updates (including DAD + address auto-configuration [1]). The random delay, δ, is important when determining the handoff bandwidth overhead and is included in our model. Overlooking this random delay factor can lead to bad network design where the available bandwidth is under-estimated thus causing the RA interval to be longer than needed. As a consequence, the move-detection delay would also be increased. It is important to emphasize that move-detection delay was previously found to be the bottleneck in the handoff process. Subsequently, we use this bandwidth calculation to guide an adaptive network model that minimizes move-detection delay by adjusting the RA interval based on recurrent switching activity and available bandwidth. This model uses Renewal Theory [20] to provide a mapping between the cell-switching distribution and the expected delay between consecutive handoffs.

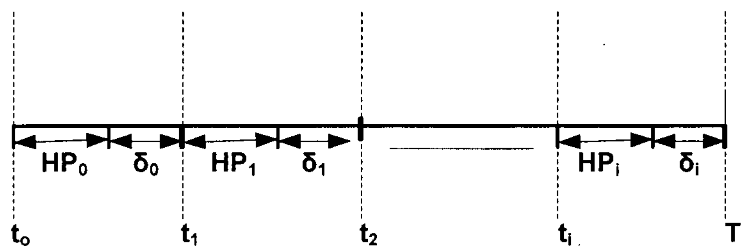


Figure 1: Recurrent cell switching can result in multiple handoffs (Handoff Process time HP = layer-2 handoff delay + layer-3 move-detection delay + layer-3 binding updates delay) in a given time period of interest (T).

1.4 Thesis Organization

This thesis is organized as follows: Chapter 2 provides background information on this research area. Chapter 3 gives a general survey on the main mobility management protocols. Further, Chapter 3 defines evaluation criteria [21] and provides a comparison of micro-mobility (i.e. mobility within the same domain) management using the different discussed protocols according to these criteria. Based on this comparison, a protocol is selected for evaluation. Chapter 4 studies the handoff process in layer-2 (using IEEE 802.11) and in layer-3, when mobility is managed at layer-3 using the selected protocol. It describes the details of the process used to characterize the handoff in layer-2 and layer-3, and the results which show where bottlenecks exist. Chapter 5 discusses the analytical model and design details of an adaptive method that we propose in order to address the identified bottlenecks. Chapter 6 validates the model proposed in Chapter 5 by first studying the analytical trends exhibited by the various parameters and discussing how well they match intuition. Next, these trends are compared to results generated by a simulator. The details of the simulator implementation are also included in this chapter. The final chapter concludes the work done in this research, describes possible future work, and summarizes the contributions.

2 BACKGROUND INFORMATION

2.1 4G Networks

2.1.1 Networking Heterogeneity

Fourth generation wireless networks will provide wireless access in a diverse environment with different types of networks (e.g. WLAN, WPAN ...etc). In such a model, the user will be identified by a user address as opposed to an IP address. Support of this type of roaming requires location management techniques as well as support for handoff management. The network heterogeneity is not only limited to access technologies using different radio technologies, but it also includes different network architectures and protocols as well as different service demands.

2.1.2 Examples of 4G networks:

Figure 2 [22] shows an example of such networks. This example shows a wireless domain (a large IP wireless access network managed by a single administrative authority) using an UMTS-like radio interface. In such networks, the base stations (BTS) are controlled by the Radio Network Controller (RNC). In Radio Access Network (RAN) the mobile movements are entirely managed at the radio layer and are transparent to the upper layers. A RAN can thus be considered the smallest IP entity in the network. In figure 1, there is also a wireless domain using Wireless LAN (WLAN) as radio access points. The smallest IP entity in this network is the access point. Between their wireless domains, there is an IP inter-domain backbone dedicated to enable the roaming between the domains, in analogy with the current GPRS networks [22].

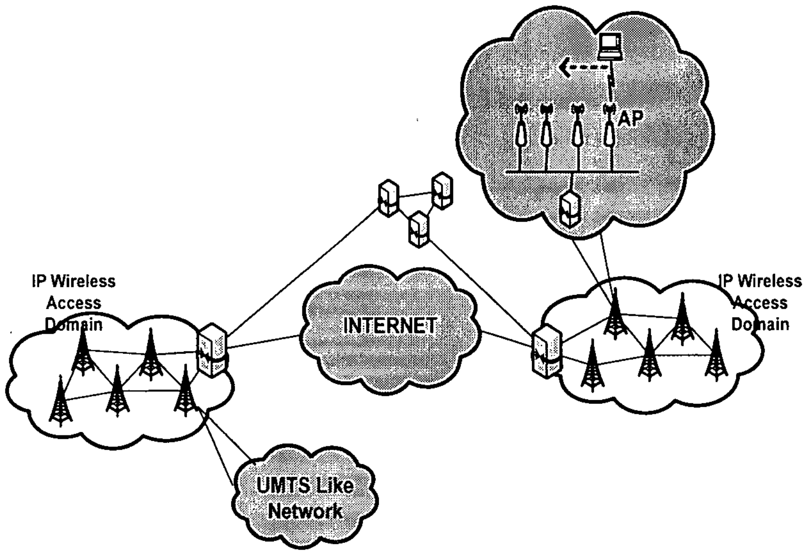


Figure 2: Fourth Generation Wireless Networks Example 1

Figure 3 shows another example of 4G networks. All networks shown provide different requirements of mobile users. In this scenario, while the user is downloading a file while at the office using an Ethernet cable, can unplug the laptop while and continues to download the file while the connectivity is transferred, first to the building's WLAN, and then to the cellular network.

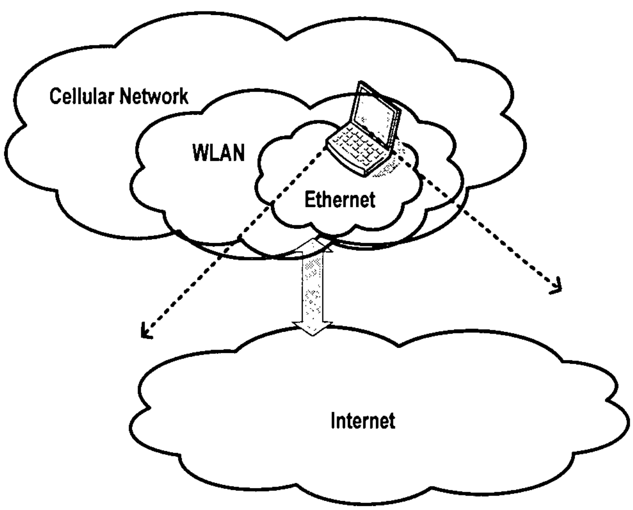


Figure 3: Fourth Generation Wireless Network Example 2

2.1.3 Roaming in 4G networks

2.1.3.1 Types of Roaming

2.1.3.1.1 Micro-mobility

This refers to the node movement within the same network/domain (Figure 4). In this case, the gateway connecting the location to the Internet does not change.

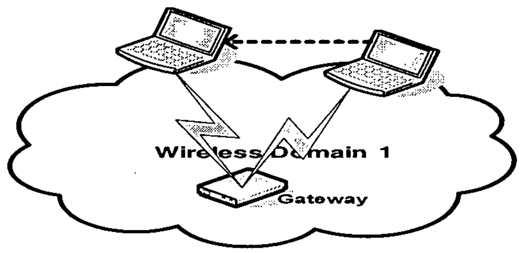


Figure 4: micro-mobility

2.1.3.1.2 Macro-mobility

This refers to the node movement between different networks/domains (Figure 5). In this case, the gateway connecting the location to the Internet is different.

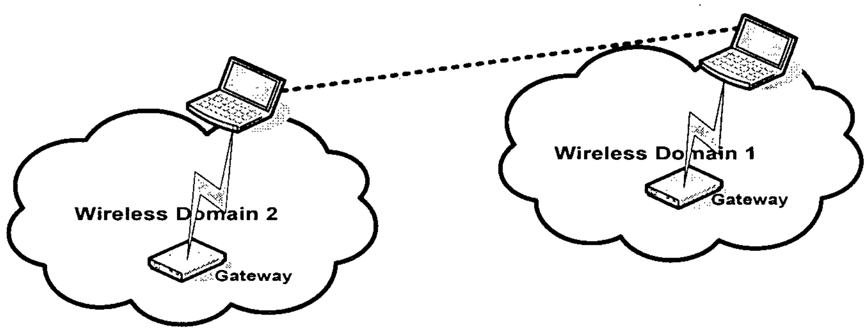


Figure 5: Macro-mobility

2.1.3.2 Levels of roaming

2.1.3.2.1 Node Mobility

This refers to the individual host or router mobility. It is important that the mobile management entity (i.e. router) can efficiently route data destined to the MN, which can be anywhere in the network, or even might not even exist in the network (for example when the MN is powered down).

2.1.3.2.2 Network Mobility

This refers to the entire network mobility [23]. This includes the mobile node as well as the access router. The network is considered mobile when a router connecting an entire network to the Internet dynamically changes its point of attachment to the Internet. Such kind of network is referred to as a mobile network.

Handoff management is more critical for mobile networks since there are multiple levels of mobility. For example when an access router moves, it's very important that all the MNs associated with this router are updated with the new location as fast as possible or otherwise, data destined to any of the MNs through the access router will be lost. Figure 6 gives an example of such scenario.

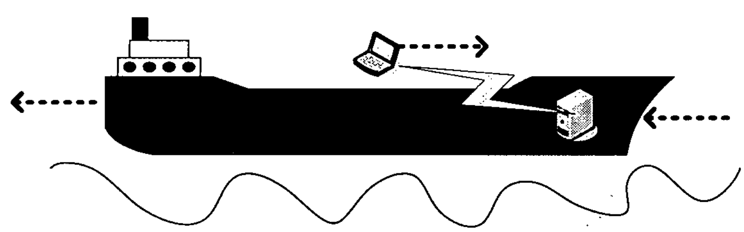


Figure 6: Network Mobility

In Figure 6, there are two levels of mobility: the host mobility, and the router mobility. In this scenario, the server on the boat is moving as the boat moves, and the host is moving as the person using it moves.

2.2 Mobility Management in 4G Networks

2.2.1 Mobility Management Issues

2.2.1.1 Handoff Management

When a mobile node (MN) moves between cells, it must change its point of attachment (or Access Point) to the network. This causes a handoff to occur (Figure 7). The handoff process involves several steps. First, Layer-2 (using IEEE 802.11 [27]) handoff occurs following a Layer-2 move-detection. Move-detection marks the point in time when a loss of connection with the current Access Point (AP) is detected, which in turn prompts Layer-2 to search for, and subsequently switch to a new AP. After Layer-2 has completed its switching, Layer-3 (using Mobile IP [1,2]) move-detection occurs which includes Layer-3 move-detection (i.e. detection of a loss of connection with the old router). The handoff process ends with Layer-3 binding updates (including address auto-configuration) and Duplicate Address Detection (DAD) [1].

Efficient handoff management is the most important issue in mobility management since handoff latencies result in session interruptions and packet loss. Handoff management becomes more critical in the case of micro-mobility since the MN is likely to experience frequent handoffs.

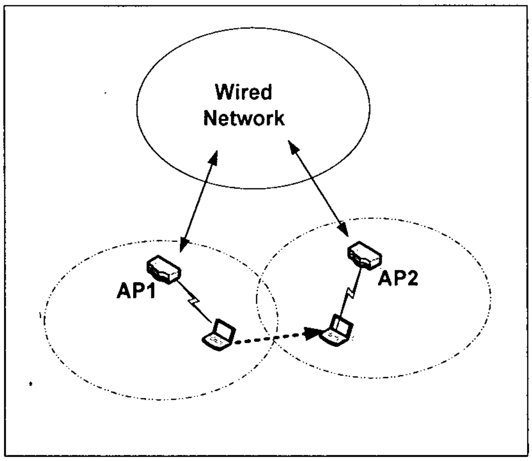


Figure 7: A mobile node switches from one domain to another causing a handoff

2.2.1.2 Intra-domain Traffic

Intra-domain traffic is the traffic exchanged within the same domain. This traffic is likely to be exchanged frequently and thus should be efficiently supported.

2.2.1.3 Network Design Requirements

Mobility introduces the need for a mechanism to forward data to the MN. This can be done through a mobility management entity that keeps track of the location of the MN and forwards data to it when its location is unknown to the corresponding MN, or through keeping track of new routes in the network that can be used to reach the MN. This adds to the complexity of the network design. Different mobility management protocols require different elements to be added to the network to implement mobility management. It is important that the mobility management protocol does not cause the network design to be very complex.

2.2.1.4 Scalability and Robustness

Because each MN adds to the load of the entity that manages its mobility, robustness of the network can be compromised when the number of MNs grows. The failure of the mobility management entity responsible for these MNs causes all of them to fail. As a result, it is important to maintain robustness and reliability as the network grows. It is important to encapsulate failures so that the failure of a single node does not spread and cause large parts of the network to fail.

2.2.1.5 Power

Mobile devices rely on limited battery power for their operations. As a result, it is important to provide a mechanism to reserve the MN's battery life when not communicating with other nodes for an extended period of time.

2.2.1.6 Security

Mobility creates a new challenge for network security. This is because the physical security of a cell (or domain) does not exist. This means that whenever the MN changes its point of attachment (or AP), it must perform authentication with the cell/domain it is attaching itself to using a log in name and a password. To protect this authentication information, there is a need for security mechanisms such as encryption techniques.

2.3 Managing Mobility at the Different Layers

Because mobility was not considered when the protocol stack was designed, mobility management functionality can be implemented at various layers, the session layer, the transport layer, the network layer, and the link layer. The following subsections discuss the different approaches used to implement mobility management at these layers and what the possible benefits and limitations of each approach are [24].

2.3.1 Mobility Management at the Link Layer

Mobility support at the link layer is always required no matter at what layer mobility management is implemented. It provides the functionality needed to detect and attach to a new network as well as routing data to the MN. However, it only provides support for mobility management within the same wireless technology.

If mobility is only managed at the link layer, ad-hoc routing schemes such as DSDV [25] can be used to allow MNs to form ad-hoc networks. To configure the network layer, protocols such as Dynamic Host Configuration Protocol (DHCP) or IPv6 auto-configuration can be used. Although this provides some mobility management support, it still does not provide the MN with a globally usable address, a requirement for mobility management across different domains. As a result, it is not sufficient to rely only on the link layer when roaming is across diverse types of physical technologies (e.g. Bluetooth [26] and IEEE 802.11 [27]).

2.3.2 Mobility Management at the Network Layer

Unlike mobility management at the link layer, which is tightly coupled with the wireless technology, mobility management at the network layer is independent of the underlying wireless technology. This is because globally usable addresses are assigned at this layer. Further, because the network layer is in the middle of the protocol stack, implementing mobility management at this layer provides mobility support to all higher layers. The main limitation of network layer solutions is the fact that they do not provide seamless handoffs, where there are no interruptions and data loss. This issue can be resolved using cross-layer solutions where the network layer can use information provided by other layers to eliminate the delays associated with the handoff process. However, this requires modifications of the other layer protocols which might not be feasible. Another possible solution is to provide algorithms to minimize these delays, which is the goal of this research.

The only mobility management protocol that has been standardized by the IETF is Mobile IP [1,2]. This protocol relies on having a mobility management entity that keeps track of the location of the MN and forwards data destined to it when needed. Mobile IP can be used to manage both macro-mobility and micro-mobility. Some protocols have been proposed to manage micro-mobility with the help of Mobile IP to manage macro-mobility. This will be discussed in more detail in chapter 3.

2.3.3 Mobility Management at the Transport Layer

When mobility is implemented at the transport layer, there is still a need to configure the network layer. Protocols such as DHCP [28] or IPv6 auto-configuration [1] can be used for that purpose. Once the network layer is configured and IP addresses are obtained, the transport layer bindings must be updated for existing connections. Because routing is done below the transport layer (in the network layer), to provide location management, there is a need for name to address mapping. This can be done using a higher layer protocols such as DNS [29]. The remaining mobility management functionality required by the transport layer is to provide dynamic rebinding of the MN's IP address. SCTP [30] as well as cellular SCTP [31] are example transport layer protocols that can be used to provide such functionality.

Transport layer solutions for mobility management can provide seamless handoffs since it is the transport layer at which congestion control is done. For example, functions such as sending rate can be utilized to reduce data loss. However, mobility management at the transport layer requires significant protocol modification as well as a considerable cooperation from other layer for location management.

2.3.4 Mobility Management at the Session Layer

Session layer mobility management can either be used as a stand a lone solution for location and handoff management or in conjunction with a transport layer mobility management scheme to provide location management only. When mobility is managed only at the session layer, after the network layer is configured and an IP address is obtained, the session layer initiates a new transport connection for the MN.

Like transport layer mobility management schemes, managing mobility at the session layer can provide seamless handoffs. The main disadvantage of this approach though is that it requires internet applications to deploy mobility management functionality required to create and use transport layer connections.

An example of session layer mobility management protocols is SLM [32].

3 A SURVEY ON LAYER-3 MOBILITY MANAGEMENT PROTOCOLS

3.1 Introduction

As an MN moves in a wireless network, its address changes in a corresponding manner. To ensure reachability of the MN at all times, there exists a need for a management entity that allows MNs to keep the same IP address regardless of their location and the type of connection used. This requires the MN to have two addresses: one permanent address by which all nodes of the network can identify the MN, and a temporary address that changes as the MN location change.

Because an MN is temporarily not reachable during a handoff, it is important that a mobility management protocol provides a mechanism to minimize the delay associated with the handoff process.

In the case of micro-mobility, the MN roams within the same domain and is therefore likely to experience frequent handoffs. Under these conditions, it is crucial to provide mechanisms to minimize delay. As a consequence, many protocols have been proposed to specifically address micro-mobility management. Such protocols rely on a base protocol such as Mobile IP to manage macro-mobility while they manage micro-mobility. Such protocols reduce the number of registrations performed by an MN by allowing the MN to avoid having to register with the home agent (HA) every time it moves within the same domain. In this case, the local AP keeps track of the MN movement and its current location. These protocols can be grouped into different categories based on their architecture and the type of issues they attempt to address.

This chapter gives an overview of the main mobility management architectures. A discussion will be provide on Mobile IP and micro-mobility specific protocols, on the criteria for comparing different protocols, and on the difference between micro-mobility management using Mobile IPv6 versus micro-mobility protocols.

3.2 Mobile IP

Mobile IP is the only mobility management protocol that has been standardized by the IETF. Because Mobile IP is a layer-3 management protocol, it allows seamless roaming between different types of networks such as third-generation (3G) networks, Bluetooth, or IEEE 802.11-WLAN. Typically, micro-mobility management protocols rely on Mobile IP to manage macro-mobility.

Mobile IP manages mobility on the network level to allow users to keep the same permanent IP address regardless of their location and the type of connection used. It redirects packets sent to the node using its IP address so that they are forwarded to its current address allowing the node to be mobile.

In help manage mobility; it is required in Mobile IP that the MN has two addresses. One of them is permanent and is used by higher level protocols and the other is temporary and is used to indicate the current location of the MN (called Care of Address (CoA)).

Thus far, there have been two important versions of Mobile IP: Mobile IPv4 which is based on IPv4 and Mobile IPv6 which is based on IPv6. Mobile IPv4 requires three functional entities in the network: the home agent (HA), the foreign agent (FA), and the mobile node (MN). Mobile IPv6, on the other hand, requires only two: the HA and MN.

3.2.1 Mobile IPv4

In Mobile IPv4, the HA keeps track of all MNs with permanent addresses on its network. For these MNs, it also keeps track of the current temporary address, the CoA. Whenever an MN moves and changes its CoA, it registers with the HA to inform it of its new location (Figure 8). This is done by sending a binding update via the Internet Control Message Protocol (ICMP) [33]. When a node needs to correspond with the MN, it sends packets via the home network. The HA then intercepts packets coming from the correspondent node (CN) and forwards them to the MN at its CoA through the current FA. This process is known as tunneling. In Mobile IPv4, all traffic destined to MN has to be tunneled via the HA. This traffic triangulation (Figure 9) unnecessarily increases the load on the HA, consumes more bandwidth and increases latency.

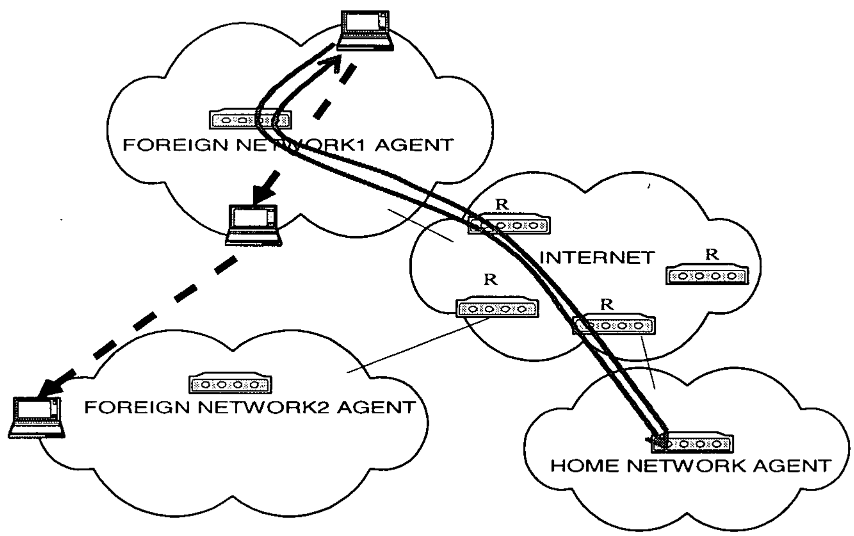


Figure 8: The registration process in Mobile IPv4 where the MN registers with its HA after obtaining a CoA from the FA in the foreign network that it's visiting.

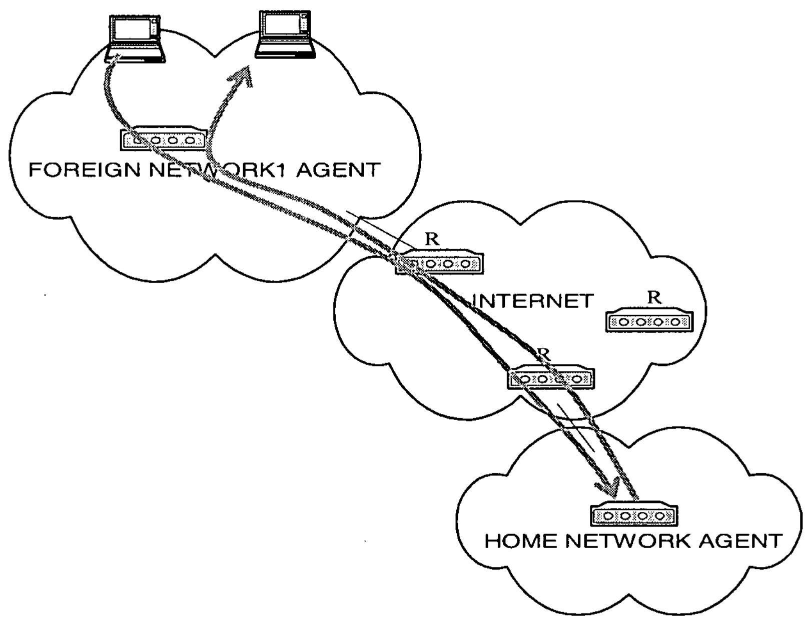


Figure 9: Intra-domain traffic and triangular routing

3.2.2 Mobile IPv6

Mobile IPv6 is an enhanced version of Mobile IPv4. It is integrated into IPv6 and provides many new improvements. For example, it eliminates the need for the FA entity. When visiting a foreign network, an MN obtains a CoA using address auto-configuration [1], and then registers this address with the HA. Any packets sent to the MN will initially be sent to the MN at its permanent address. The HA then intercepts these packets and tunnels them to the MN. After this, the MN sends the CN a binding update allowing the CN to correspond directly with the MN (Figure 10). This eliminates the triangular routing problem present in Mobile IPv4.

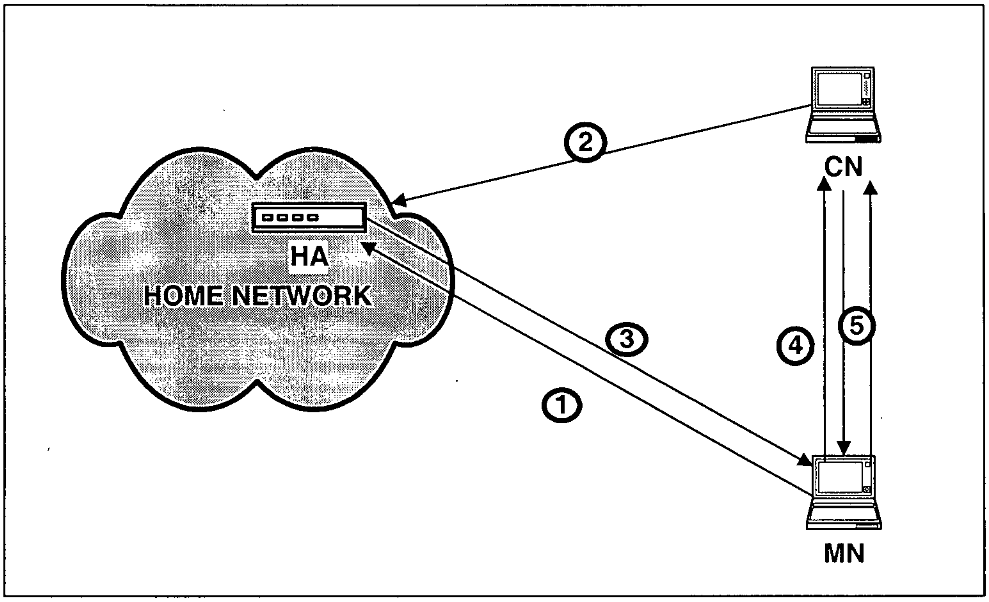


Figure 10: Communication between an MN and a CN in Mobile Ipv6 follows several steps: The MN registers its CoA with the HA (1). Packets destined to the MN are initially sent to the MN's permanent address (2). Packets are tunneled by the HA to the MN (3). The MN can then correspond with the CN directly (4 and 5).

Mobile IPv6 has the following improvements over Mobile IPv4 [1]:

- *Intra-domain Traffic:* Mobile IPv6 does not suffer from the triangular routing problem because it has built in support for "Route Optimization" [34].
- *Addressing:* IPv6 has a large address space (i.e. 128 bytes) that allows having the MN's CoA in the header of the IP packets that the MN send. Mobile IPv6 uses this feature to allow the MN to use its CoA as it source address in the IP header of the packets it sends allowing these packets to be processed by ingress filtering [35] routers. Allowing the use of the CoA as the MN's source address also simplifies routing of multicast packets sent by the MN since it allows the home address to be used but still be compatible with multicast routing that is partially based on the packet's Source Address [1].

- ***Eliminating the need for FAs:*** In Mobile IPv6, an MN visiting a foreign network can obtain a temporary address or CoA either through the IPv6 Neighbor Discovery [36] or Address Auto configuration. As a result, there is no need any more for FAs.
- ***Security:*** Mobile IPv6 utilizes IPsec [37], [38], and [39]. This includes sender authentication, data integrity protection, and replay protection for Binding Updates. Mobile IPv4 on the other hand, defines its own security mechanisms that are based on statically configured "mobility security associations" [1].
- ***Move-Detection:*** The movement detection mechanism in Mobile IPv6 provides bidirectional confirmation of a mobile node's ability to communicate with its default router in its current location.
- ***Inter and Intra-domain Traffic:*** Mobile Ipv6 reduces latency and saves bandwidth by avoiding IP encapsulation as much as possible. A side from the first few packets that have to be intercepted and tunneled by the MN's HA, all packets sent to the MN while away from home are tunneled using an IPv6 routing header rather than IP encapsulation. Also, as explained earlier, Mobile Ipv6 eliminates the need for triangular routing.
- ***Robustness and Simplicity:*** The use of Neighbor Discovery improves the robustness of the protocol and simplifies implementation of Mobile IP. Further, the dynamic HA address discovery mechanism in Mobile IPv6 uses IPv6 "anycast" and returns a single reply to the MN instead of returning a separate reply from each HA as in Mobile IPv4. This is more efficient and more reliable.

3.3 Micro-mobility Specific Architectures

Micro-mobility protocols [22] manage intra-domain mobility. They rely on a protocol such as Mobile IP to manage macro-mobility. As a result, when an MN roams within the same domain, it does not have to register with the HA every time it changes its local AP. This amounts to savings in the registration process (i.e. binding delay) regardless of the methodology of movement tracking used by the micro-mobility protocol (this could be an update in the routing caches such as in Cellular IP [40, 41, 42,43] or registration with a local AP such as in Mobile IPv4 Regional Registration [44]).

3.3.1 Mobile IP Extensions

These protocols are extensions to Mobile IP and manage mobility through a hierarchy of FAs. The FA that connects the domain to the Internet is called Gateway Foreign Agent (GFA). The following are examples of these protocols:

3.3.1.1 Mobile IPv4 Regional Registration (RR) [44]:

When the MN changes its location in RR, only the FA that's directly connected to it in the hierarchy keeps track of this movement.

3.3.1.2 Transparent Hierarchical Mobility Agents (THEMA) [45]:

This protocol separates the HA and FA functionality from the link layer. To reduce packet loss during a handoff, the protocol allows the link-layer termination point to play the role of a surrogate agent, which tunnels packets to the real mobility agent (MA) and de-tunnels packets that arrive from the MA. The creation of the tunnel is transparent to the MN.

3.3.1.3 Fast Handoff [46]:

Fast handoff is similar to RR; however, it reduces the move-detection time by interacting with the radio layer to detect a handoff before it occurs. The protocol assumes that it is possible to obtain a signal from the radio layer indicating that a handoff is about to happen as well as providing the address of the new AP. This way, the MN can register with the new AP through the old AP before the handoff occurs.

3.3.1.4 Agent Assisted Handoff [47]:

Agent Assisted handoff is similar to RR, except it lacks the multi-levels hierarchy. Similar to the Fast Handoff protocol, it speeds up the handoff process by anticipating the handoff through interacting with the radio layer. This handoff, however, is initiated by the old and new FAs. After a short period of negotiations, one of the FAs sends a regional registration to the GFA on behalf of the MN. During the handoff, traffic destined to the MN can be bicasted to the two FAs. Once the MN is in the new network, the new FA sends an RA so that the MN can perform a normal registration.

3.3.1.5 TeleMIP [48]:

TeleMIP is similar RR, however, it supports only a two-level hierarchy and implements a load balancing algorithm that allows FAs to be connected to more than one GFA.

3.3.2 Host Based Routing

In a host-based routing architecture, the main mobility management entity is called the Mobile Agent (MA). MAs maintain a routing cache that contains the next hop to reach a given MN and the next hop to reach the gateway. The routes from the MN to the gateway and vice versa are established and maintained by the hop-by-hop transmission of two advertisement packets from the MN and the gateway. The MAs update their routing cache when these advertisement packets are received. The following are examples of Host Based Routing architecture protocols:

3.3.2.1 Cellular IP (CIP) [40,41,42,43]:

A cellular IP domain is composed of several MAs. One of the MAs acts as an FA and as a gateway to the Internet. The MAs are built on IP; however, IP routing is replaced by CIP routing and location management. In CIP networks, the MN transmits a route update packet to the gateway after the radio handoff is completed to establish new routes. To reduce handoff latencies, the protocol defines a "semi-soft" handoff that is based on interactions with the radio layer to detect a handoff before it occurs. Once an MN detects that a radio handoff is about to happen, it can send a special packet to request bi-casting traffic to the old and new APs.

3.3.2.2 Handoff Aware Wireless Access Internet Infrastructure (HAWAII) [49,50]:

HAWAII supports a hierarchical network topology. It defines two different handoff mechanisms that are based on the reception of a trigger from the radio layer. These mechanisms are adapted to different radio technologies that can be chosen to optimize various aspects of the network such as handoff latency, packet losses or packet reordering.

3.3.3 Multicast Based Routing

In Multicast based routing protocols, an MN is assigned a multicast-CoA. When visiting a foreign network, the MN joins this CoA through the local router in that network.

Packets destined to this MN are always sent to the multicast CoA allowing the MN to receive these packets at all locations that subscribe to this address. Multicast protocols do not use bandwidth and network resources efficiently. This is because multicast packets will be sent to and processed by all routers belonging to the multicast address. Examples of Multicast based routing protocols include the following:

- Dense Mode [51,52,53,54]
- Sparse mode [51,55]
- Efficient micro-mobility Multicast based architecture [56].

3.3.4 MANET Based

MANET based protocols were originally designed to manage mobility in ad-hoc networks. In this architecture the routing is multi-hop based. Routes are updated as the MN moves and network connectivity changes. The following are examples of MANET based architecture protocols:

3.3.4.1 Edge Mobility Architecture (EMA) [57]:

EMA is a prefix-based routing protocol. It routes packets based on the prefix. This prefix represents IP blocks that are assigned to each Access Router in the domain as well as routes to the MN. The handoff is initiated based on the reception of a handoff trigger signal from the radio layer.

3.3.4.2 Ad-Hoc on Demand Distance Vector (AODV) [58]:

AODV builds and maintains routes between nodes only as needed by source nodes. The protocol also forms tree network topologies that connect multicast group members.

3.4 Protocol Comparison

3.4.1 Comparison Criteria

Whether managing micro-mobility or macro-mobility, there are several issues that mobility management protocols should be concerned with. Section 2.2.1 discussed these issues. The following is a summarized list:

- *Efficient handoff management:* This is the most important issue in mobility management since handoff latencies result in session interruptions and packet loss.
- *Simplicity (or network design requirements):* A simple network design reduces development complexity and allows for easier future enhancements.
- *Scalability and robustness:* It's important to maintain robustness and reliability as the network grows.
- *Security:* Protocols need to provide security mechanisms to prevent threats against mobility management mechanisms.
- *Power:* Because mobile devices have limited battery life, it's preferred that a mobility management protocol provides a mechanism such as a power-save mode to save power when possible.

3.4.2 Mobile IPv6

- *Handoff Management:* Unlike Mobile IPv4, Mobile IPv6 provides an efficient handoff management mechanism in which only the first few packets in a session have to be tunneled via the HA. This allows direct communications with the CN for the rest of the session or until the MN switches its point of attachment to the network. This mechanism limits tunneling and, as a result, reduces triangulation latencies. In the case of micro-mobility, however, the protocol generates more signaling traffic when compared to other micro-mobility protocols since the MN has to register with its HA every time it moves within a domain. Both Mobile IPv4 and Mobile IPv6 do not provide a mechanism to reduce handoff latencies.

- *Simplicity:* Mobile IPv6 is simpler to implement when compared to micro-mobility protocols since the same protocol handles both macro-mobility and micro-mobility.
- *Scalability and Robustness:* Avoiding single points of failure by having a flat hierarchy makes the protocol more scalable and reliable. Further, Mobile IPv6 enhances reliability and robustness through the support of multiple HAs in the home network. If an HA fails, any MN associated with this FA can contact other HAs through the Neighbor Discovery algorithm.
- *Security:* Because IPv6 has a large address space (128 bits), there is room to have the MN's CoA in the header of the IP packets that the MN sends. As a result, these packets can be processed by ingress filtering [59] routers. Further, Mobile IPv6 allows bi-directional confirmation for binding updates and utilizes IPsec [37], [38], and [39] which include sender authentication and data integrity protection.
- *Power:* Paging extensions have been added to Mobile IPv6 [60] that allow the MN to be in power-save mode in order to save on battery power consumption.

3.4.3 Micro-mobility Protocols

- *Handoff Management:* Micro-mobility protocols allow the MN to avoid having to register with the HA when roaming within the same domain. This reduces the amount of signaling in the handoff and registration latencies.
- Fast Handoff, Agent Assisted Handoff, CIP, HAWAII, and EMA reduce handoff latencies associated with the handoff by interacting with the radio layer to anticipate the handoff. This allows the IP handoff to start before the radio handoff occurs thus reducing the amount of packet loss. Other Micro-mobility protocols do not address this issue.
- *Simplicity:* Micro-mobility protocols that are based on Mobile IP tend to be simpler to implement than others. On the other hand, all Micro-mobility protocols introduce more complexity to the network since they use two protocols to manage mobility (i.e. Mobile IP to manage Macro-mobility and another protocol to manage Micro-mobility).
- *Scalability and Robustness:* Protocols that have a hierarchical network structure such as RR, CIP and HAWAII have single points of failure which is a major issue for scalability and robustness.
- EMA relies on TORA to manage mobility which is an ad-hoc network protocol that provides more than one route to each destination. As a result, many nodes maintain redundant routes about other nodes or subnets. Most of these routes are not needed since a large part of the network is wired. As a result, the tables maintained by each node become very large causing unnecessary load as the network grows. This represents a major scalability issue for this protocol.

- *Security:* Most Micro-mobility protocols rely on existing security mechanisms in Mobile IP; however, protocols such as HAWAII and MMP, in which the AP sends a registration requests on behalf of the MN, do not include a Mobile-Home Authentication extension [51].
- *Power:* Mobile IPv4 Regional Registration, Cellular IP and HAWAII define paging mechanisms that allow the MN to be in a power-save mode in order to save on battery power consumption. The other discussed micro-mobility management protocols do not define such mechanisms.

3.4.4 Conclusions

The micro-mobility protocols Fast Handoff, Agent Assisted Handoff, CIP, HAWAII, and EMA try to reduce handoff latencies by anticipating a handoff using triggers from the radio layer or the link layer. On the other hand, all of these protocols, with the exception of Agent Assisted Handoff, suffer from scalability and reliability issues making it difficult to deploy these protocols in large networks.

Agent Assisted Handoff does not have a multiple level hierarchy and, as a result, does not suffer from these scalability issues. Further, Agent Assisted Handoff defines a network initiated handoff which is faster than the MN initiated handoff. Like other Mobile IP based protocols discussed, this protocol is not compatible with Mobile IPv6. As a result, it suffers from a major drawback in that it does not benefit from all the improvements introduced to Mobile IPv6.

None of the protocols discussed suggest any mechanism to reduce DAD latencies. These latencies can be reduced by not delaying the DAD process and performing it in parallel with other steps in the handoff [1]. Further, if the MN determines that the overhead of performing DAD outweighs its benefits, it can choose not to perform DAD [1].

In general, the use of micro-mobility protocols introduces new complexity to the network design. Further, most micro-mobility protocols that try to reduce handoff latencies are not compatible with Mobile IPv6 and do not benefit from all the improvements of this protocol. As a result, a simple solution to handoff management in the case of micro-mobility would be the use of Mobile IPv6 but with optimizations that address handoff latencies.

4 CHARACTERIZATION OF THE HANDOFF PROCESS

4.1 Introduction

This chapter discusses the characterization process of the handoff in Mobile IPv6 and in IEEE 802.11 networks using a simulator that is based on OPNET®. The results of the characterization are then presented in the form of a detailed breakdown of the delay contributions associated with a handoff in Layer-2 and Layer-3.

4.2 Layer-2 (IEEE 802.11) Handoff

4.2.1 Move-Detection

This is when the MN discovers that it has moved to a different location. The procedure in this stage is affected by whether the handoff was network initiated or MN initiated. In the first case, the MN discovers that it has moved when the AP sends it a disassociation message. The later case, however, is the most common one, although it is not as fast. When handoff is MN initiated, movement can be detected in two different ways depending on whether or not it is transmitting at the time of the handoff. If the MN is transmitting frames, it will detect its movement after a number of failed frame transmissions. On the other hand, if the node is not transmitting at the time of handoff, it relies on beacon reception to detect its movement. If the node does not hear a certain number of beacons, it assumes that it has moved. In both cases, it is critical to set the movement detection parameters properly since they affect the length of the move-detection phase and other relevant factors. For example, the number of failed transmissions that indicates move-detection should take into consideration collision as a factor for failure. Velayos and Karlsson [9] reported that 3 consecutive collisions is a rare event, so 2 consecutive retransmissions should eliminate collision as a possible reason for failure. In the second case, the move-detection time depends on the beacon interval. A short beacon interval will reduce the move-detection phase but will consume more bandwidth and will need more processing power. Currently, the standard beacon interval is 100ms. Velayos and Karlsson showed that this beacon interval can be reduced to 60ms without noticeably increasing used channel capacity.

4.2.2 Searching for a new AP

Once the MN detects that it has changed its location, it starts looking for a new AP to associate to. The IEEE 802.11 standard defines two search mechanisms: active scanning, and passive scanning. In passive scanning the MN listens to beacons that are sent periodically by APs. These beacons provide both timing and advertising information. Once it scans all channels, the MN uses the information obtained from beacon frames and creates a prioritized list of the available APs to elect from. Passive scanning is not efficient compared to active scanning. In passive scanning, the periods that the MN uses to listen to beacons should be longer than the beacon period. However, the MN does not know this beacon period until it reads the timing information from the first beacon it receives. Once it identifies that, a MN waits for the rest of the beacon period before switching to a new channel to allow for beacons that might be received from other APs from different WLANs that are using the same channel. Further, since the standard specifies that the whole set of available channels must be scanned; the MN has to wait until it receives beacons from all channels. This means that, the MN has to wait for: (beacon period * number of channels) seconds

Active scanning is more efficient since the MN does not have to wait for beacons. It sends probe requests on each channel, sets a timer, and waits for responses. The wait time is bounded by two values: minChannelTime and maxChannelTime. If no response has been received by minChannelTime, the MN abandons the channel and starts scanning the next channel. If one or more responses are received by minChannelTime, the MN waits until maxChannelTime. It then stops accepting probe responses and starts scanning the next channel. After all channels have been scanned, the MN processes all the responses it received and selects an AP to associate to based on the information available in these responses. The values of minChannelTime and maxChannelTime have not been specified by the standard.

4.2.3 Switching to a new AP

Once the MN finds an AP to associate with, it starts the switching phase. This phase involves two steps, the authentication and reassociation to the new AP. The MN first sends an authentication request to the AP. The AP responds to this request indicating acceptance or rejection. There are two authentication schemes specified by the IEEE 802.11 standard: open system and shared key. Open system is the default and requires no authentication procedure. Shared key specifies a four step handshake authentication procedure. Once the MN has been authenticated, it sends a reassociation request to the AP. The AP responds by sending a reassocaiation response indicating acceptance or rejection.

4.3 Layer-3 (Mobile IPv6) Handoff

4.3.1 Move-Detection

A MN listens to RAs either periodically or by sending a solicitation request. The move-detection in Layer-3 occurs when the MN does not receive a certain number of RAs. Once Layer-2 handoff is complete and a connection has been established, the MN can receive RAs again. It learns the address of the new router by analyzing these RAs [5].

4.3.2 Handoff Execution

Once an MN detects that it has moved, it performs DAD to verify the uniqueness of its new link local address. It then generates a new COA that is based on the information contained in the RA. This is accomplished through stateless or statefull address autoconfiguration [61]. Once the COA is obtained, the MN then performs DAD. To do this, the MN sends a number of neighbor solicitation messages to its new address and waits for at least one second to hear a response. The initial solicitation message is delayed by a random factor between 0 and a preset maximum solicitation delay value. Once the COA is formed, the MN can update its HA and any CNs of its new location. This is done through the binding update process. The MN can also request an acknowledgment by setting a flag in the binding update [5].

4.4 Handoff Delays and Existing Bottlenecks

4.4.1 Experimental Setup

Experimentation is performed using OPNET® Simulator v.11. It implements a realistic scenario where 36 MNs roam randomly amongst 9 APs (Figure 26). This model extends an existing scenario provided in OPNET®. The number of MNs in any one handoff varies from 1 to 4 MNs.

The network is configured as follows:

- The Physical Layer (i.e. Layer-1) is implemented using Direct Sequence.
- Link Layer (i.e. Layer-2) is implemented using IEEE 802.11.
- Network Layer (i.e. Layer-3) is implemented using Mobile IPv6.
- All MNs run four applications simultaneously: Voice over IP, Email, Telnet and FTP.
- Initially each AP serves 4 MNs, for which the initial access point will be their HA.
- The channel bandwidth used is 11Mbps.
- The RA interval has a uniform distribution with a minimum outcome of 0.5s and a maximum outcome of 1s.
- The number of missed RAs that indicate a Layer-3 move-detection is 3.
- The Beacon Period is 60 ms
- The number of missed beacons that indicate a Layer-2 move-detection is 2.
- Layer-2 scanning is implemented using passive scanning.

4.4.2 Simulator Design

The simulator extends existing implementations of IEEE 802.11 and Mobile IPv6 in OPNET®. The next subsections describe the simulator design.

4.4.2.1 Layer-2: IEEE 802.11

Figure 27 shows the simulator design for IEEE 802.11. This extends the existing implementation of the protocol in OPNET®. The main modifications to this implementation were the following:

- The addition of the "SWITCH" state (shown in Figure 27). This state implements the functionality needed for simulating the switching action from one AP to another. Current implementation of IEEE 802.11 in OPNET® does not consider the association request and association response frames exchange between the MN and the new AP, which is part of the procedure of a handoff in IEEE 802.11.
- The addition of a performance evaluation tool that times the various operations involved in the handoff process. Implementation of this tool was done by embedding its functionality into the various states of the state machine. The timing results measured for various operations at the different states were also manipulated by this tool to calculate the final required performance metrics for the handoff process in IEEE 802.11.

4.4.2.2 Layer-3: Mobile IPv6

Figure 28 shows the simulator design for Mobile IPv6. This extends the existing implementation of the protocol in OPNET®. The main modification to this implementation is the addition of a performance evaluation tool that times the various operations involved in the handoff process in layer-3 using Mobile IPv6. Implementation of this tool was done by embedding its functionality into the various states of the state machine. The timing results measured at the different states were also manipulated by this tool to calculate the final required performance metrics for the handoff process in Mobile IPv6.

4.4.3 Results

Figure 11 shows a breakdown of delays for various phases of the Layer-2 handoff. The x-axis is the handoff number (i.e. handoff in order of occurrence) while the y-axis is a delay breakdown of the three phases of Layer-2 as discussed in Section 4.2. The delay associated with the switching phase has been shown to be approximately 1ms [9,10]. As a result, they have been modeled as a delay of 1ms. This delay is negligible compared to the delays associated with the search and switching phases. From Figure 11, it can be seen that the search phase contribution is an average value of 75% of the overall Layer-2 delay. These results seem to differ from what is presented in [11] but are in line with what is presented in [9] and [10].

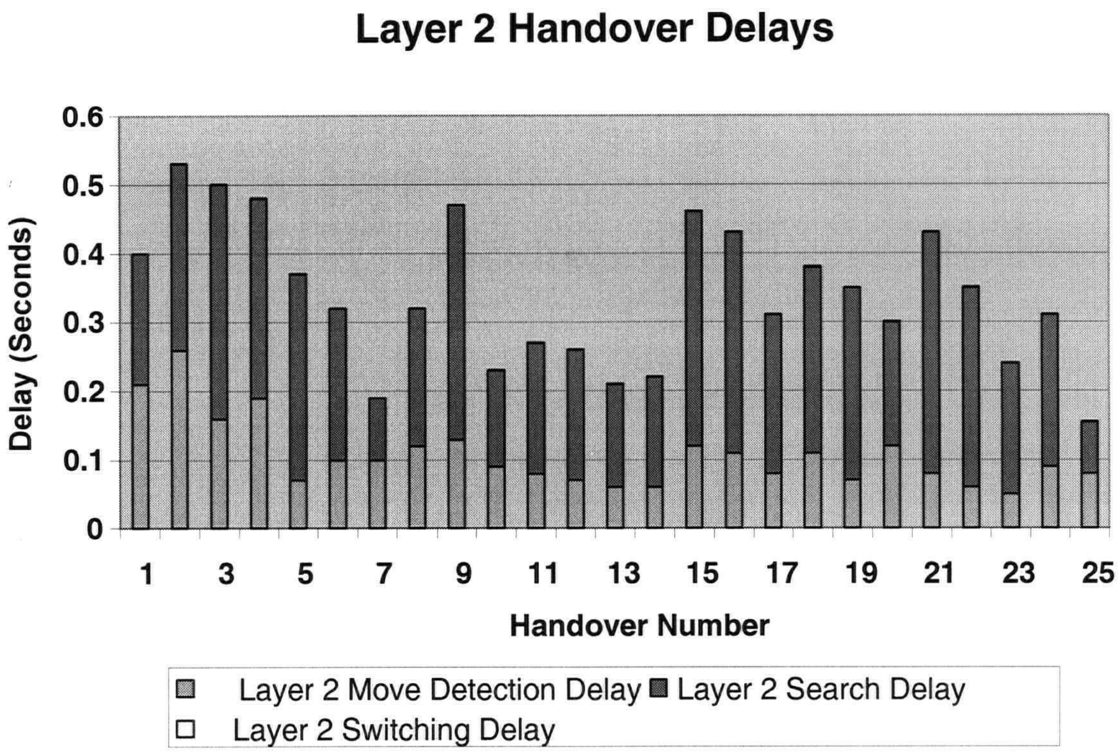


Figure 11: Layer-2 handoff delays

Figure 12 shows a breakdown of Layer-3 handoff delays. The x-axis is the handoff number and the y-axis is a delay breakdown of the two phases of Layer-3 handoff as discussed in Section 4.3. This plot shows that on average, the delay associated with move-detection is 63% of the total Layer-3 delay.

Figure 13 shows a comparison of Layer-2 handoff delays to Layer-3 handoff delays. This plot shows that on average, Layer-3 contributes over 90% of the total handoff delay. This means that the major contributor to the overall Layer-2 and Layer-3 handoff delay is the Layer-3 move-detection phase which, on average, contributes 57% of the total handoff delay. The second largest contributor is the binding delay and the DAD process which, on average, contributes 33%. These comparison results of Layer-2 vs.

Layer-3 handoff delay are different than what is presented in [11] since they considered Layer-2 handoff delay as part of Layer-3 handoff delay.

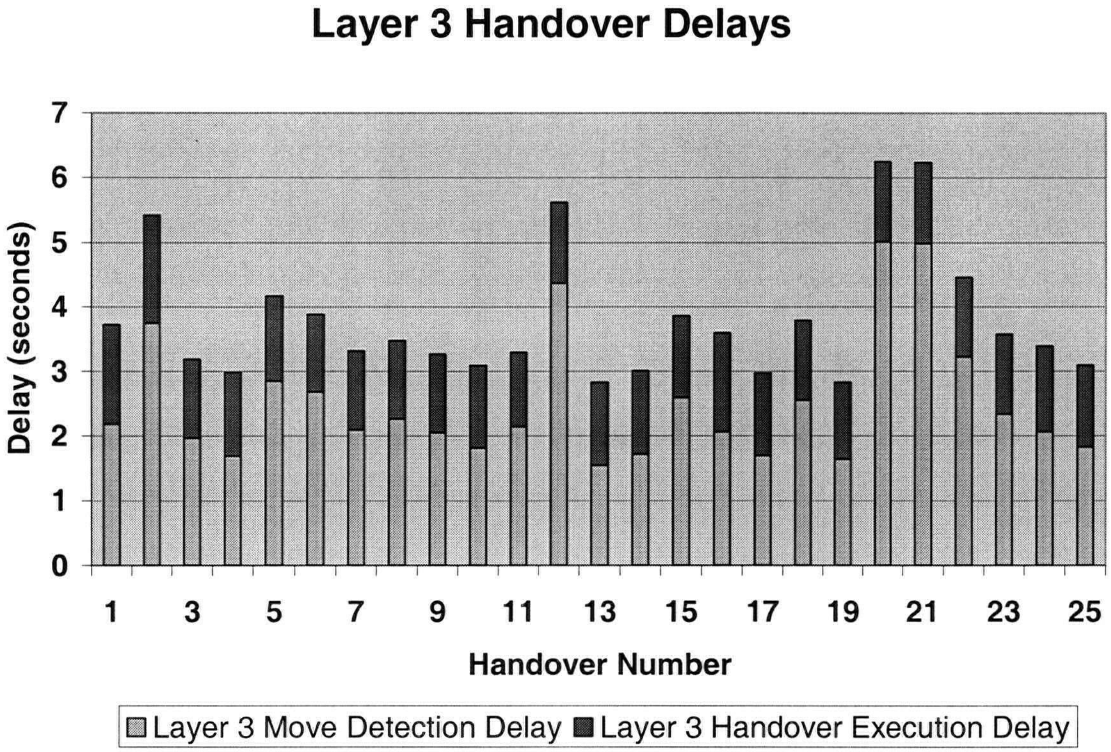


Figure 12: Layer-3 handoff delays

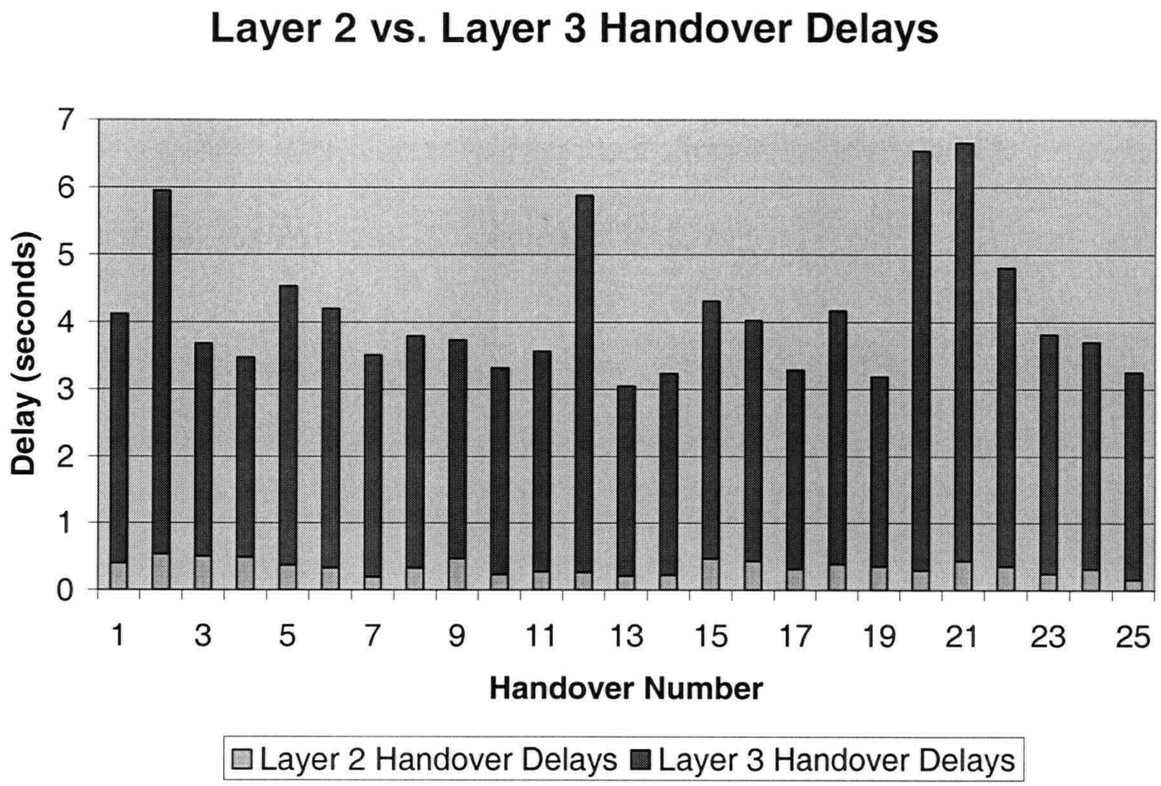


Figure 13: Layer-2 vs. Layer-3 handoff delays

5 A NOVEL HANDOFF INITIATION METHOD TO ADDRESS HANDOFF DELAYS

5.1 Introduction

In the previous chapter we characterized the handoff process for Mobile IPv6 in IEEE 802.11 wireless networks [27] showing that, on average, Layer-3 contributes more than 90% of the handoff delays, while Layer-2 contributes less than 10%. Further, we showed that the most prominent source of delay is Layer-3 move-detection accounting for almost 60% of the combined Layer-2 and Layer-3 handoff delays.

It is known that move-detection delay is directly dependent on the length of RA. Due to the limited bandwidth, care must be taken when reducing the RA interval when trying to reduce the move-detection delay. This chapter studies the move-detection in a realistic scenario that considers the move-detection delay in the context of recurrent handoffs. It provides an analysis of move-detection in terms of bandwidth and cell-switching behavior and proposes an adaptive method for minimizing move-detection delay given certain bandwidth requirements.

5.2 The Move-Detection Problem in Recurrent Handoffs

The motion of an MN amongst overlapping cells is sporadic. Motion can be in a straight line, from one cell to another, or it can be along the boundary of two cells. Further, motion can change in any direction at any time. Aside from physical motion, the signal strengths of adjacent cells continuously fluctuate with time due to noise factors. When combined, these factors contribute to the sporadic entry, exit, and re-entry of an

MN into adjacent cells that overlap. For ECS (where the MN switches to a new router through a Layer-3 handoff as soon as it starts to receive RA from the new router), this results in somewhat "meta-stable" behavior of handoffs between cells. As a result, ECS is likely to result in recurrent switching between the two cells. . As an example, Figure 14 shows the overlap of two adjacent cells. An MN at location "a" is clearly in Cell 1 from a signal strength perspective; however, an MN at location "b" is exposed to almost equal signal strength from both Cell 1 and Cell 2. As a result, ECS is likely to result in frequent switching between the two cells.

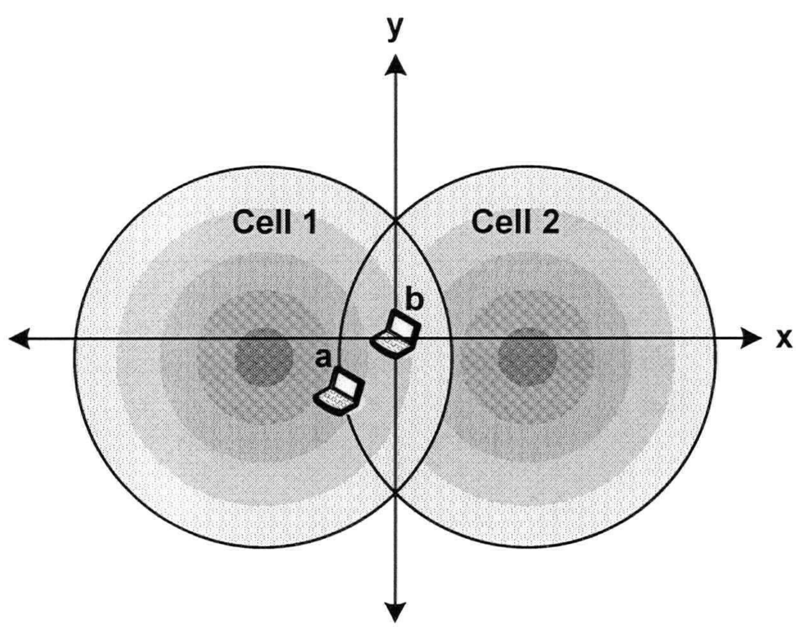


Figure 14: In "real" networks, the transition between two cells is not clearly distinguishable. As a result, the node at location "b" is likely to result in recurrent handoffs when ECS is used.

To model this, we propose that all recurrent cell switching behavior (switching rate) can be modeled as a single random variable with a uniform distribution. The inverse of this random variable (ρ) is the cell switching interval θ, i.e. $\theta = 1/\rho \in [\beta,\eta]$, where β and η are the minimum and maximum values respectively. This value, θ, represents the delay between consecutive cell switches.

5.3 Proposed Model

5.3.1 Assumptions

We assume a wireless network model in which an MN roams amongst a number of domains with overlapping coverage areas (or cells). The mobility analysis carried out here pertains to the IP layer (i.e. layer-3) and does not assume any specific link layer technology.

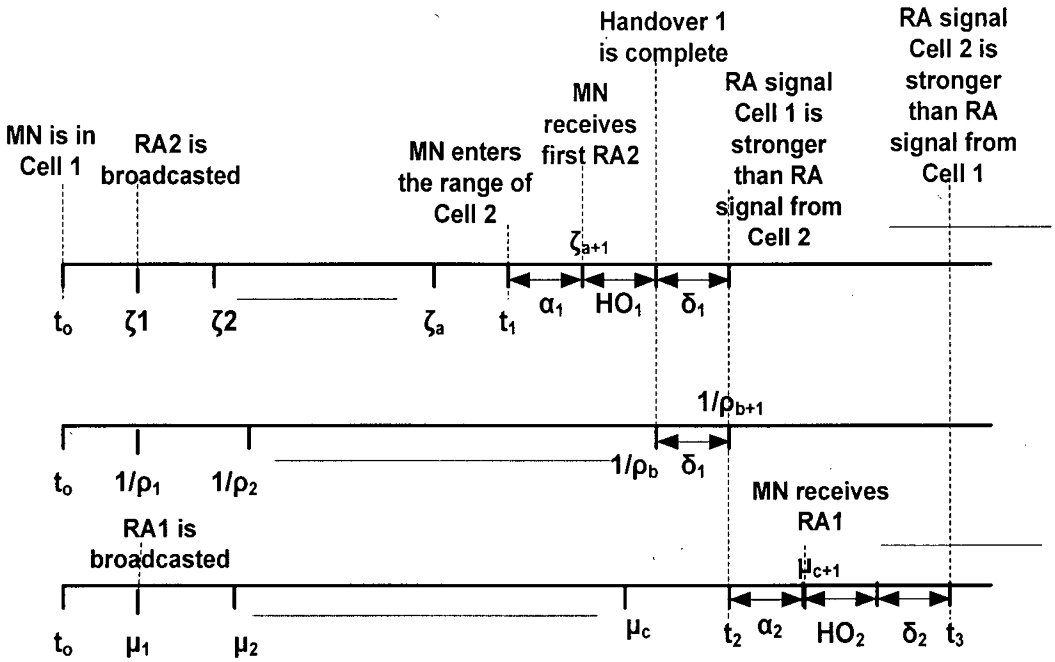
5.3.2 Modeling Move-Detection

The specifications in [36] suggest that the RA interval should be a random value that falls between a specified minimum and maximum and is uniformly distributed in that region. Further, it is recommended that these intervals issued are independent of each other thus preventing multiple nodes from transmitting at the same time.

Move-detection Delay is modeled as in [19] using Renewal theory. Assuming that the RA interval, ξ, has a uniform distribution in the region $[a,b]$, then the expected value for the move-detection delay, α, is given by,

$$E(\alpha) = \frac{a^2 + b^2 + ab}{3(a+b)} \tag{1}$$

5.3.3 Timing Behavior



ζ: RA2 interval (for Cell 2)
μ: RA1 interval (for Cell 1)
ρ: Switching rate
α_1: Move detection time of RA2
HO_1: Execution time for handoff with Router 2
δ_1:: Random delay after handoff with Router 2 is complete and before the switching process to Router 1
α_2: Move detection time of RA1
HO_2: Execution time for handoff with Router 1
δ_2:: Random delay after handoff with Router 1 is complete and before the switching process to Router 2

Figure 15: Timing diagram corresponding to consecutive handoffs between adjacent cells: The top time line represents the sequence of events for a handoff from router1 in Cell 1 to router 2 in Cell 2. The second time line represents the mapping of the random delay δ to the switching factor ρ. The bottom timeline shows the next handoff from router 2 back to router 1.

To determine the bandwidth requirements of an MN with recurrent cell-switching behavior, we must derive the relationship between consecutive handoffs between Cell 1 and Cell 2 which occur with a rate of ρ. This represents the canonical case of recurrent cell switching and is sufficient for deriving the relationships between bandwidth, the RA interval, and the cell-switching rate.

Figure 15 shows the timing relationship for two consecutive handoffs. In this figure, there are three timelines. The top timeline shows the sequence of events for a handoff from router1 in Cell 1 to router2 in Cell 2. The bottom timeline shows the next handoff from router2 back to router1. The middle timeline relates the two handoffs as a consequence of the MN switching between cells at a rate ρ.

The reference point t_o marks the time when the MN resides in Cell 1. From this point on, router2 transmits consecutive RA broadcasts $\xi_1, \xi_2, \ldots, \xi_{a+1}, \ldots$ such that ξ_{a+1} is the time of the first RA received by the MN after it enters Cell 2 at t_1. The time between t_1 and ξ_{a+1} is the move-detection latency for ECS and is specified by α_1. *HO* time includes layer-2 handoff, and layer-3 binding updates. After the time period *HO* + move-detection latency, the handoff is process is completed. Note that part of the *HO* time (i.e. layer-2 handoff) happens before layer-3 move-detection, and the other part (i.e. layer-3 binding updates) happens after layer-3 move-detection. For simplicity of representation, the *HO* time period is placed after the move-detection period in the timeline. After handoff is complete, the *MN* will begin to encounter RAs from Cell 1 after some random delay period, δ_1.

The time at which δ_1 ends corresponds to t_2 in the bottom timeline (when the MN re-enters Cell 1). In a similar way to the top timeline, RAs broadcasted by Cell 1 are shown in the bottom timeline as $\mu_1, \mu_2, \ldots, \mu_{c+1}, \ldots.$, where μ_{c+1} is the time when the first RA is received by the MN after entering Cell 1 at t_2. In this case, the move-detection time is denoted by α_2. We will model the switching interval, δ, as a random variable. To determine its expected value, we first start with the switching rate, ρ. We have a stochastic sequence $\{1/\rho_1, 1/\rho_2, ..., 1/\rho_b\}$ where $1/\rho_1, 1/\rho_2, ..., 1/\rho_b$ are Independent and Identically Distributed (IID) positive stochastic variables; and $1/\rho_b$ denotes the b^{th} switching interval.

We will use the notation δ instead of δ_1 in order to simplify the representation. Further, for simplicity, as mentioned in section 5.2, we will denote $1/\rho$ by θ. In a similar way to Section 5.3.1, Renewal Theory is used. If we set

$$J_b = \sum_{i=1}^{b} \theta_i ,$$

and

$$X_t = \sup\{b : J_b \leq t\}$$

then $\{X_t, t \geq 0\}$ is a renewal process. This is a process that renews itself on the occurrence of each event.

The random delay after handoff with router 2 (as shown in Figure 15), δ, for θ_{b+1} (i.e. ρ_{b+1}) can be expressed by,

$$\delta(t) = J_{X_t+1} - (t_1 + \alpha + HO) \qquad (2)$$

so that,

$$\delta(t) = J_{X_t+1} - t \qquad (3)$$

where $t = (t_1 + \alpha + HO)$. By renewal theory,

$$\lim_{t\to\infty} P(\delta(t)) < y) = \frac{\int_0^y (1 - F_\theta(s))ds}{E(\theta)} \qquad (5)$$

where $F_\theta(t)$ is the distribution function and $E(\theta)$ is the mean of θ. A formal proof of Equation (5) can be found in [62].

We can assume that $\delta(t)$ is convergent in distribution to δ, i.e.,

$$\lim \delta(t)_{t\to\infty} = \delta \tag{6}$$

then

$$P(\delta < y) = \frac{\int_0^y (1 - F_\theta(s))ds}{E(\theta)} \tag{7}$$

If $D(\delta)$ is the variance of δ, and $f_\delta(y)$ is the probability density function, then, by renewal theory [63]:

$$f_\delta(y) = \frac{dP(\delta < y)}{dy} = \frac{1 - F_\theta(y)}{E(\theta)}, \tag{8}$$

$$E(\delta) = \frac{E(\theta^2)}{2E(\theta)} = \frac{D(\theta) + [E(\theta)]^2}{2E(\theta)} \tag{9}$$

Assuming that the switching interval θ has a uniform distribution over $[\beta, \eta]$ where β denotes the maximum limit and η denotes the minimum limit, then,

$$E(\theta) = \frac{\beta + \eta}{2}, \tag{10}$$

$$D(\theta) = \frac{(\eta - \beta)^2}{12}, \tag{11}$$

$$f_\delta(y) = \begin{cases} \dfrac{2}{\eta + \beta}, & 0 \le y < \beta \\ \dfrac{2(\eta - y)}{\eta^2 - \beta^2}, & \beta \le y < \eta \\ 0, & otherwise \end{cases} \tag{12}$$

$$E(\delta) = \frac{\beta^2 + \eta^2 + \eta\beta}{3(\beta + \eta)} \tag{13}$$

Equation (13) expresses the expected value of δ as a function of the switching rate range. It is not surprising that this equation has the same form as Equation (1) [19] since both δ and α are renewal processes and are assumed to have a uniform distribution.

5.3.4 Router Advertisement and Move-Detection Delay Calculation

To determine the bandwidth requirements of the system, the expected number of handoffs in a given period of time, T, must be determined. From Figure 15, the expected number of handoffs can be calculated as:

$$E(N) = \frac{T}{E(\alpha) + E(HO) + E(\delta)} \tag{14}$$

where N is the number of handoffs. Further,

$$\frac{K}{b} \leq BW_{RA} \tag{15}$$

where, K is the length of a RA packet, and b is the upper limit of the RA interval. In other words, K/b must be bounded by BW_{RA}.

The total number of bits consumed is:

$$B = B_D + E(N) * B_{HO} + B_{RA} \tag{16}$$

where B is the total bits consumed, $E(N)$ is the expected number of handoffs, B_D is the number of bits consumed by ongoing data calls, B_{HO} is the number of bits consumed by handoff calls, and B_{RA} is the number of bits consumed by RAs. We know that:

$$B_{RA} = P * K$$

where P is the number of RAs in a time period T, and K is the number of bits in an RA.

Next, we will assume an RA interval of $[0,b]$ with ξ (i.e. the RA interval) uniformly distributed in this region. The value of the upper limit for the RA interval, b, is very important since it represents the minimum possible upper limit of the RA interval given a certain bandwidth (as in Equation (15)). If we were to use Equation (15) with the smallest possible value of b, the bandwidth consumed by the RAs then is:

$$BW_{RA} = \frac{P*K}{T} = \frac{P*K}{P*b} = \frac{K}{b}$$

The total bandwidth over period T can be calculated as follows:

$$BW = BW_D + E(N)*BW_{HO} + E(N)*BW_{RA},$$

$$=> BW = BW_D + E(N)*BW_{HO} + K/b \tag{17}$$

such that BW_D is the bandwidth consumed by ongoing data calls, $E(N)*BW_{HO}$ is the bandwidth consumed by handoff calls (i.e. N number of handoff calls), and k/b is the bandwidth consumed by RAs.

$$b = \frac{K}{BW - BW_D - E(N)*BW_{HO}} \tag{18}$$

Setting $a = 0$ in Equation (1), the expected value of α is,

$$E(\alpha) = \frac{a^2 + b^2 + ab}{3(a+b)} = \frac{b}{3} \tag{19}$$

Next, we will assume that HO follows a uniform distribution in region the $[m_1, m_2]$ such that its expected value is,

$$E(HO) = \frac{m1 + m2}{2} = m \tag{20}$$

Substituting Equations (13), (19), and (20), into Equation (14), we get the expected value of N:

$$E(N) = \frac{T}{\frac{b}{3} + m + \frac{\beta^2 + \eta^2 + \eta\beta}{3(\beta+\eta)}} \tag{21}$$

Now we can express the upper limit of RA, b, in terms of available bandwidth and the switching rate,

$$b = \frac{K}{BW - BW_D - E(N) * BW_{HO}}$$

$$b = \frac{K}{BW - BW_D - \frac{BW_{HO} * T}{\frac{b}{3} + m + \frac{\beta^2 + \eta^2 + \eta\beta}{3(\beta + \eta)}}} \tag{22}$$

The expected value of the move-detection delay in terms of available bandwidth and the switching-rate can now be expressed by substituting Equation (22) into Equation (19).

By making the following substitutions,

- $C_1 = 3 * K$
- $C_2 = 3(BW - BW_D)$
- $C_3 = 3 * BW_{HO} * T$
- $C_4 = 3 * m + \frac{\beta^2 + \eta^2 + \eta\beta}{(\beta + \eta)}$

Equation (22) becomes a quadratic equation as follows:

$$b^2 * C_2 + b * (C_4 - C_3 - C_1) - C_1 * C_4 = 0\,, \tag{23}$$

Using Equation (23), we can now solve for *b*. The value of *E(α)* can be calculated using Equation (19). The resulting representation of *b* gives the expected value of the upper limit of the RA interval for a given bandwidth and switching-rate. This value can be used to dynamically adjust the RA interval based on the cell-switching behavior and available bandwidth.

Equations (14), (21), and (22) take the cell switching rate into account when calculating the expected number of handoff calls. A degenerate case of this, when $\delta = 0$, is a straightforward way of extending single-handoff models (such as in [9]) to account for consecutive handoffs. For this case, calculating the bandwidth of multiple handoffs would require a simple summation of handoff delays (i.e. move-detection and handoff execution *HO*) but would exclude the random delay between consecutive handoffs.

6 MODEL VALIDATION

6.1 Introduction

This chapter validates the model proposed in Chapter 5 by providing an implementation of the model and an intuitive discussion of the various trends of the model. These trends are then compared to results from simulation. Further, detailed of the design and implementation of the simulator is provided.

Explicitly, validation includes two stages:

- In the first stage, trends are extracted from the analytical model through an implementation of the model. These trends are important is validating the model through intuitive arguments. The analytical model is implemented using MATLAB® v7.0 with key parameters being generated by an OPNET® v11.0 based simulator for IEEE 802.11.
- The second stage verifies theoretical result against results generated by a simulator also implemented using MATLAB® v7.0. This simulator is used in conjunction with the OPNET IEEE 802.11 handoff simulator to form the full system simulator.

Figure 16 and Figure 22 show the top level design of the verification stages.

6.2 Implementation of the Analytical Model

This section provides performance results of the proposed method using a direct implementation of the analytical model in MATLAB®. The system used for evaluating the proposed method is composed two main parts: link layer (using IEEE 802.11) simulator that extends the existing implementation in OPNET®, an implementation of the analytical model in MATLAB®, as well as an evaluation tool that generates performance metrics of the model. The next few subsections discuss the details of this system and the resulting performance metrics.

6.2.1 Approach

Figure 16 shows a block diagram of the system used for the theoretical evaluation.

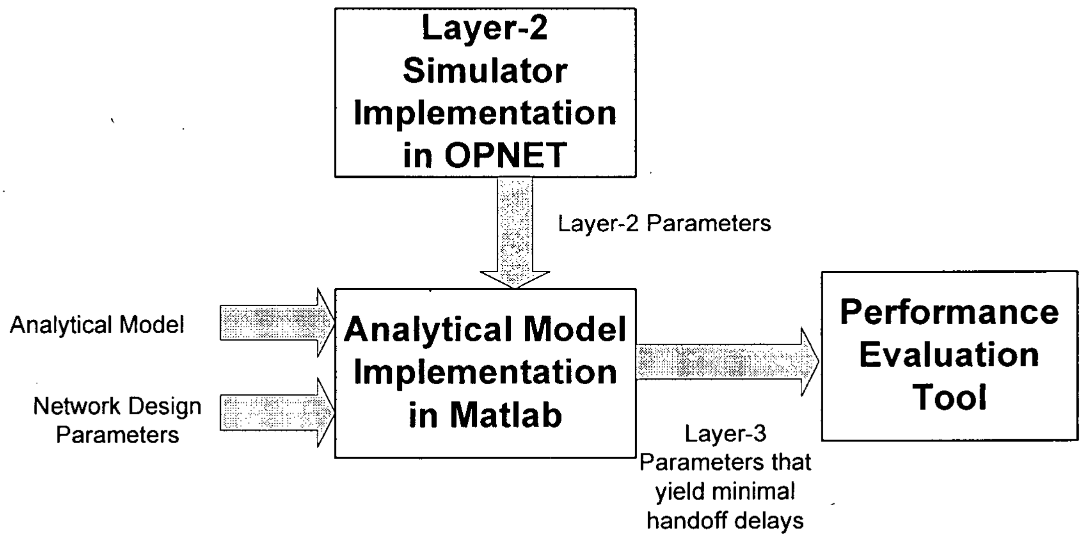


Figure 16: Theoretical evaluation system setup

6.2.1.1 Layer-2 Handoff Simulator

This is an OPNET® based simulator that is used to simulate Layer-2 handoff. The simulator generates Layer-2 configuration parameters (i.e. Expected value of the handoff execution time, and Expected value of the handoff bandwidth (BW_{HO})). These parameters are then used as inputs to the Analytical Model simulator. To generate these parameters, we implemented a simple experiment that simulates an MN performing handoffs between two cells with IEEE 802.11 used for layer-2.

6.2.1.2 Analytical Model Implementation in MATLAB®

This part of the system uses Layer-2 configuration parameters generated by the Layer-2 handoff simulator, and the configurable network design parameters (i.e. total available bandwidth BW, data bandwidth BW_D) as inputs to the mathematical model. The system generates the Layer-3 configuration parameters (the maximum RA interval that would provide the minimum move-detection delay) needed to evaluate the performance of the proposed method.

6.2.1.3 Performance Evaluation Tool

The performance evaluation tool is a MATLAB® tool that uses the information provided by the analytic model implementation to generate performance metrics for the proposed model.

6.2.2 Performance Evaluation

6.2.2.1 Assumptions

The move-detection parameters assume that the underlying network employs IEEE 802.11 APs. We assume that APs perform layer-3 functions of a router (in addition to 802.11 MAC) including the issue of RAs and that APs/Routers in different cells have identical configurations and broadcast RAs within the same frequency range.

Further, we assume that the MN does not perform DAD [1] and that the time needed to perform binding updates [1] is negligible. As a result, the value of the layer-3 handoff execution time is negligible and the expected value of the overall handoff execution time reported in Table 1 represents the layer-2 (i.e. IEEE 802.11a) handoff time. The handoff time consists of the layer-2 move-detection, layer-2 search, and the layer-2 handoff execution time.

The length of the RA packet in Mobile IPv6 is 512 bits divided as follows:

- A 16 byte RA header (as defined in [36]),
- An 8 byte source/target Link-layer Address (as defined in [36]),
- A 32 byte prefix option (as defined in [36] and modified in [1]),
- An 8 byte advertisement interval option (as defined in [1]).

Table 1: Various constants used throughout the experiments

Length of RA packet (K) in Mobile IPv6	512 bits
Time period of interest in seconds (T)	100 seconds
Expected value of handoff execution time (m) assuming an 802.11 wireless model	0.2 seconds
Expected value of handoff bandwidth (BW_{HO})	1 Kbps

To simplify the representation of results, we assume that the minimum and maximum values of the switching interval are identical and consequently the expected value of the switching interval has the same value as the minimum and maximum. Further, we assume that the overall available bandwidth, data bandwidth, and handoff bandwidth all have uniform distributions. As a result, their expected value will be the average of their respective minimum and maximum values.

6.2.2.2 Results

Chapter 5 provided an analytical model for calculating the RA interval limit and the expected value of the move-detection latency. This section uses Equations (21) and (22) to provide numerical results. Equation (21) is used to demonstrate the relationship between the switching interval limit (i.e. 1/switching rate) and the number of handoff calls. Equation (22), on the other hand, is used to demonstrate the relationship between the allowed value for the RA interval upper limit b (which is linearly proportional to the move-detection delay as in Equation (19)), total available bandwidth (BW), the data bandwidth (i.e. the bandwidth consumed by data calls, BW_D), the handoff bandwidth (i.e. the bandwidth consumed by handoff calls, BW_{HO}), and the switching interval [η, β]. The handoff bandwidth is assumed to be uniformly distributed. Without loss of generality, it is assumed to have the expected value reported Table 1.

Figure 17 uses Equation (21) to show the expected value of the number of handoff calls in a period of time T as a function of the Switching Interval (i.e. inverse of switching rate) when the switching rate has a uniform distribution. The value of the RA limit b is assumed to be 0.2 seconds.

Figure 17 shows that, as the value of the switching interval increases beyond 5 seconds, the expected value of the number of handoff calls approaches zero. On the other hand, as the switching interval becomes less than 2.2 seconds, the expected value of the number of handoff calls increases dramatically. As an example, when the number of handoff calls is 200, the switching interval is 0.5 seconds (i.e. switching rate = 2 switches/second). These results match intuition since the number of handoff calls should increase when the MN switches between cells more frequently.

Figure 18 uses Equation (22) to show how the RA interval changes for different switching rates. From Figure 18, it is evident that the RA interval becomes roughly constant when the switching interval is greater than 12 seconds. As the switching interval approaches zero, on the other hand, there is a "spike" in the RA interval limit.

The value of the total available bandwidth and the data bandwidth are normally dependent on the physical layer technology used. However, without loss of generality, we assume the values presented in Table 2.

Table 2:Constants used for the experiment in Figure 18

Total available bandwidth (BW)	1.43 Mbps
Data bandwidth (BW_D)	35 Kbps

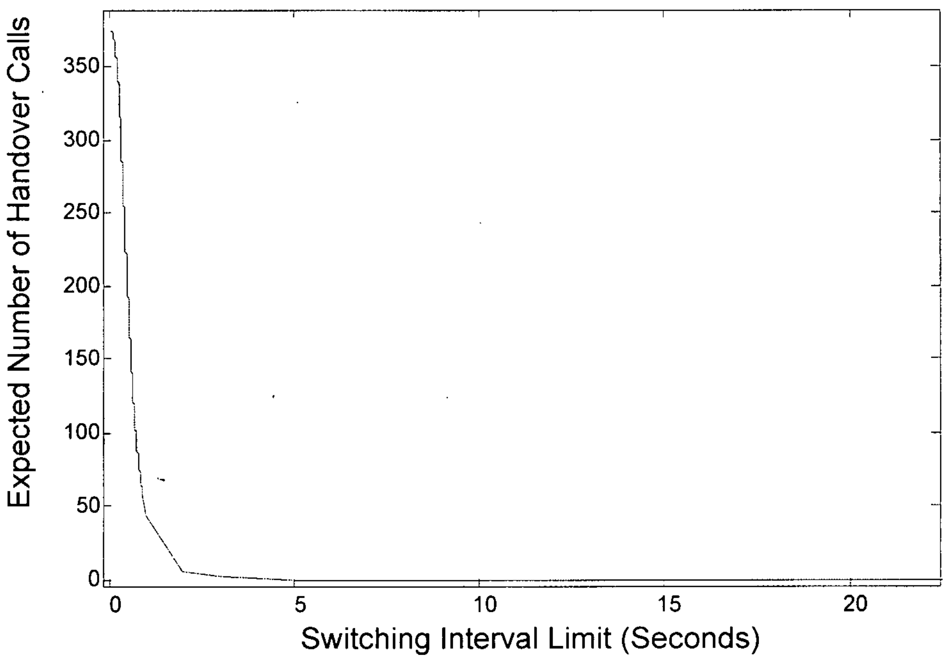


Figure 17: Effect of switching interval on the expected value of the number of handoff calls for a given RA interval value.

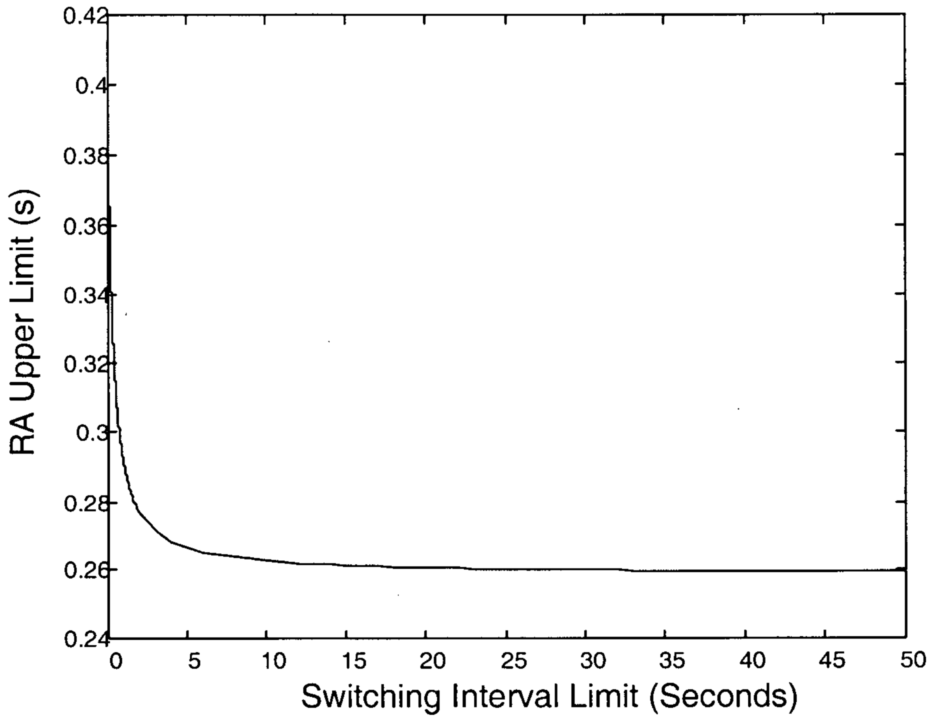


Figure 18: Effect of the switching interval on RA interval.

Figure 19 shows the effect of the total available bandwidth (*BW*) on the RA interval for various switching rates. The data bandwidth (BW_D) is assumed to be 35 Kbps. From this figure, there are several important trends that can be noticed. First, it can be seen that smaller switching intervals (i.e. higher switching rates) result in a longer RA limit (for a given *BW*). This is due to the fact that smaller switching intervals will result in a larger number of handoffs which will consume more bandwidth. This will result in less bandwidth being available for RAs. Consequently, the RA interval limit must then become shorter.

Another important trend from Figure 19 is that the RA limit decreases with an increase in the total bandwidth (for a fixed switching rate). When more bandwidth is made available, RAs can consume more bandwidth by an increase in frequency.

A third trend that can be noticed from Figure 19 is that the difference between curves narrows as the bandwidth increases. When the total bandwidth is large, there is an abundance of bandwidth available for RAs. As a result, no matter what the switching rate is (and what the number of handoff is), RAs will have plenty of bandwidth to operate at a higher frequency. The opposite is true when the total bandwidth is small; the amount of bandwidth left available to RAs is extremely sensitive to the bandwidth used by frequent handoffs.

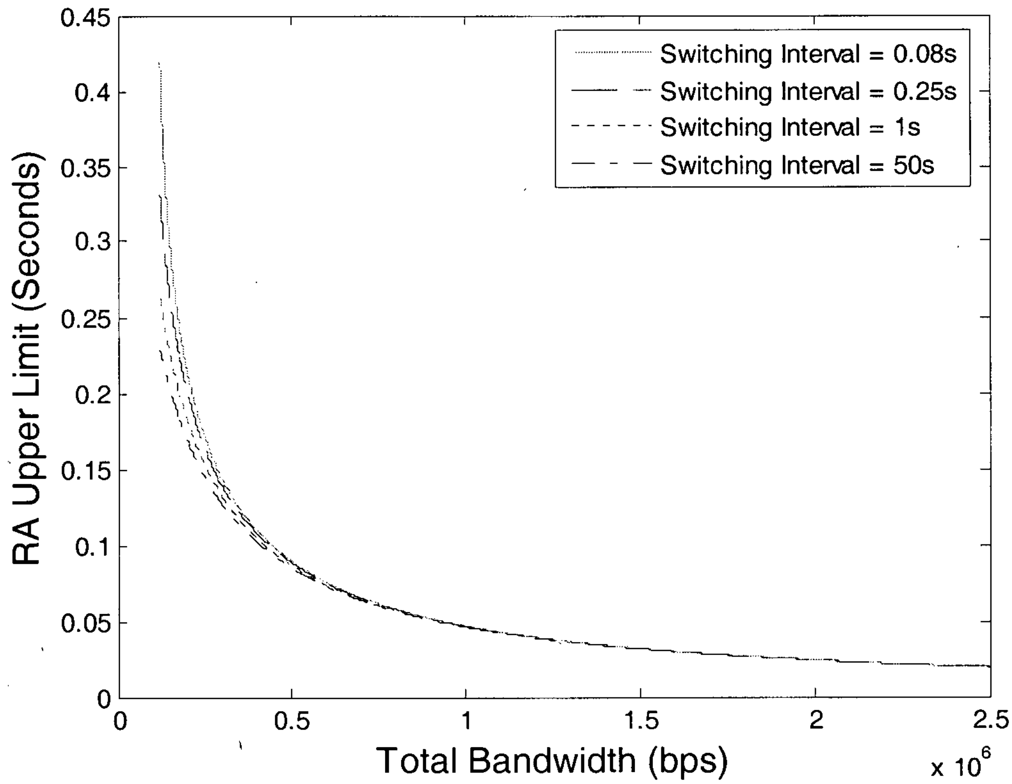


Figure 19: Effect of the total available bandwidth on RA interval for different switching intervals/rates

Figure 20 shows the effect of the data bandwidth, BW_D, on the RA interval limit for different switching intervals. As BW_D becomes larger, the available bandwidth for the RA interval becomes more limited. Further, it can be seen from the graph that this effect becomes more pronounced when the switching interval becomes shorter (i.e. switching rate becomes higher). As the switching rate increases, the number of handoff calls becomes higher. As a result, the bandwidth consumed by handoffs is higher. Consequently, the bandwidth for RAs is lower so that its interval limit becomes longer.

Figure 21 shows the most significant result in this analysis. It shows the improvement introduced using our model by comparing the resulting move-detection delay when the RA interval limit is adjusted using two different adaptive network schemes. The solid line uses simple, single handoff, models that ignore the delay between consecutive handoffs (i.e. $\delta = 0$). The dashed line represents our proposed approach that includes the effects of recurrent switching between cells. As an example, we consider an adaptive network were the total bandwidth is limited to 110 Kbps. In this case, the simple model would choose an RA such that the resulting move-detection delay is 320 ms (point A). On the other hand, our model would result in a move-detection delay of 105 ms (point B). By using our model (that accounts for recurrent handoffs), the adaptive network is able to choose a more suitable RA resulting in a 65% improvement in move-detection delay. Note that these numbers will vary as the configuration parameters (such as the total bandwidth and data bandwidth) vary.

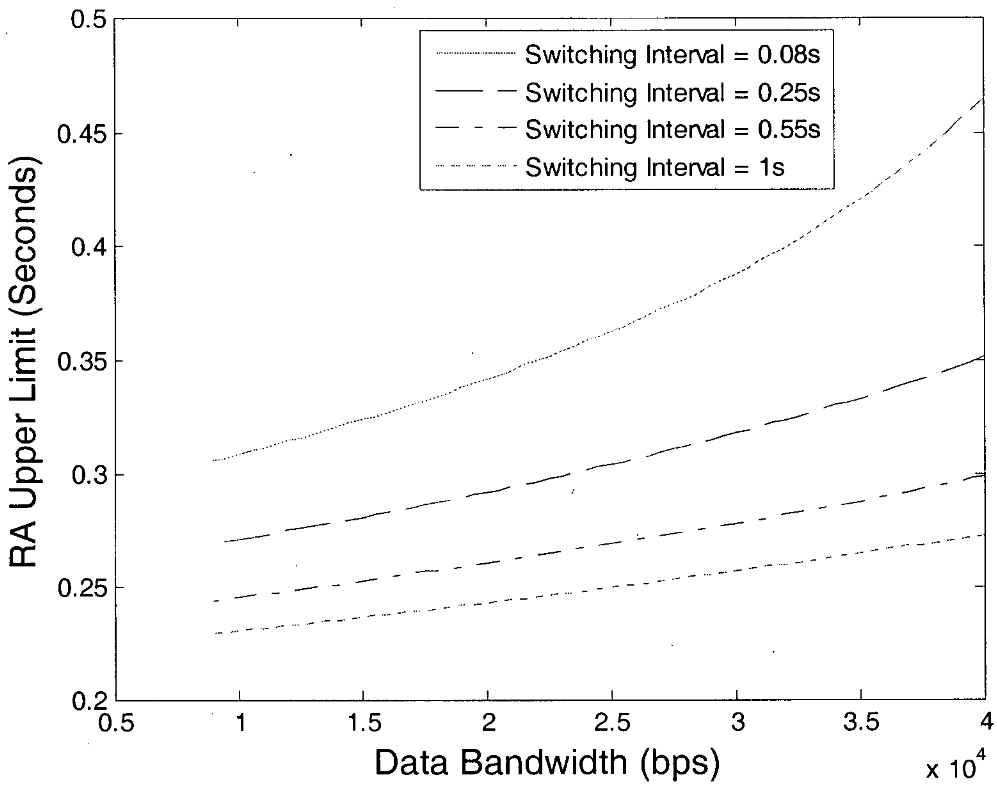


Figure 20: Effect of data bandwidth on RA interval for different switching intervals/rates

Because simple models do not account for the random delay (switch interval or switching rate) between consecutive handoffs, they erroneously determine that handoffs will occur at a higher rate. Thus, they determine that there will be less bandwidth available for RAs than our model. Consequently, they set the upper limit of the RA interval higher than our model. By doing so, the expected value of the move-detection delay will be longer for simple models when compared to our model.

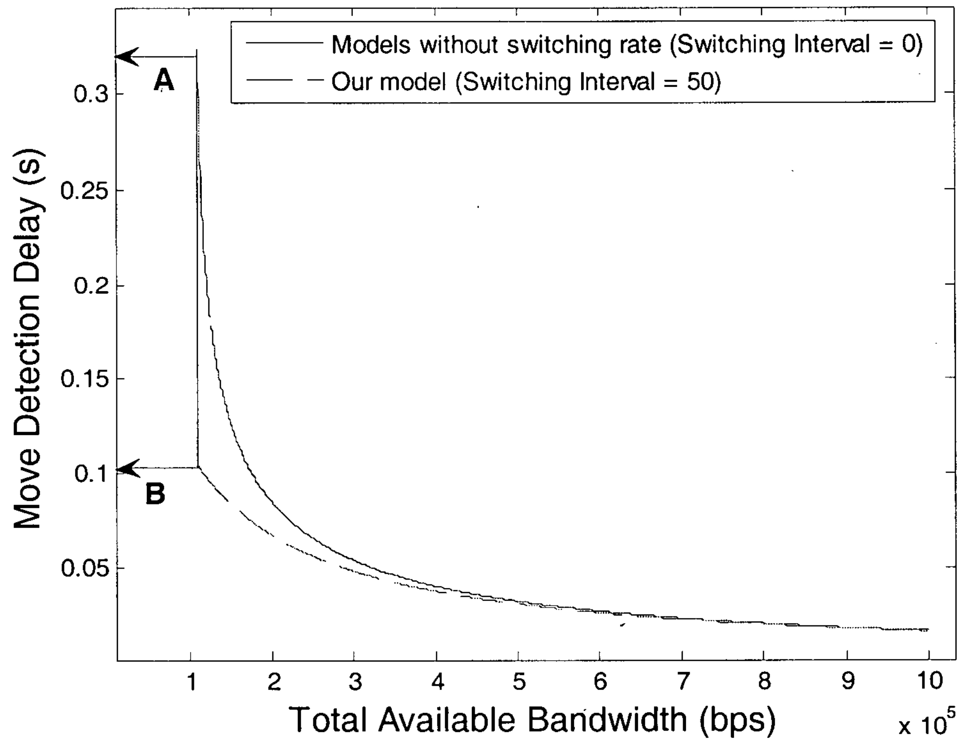


Figure 21: Comparison of models in terms of the resulting move-detection delay (i.e. effect of accounting for the switching interval/rate when adjusting the RA limit as a function of the bandwidth).

6.3 System Simulator

The purpose of the system simulator is to verify the theoretical results obtained in the previous section. Figure 22 shows a block diagram that illustrates the approach used to implement the simulator. The simulator simulates a network where an MN performs recurrent handoffs between two cells. The system uses an IEEE 802.11 link layer and a Mobile IPv6 network layer. The first part of the simulator simulates IEEE 802.11 handoffs. The second part simulates the proposed model in a realistic network and is implemented in MATLAB®. The third part is a performance evaluation tool that provides performance metrics for the simulations of the proposed model. The next few subsections describe the different parts of the system simulator.

6.3.1 Approach

Figure 22 shows a block diagram of the system simulations. In addition, a discussion of each component is provided in the following subsections.

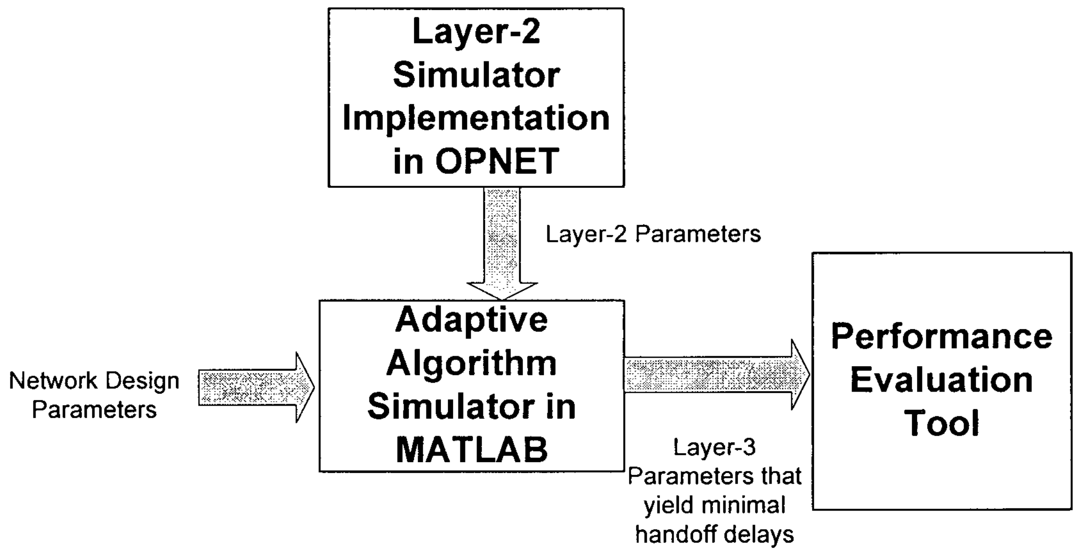


Figure 22: System Simulator

6.3.1.1 Layer-2 Handoff Simulator

The layer-2 handoff simulator is based on OPNET® and is used to simulate a Layer-2 (i.e. IEEE 802.11) handoff. It uses the same implementation and network design used in section 6.2.1.1.

6.3.1.2 Adaptive Model Simulator

The adaptive model simulator is implemented in MATLAB®. It simulates an MN roaming between two cells, and uses the proposed adaptive model for adjusting the

RA interval based on the available bandwidth. It uses the Layer-2 configuration parameters generated from the OPNET® IEEE 802.11 simulator. These parameters are identical to the ones used in section 6.2.1.1. Layer-3 configuration parameters used are shown in the various tables and graphs in section 6.3.2. Appendix B contains a pseudo code that provides a high level description of the simulator.

6.3.1.3 Performance Evaluation Tool

Similar to the tool used in the theoretical results section, this is a MATLAB® based performance evaluation tool that uses information provided by the network simulator to generate performance evaluation metrics for the proposed method.

6.3.2 Performance Evaluation

6.3.2.1 Assumptions

The assumptions made in this part of the evaluation process are identical to the assumptions stated in section 6.2.2.1.

6.3.2.2 Results

Simulation results are meant to compare the main performance metrics obtained from the analytical model implementation to the ones obtained from the simulator.

Figure 23 shows the number of handoffs as a function of the switching interval for both the analysis and the simulation. In both cases, the number of handoffs decreases as the switching interval increases and they both follow an exponential decay. Further, the variation is more dramatic when the switching interval is smaller than 5 seconds. The general trend is similar in both curves. The difference between these curves is due to sources of error introduced through simulation because of randomness and due to the discrete nature of the simulation.

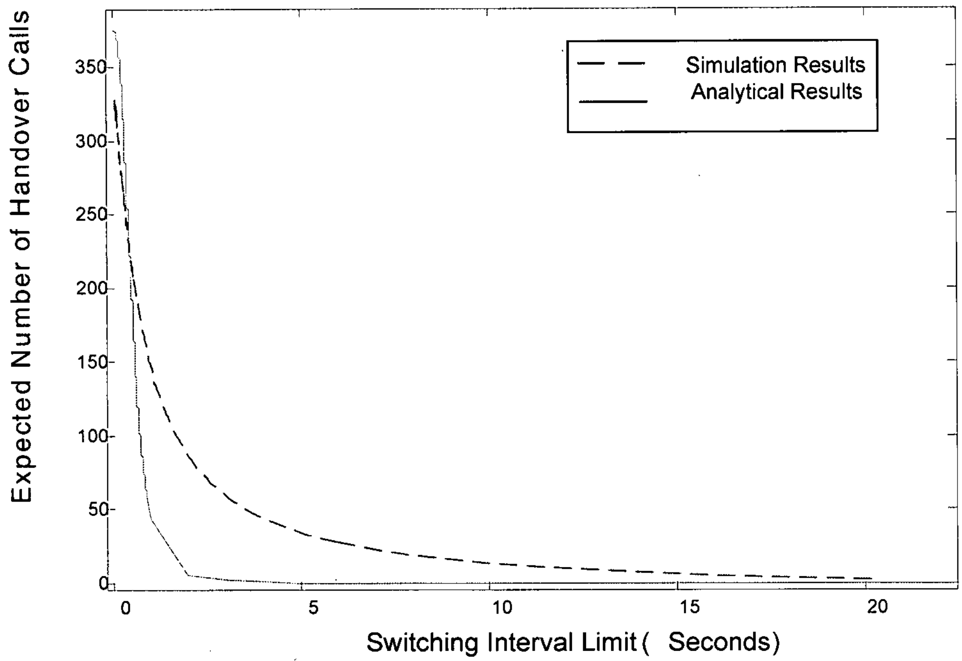


Figure 23: The effect of the switching interval on the expected value of the number of handoff calls for a given value of the RA interval. Both analytical and simulation results are provided.

Figure 24 shows the RA upper limit value as a function of the switching rate for both the simulation and the analytical model. In both cases, the RA interval upper limit decreases as the switching interval increase. In a similar way to the results in Figure 23, the discrepancy between simulation results and analytic results is likely due to the randomness of the simulation and because of its discrete nature.

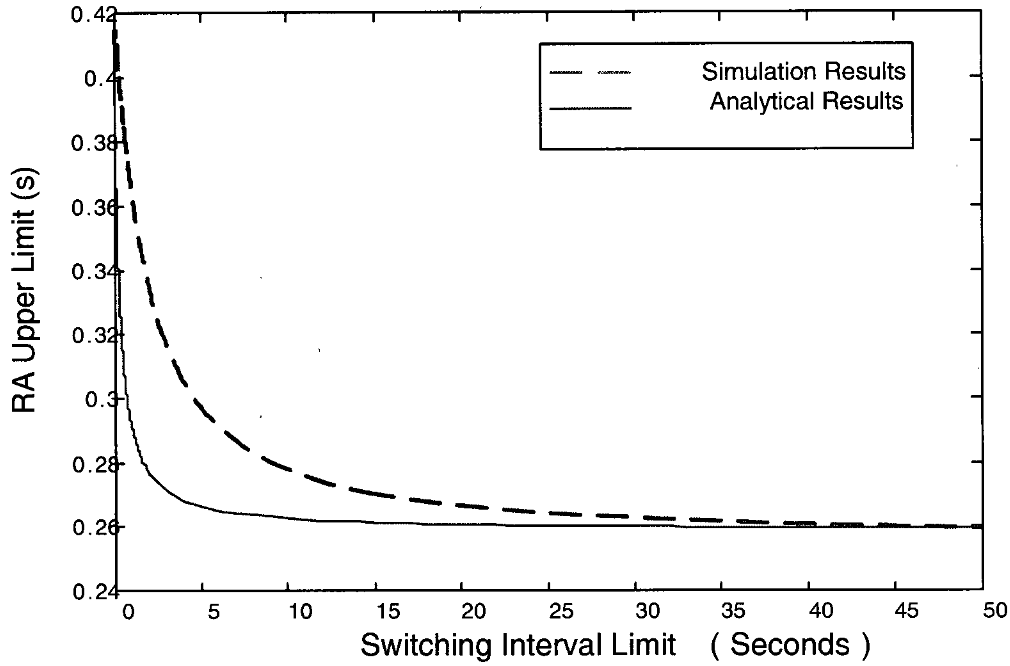


Figure 24: Effect of switching interval on the RA interval. Both analytical and simulation results are provided.

Figure 25 shows the move-detection delay as a function of the BW when different RA interval estimates are used and for both the simulation and the analysis. It compares our method to other methods that do not consider the switching rate. It shows that our method achieves a better estimate for the upper limit of the RA interval based on the available bandwidth. For simulation, our estimate yields a 56% reduction in move-detection delay over existing methods. For the analytical model, our method achieves a 65% reduction in move-detection delay over existing methods.

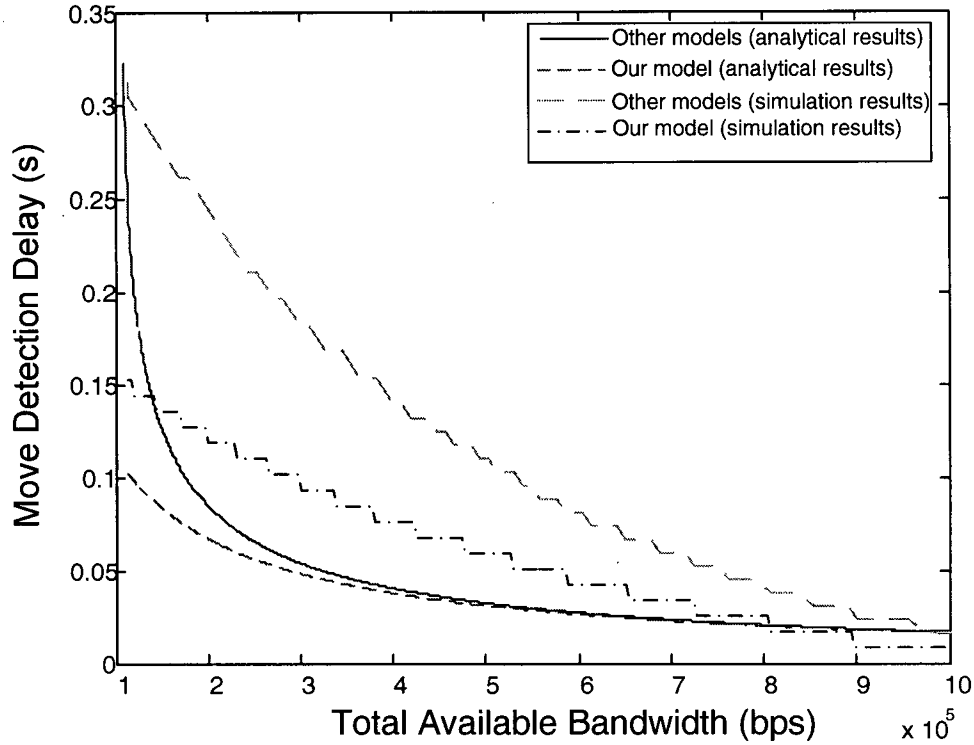


Figure 25: Comparison of Models in terms of the resulting move-detection delay (i.e. effect of accounting for the switching interval/rate when adjusting the RA limit as a function of the bandwidth). Both analytical and simulation results are provided.

7 CONCLUSIONS AND FUTURE WORK

7.1 Conclusion

This research investigated the issue of interruptions caused by handoff delays in an environment where recurrent handoffs are expected. The primary focus is on micro-mobility since it is for this case that the MN would likely experience frequent handoffs. Although the analysis provided in this work is intended for micro-mobility, it would equally apply to macro-mobility.

The goal of this research is to determine where bottlenecks exist in the handoff process (both in layer-2 and layer-3) and to provide a solution that addresses these bottlenecks. To accomplish this, we first selected an appropriate micro-mobility management protocol. This entailed a survey of the available micro-mobility management protocols as well as an evaluation of the discussed protocols based on specific mobility management criteria. Mobile IPv6 was selected as the most appropriate solution for micro-mobility management. We then identified the delay bottlenecks in the handoff process in both layer-2 using IEEE 802.11 and layer-3 using Mobile IPv6. This was achieved by providing detailed delay measurements of the handoff process using an OPNET® simulation environment. The layer-2 handoff was separated into the following three phases: move-detection; the search for a new AP; and switching to a new AP. Layer-3 handoff was separated into two phases: move-detection; and handoff execution. It was shown that the main bottleneck of handoff delays is the Layer-3 move-detection phase which, on average, contributes 57% of the total delay. The second major contributor to these delays is the binding delay and the DAD process performed in the

Layer-3 handoff. This contribution is, on average, 33% of the total delay. It was also shown that, on average, Layer-2 handoff delay contributes less than 10% of the overall layer-2 and layer-3 handoff delay.

After identifying move-detection as the prominent source of delay in the handoff process, we proposed an approach that minimizes this delay. We modeled a realistic scenario for recurrent handoffs and proposed a model for describing the recurrent cell switching activity of an MN. Renewal Theory was used to provide a mapping between the cell switching distribution and the expected delay between consecutive handoffs. This model analyzed the RA interval as a function of the bandwidth and the cell switching behavior. This analysis was used to derive an adaptive network capable of dynamically adjusting the RA interval as a function of the cell-switching behavior and available bandwidth in order to minimize the move-detection delay. Numerical results showed a reduction in move-delay by as much as 65% over adaptive networks that only consider a single handoff model. Simulation results showed a reduction in move-delay by as much as 56%.

7.2 Future Work

This research work used Renewal Theory to solve mobility related issues. In particular, it solved the issue of interruptions and data lost during the handoff process. The analysis performed in this work assumed a uniform distribution for the handoff bandwidth, the total bandwidth, and the data bandwidth. Possible future work could involve the same analysis as Chapter 5 but for different distributions. Similarly, it would be interesting to explore how the system would change if various components of the bandwidth such as the total bandwidth or data bandwidth fluctuated as time progresses.

Bandwidth calculations were performed from the router's perspective and it was assumed that a single MN performs the handoffs between two cells (routers). Possible future work could investigate the impact of using a different number of MNs for each distribution. Further, it could include an investigation of how the analysis would differ if each router within a system did not have the same configuration as assumed in this work.

7.3 List of Contributions

The following list summarizes the key contributions of this thesis:

1. We provided a survey on the most common micro-mobility management protocols and evaluated each protocol according to a predefined set of objectives expected in a mobility management protocol. Of the protocols discussed, this evaluation showed that Mobile IPv6 is the most appropriate solution for micro-mobility management.
2. We studied the handoff process in layer-2 using IEEE 802.11 and in layer-3 using Mobile IPv6 and provided a detailed characterization of its various stages. This characterization identified that the main bottleneck in handoff delays is move-detection.
3. We proposed an adaptive network model that dynamically adjusts the RA interval given a set network bandwidth and data bandwidth in order to minimize handoff delays. Analytical results showed that our model reduces move-detection delay by as much as 65% over other models that do not consider the delay between consecutive handoffs. Simulation results showed reductions in move-detection delay by as much as 56% reduction.

BIBLIOGRAPHY

[1] D. Johnson and C. Perkins, "Mobility Support in IPv6," IETF draft, July 2003.

[2] C.E. Perkins, "IPmobility support for IPv4", RFC 3344.

[3] C.E. Perkins, "Mobile IP – Design Principles and practices", Addison-Wesley, 1998, ISBN 0-13-850505-5.

[4] T.W. Andersen, A. Lildballe, and B. Nielsen, "Handoff Initiation in Mobile IPv6", Fourth International Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark, September 2001.

[5] S. Dixit, and R. Prasad, "Wireless IP and Building the Mobile Internet", Artech House, Incorporated, P393-P413, 2002.

[6] Q. Zhao, L. Feng, Z. Li, and J. Yang, "Movement detection delay analysis in mobile IP", Computer Communications 28(5): 550-556, 2005.

[7] M. Siksik, H. Alnuweiri and S. Zahir, "Micro-mobility Management in Mobile IPv6 Networks: Issues and Solutions", Wirelesscomm. International Conference on Wireless Networks, Communications and Mobile Computing, Maui, Hawaii, USA, June 2005.

[8] M. Siksik, H. Alnuweiri and S. Zahir, "A detailed Characterization of the Handover Process Using Mobile IPv6 in 802.11 Networks", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, Canada, August 2005.

[9] Velayos, H. and G. Karlsson, "Techniques to Reduce IEEE 802.11b MAC Layer Handover Time", Laboratory for Communication Networks, KTH, Royal Institute of Technology, Stockholm, Sweden, TRITA-IMIT-LCN R 03:02, April 2003.

[10] A. Mishra, M. ho Shin, and W. A. Arbaugh, "An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process", Submitted to ACM Computer Communications Review, Volume 33, Issue 2, April 2003.

[11] N. Montavont and T. Noel, "Handover Management for Mobile Nodes in IPv6 Networks", IEEE Communication Magazine, August 2002.

[12] S. Seshan, H. Balakrishnan, and R. H. Katz, "Handoffs in Cellular Wireless Networks: The Daedalus Implementation and Experience", ACM/Baltzer Journal on Wireless Networks, 1995.

[13] C. Tan, S. Pink, and K. Lye, "A Fast Handoff Scheme for Wireless Networks", Proceedings of the Second ACM International Workshop on Wireless Mobile Multimedia, ACM, August 1999.

[14] N. A. Fikouras and C. Gorg, "A Complete Comparison of Algorithms for Mobile IP Hand-offs with Complex Movement Patterns and Internet Audio", In Proceedings of the Fourth International Symposium on Wireless Personal Multimedia Communications (WPMC), Aalborg, Denmark, September 2001.

[15] N. A. Fikouras, and C. Gorg, "Performance Comparison of Hinted and Advertisement Based Movement Detection Methods for Mobile IP Hand-offs", 7th European Conference on Fixed Radio Systems and Networks, Sept. 2000.

[16] N. Fikouras, K. EL Malki, S.R. Cverkovic, and M. Karner, "Performance Analysis of Mobile IP Handoffs", Microwave Conference, Asia Pacific3, 770-773, 1999.

[17] T. W. Andersen, A. Lildballe, and B. Nielsen, "Handoff Initiation in Mobile IPv6", Fourth International Symposium on Wireless Personal Multimedia Communications, Aalborg, Denmark, September 2001.

[18] S. Dixit, and R. Prasad, "Wireless IP and Building the Mobile Internet", Artech House, Incorporated, P393-P413, 2002.

[19] Q. Zhao, L. Feng, Z. Li, J. Yang, "Movement detection delay analysis in mobile IP", Computer Communications 28(5): 550-556, 2005.

[20] E. Parzen, "Stochastic Process", Holden-Day, 1962.

[21] C. Williams, Localized Mobility Management Goals, Internet-Draft, July 2004.

[22] P. Reinbold and O. Bonaventure, "A survey of IP micro-mobility protocols", IEEE Communications Surveys and Tutorials, 2003.

[23] V. Devarapalli, R. Wakikawa, A. Petrescu, P. Thubert, "Network Mobility (NEMO) Basic Support Protocol", Ineternet Draft, June 2004.

[24] W.M. Eddy, "At What Layer Does Mobility Belong?" IEEE Communications Magazine, Volume 42, Issue 10, pp. 155-159, 2004.

[25] C. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers", ACM Computer Communications Review, 24(4), October 1994.

[26] Bluetooth Core Specification Version 2.0, Bluetooth SIG, November 8, 2004.

[27] IEEE. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Standard 802.11, 1999.

[28] R. Droms, "Dynamic Host Configuration Protocol", RFC 2131.

[29] P. Mockapetris, "Domain Names--Concepts and Facilities", RFC 1034.

[30] R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson, "Stream Control Transmission Protocol", RFC 2960.

[31] I. Aydin and C. Shen, "Cellular SCTP: A Transport-Layer Approach to Internet Mobility", Internet Draft, October 2003.

[32] B. Landfeldt, T. Larsson, Y. Ismailov, and A. Seneviratne, "SLM, A Framework for Session Layer Mobility Management", Proceedings of IEEE ICCCN, October 1999.

[33] J. Postel, "Internet Control Message Protocol", RFC792,.

[34] C. Perkins and D. B. Johnson, "Route optimization in Mobile IP", Internet-Draft, November 1997.

[35] P. Ferguson and D. Senie, "Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing", RFC 2267.

[36] T. Narten, E. Nordmark, and W. A. Simpson, „Neighbor Discovery for IP version 6 (IPv6)", Internet-Draft, July 1997.

[37] S. Kent and R. Atkinson, "IP Authentication header", Internet-Draft, October 1997.

[38] S. Kent and R. Atkinson, "IP Encapsulating Security Payload (ESP)", Internet-Draft, October 1997.

[39] S. Kent and R. Atkinson, "Security architecture for the Internet Protocol", Internet-Draft, November 1997.

[40] Z. D. Shelby, D. Gatzounas, A. Campbell , and C. Wan, "Cellular IPv6", Internet Draft, November 2000.

[41] A. T. Campbell, J. Gomez, S. Kim, A. G. Valko, C-Y. Wan, and Z. Turanyi, Cellular IP, Internet draft, draft-ietfmobileip-cellularip-00.txt, January 2000.

[42] A. T. Campbell, J. Gomez, S. Kim, A. G. Valko, C-Y. Wan, and Z. Turanyi., "Design, Implementation, and Evaluation of Cellular IP", IEEE Personal Communications, 7(4):42–49, August 2000.

[43] A. G. Valko, "Cellular IP: A New Approach to Internet Host Mobility", Computer Communication Review, 29(1):50–65, January 1999.

[44] E. Gustafsson, A. Jonsson, and C. Perkins, "Mobile IPv4 Regional Registration", Internet draft, November 2004.

[45] P. McCann, T. Hiller, J. Wang, A. Casati, C. Perkins, and P. Calhoun, "Transparent Hierarchical Mobility Agents (THEMA)", Internet Draft, March 1999.

[46] K. El-Malki and H. Soliman, „Fast Handoffs in Mobile IPv4", Internet draft, , September 2000.

[47] P. Calhoun et al., "Agent Assisted Hand-off," Internet draft, November 2000.

[48] S. Das et al., "TeleMIP: Telecommunication Enhanced Mobile IP Architecture for Fast Intra-Domain Mobility", IEEE Pers. Commun., August 2000.

[49] R. Ramjee, T. La Porta, and L. Li, "Paging support for IP mobility using HAWAII", Internet Draft, June 1999.

[50] R. Ramjee, T. La Porta, S. Thuel, and K. Varadhan, "IP micro-mobility support using HAWAII", Internet Draft, June 1999.

[51] P. Eardley, A. Mihailovic, and T. Suihko , "A Framework for the Evaluation of IP Mobility Protocols", PIMRC '00, London, September 2000.

[52] S. Seshan, H. Balakrishnan, and R. H. Katz, "Handoffs in Cellular Wireless Networks: The Daedalus Implementation and Experience", ACM/Baltzer Journal on Wireless Networks, 1995.

[53] J. Mysore, and V. Bharghavan, "A New Multicasting-based Architecture for Internet Host Mobility", Proceeding of ACM Mobicom, September 1997.

[54] C. Tan, S. Pink, and K. Lye, "A Fast Handoff Scheme for Wireless Networks", Proceedings of the Second ACM International Workshop on Wireless Mobile Multimedia, ACM, August 1999.

[55] A. Mihailovic, M. Shabeer, A.H. Aghvami, "Multicast for Mobility Protocol (MMP) for emerging Internet networks", Proceedings of PIMRC2000, London, UK, September 2000.

[56] A. Helmy, M. Jaseemuddin, and G. Bhaskara, "Multicast-based Mobility: A Novel Architecture for Efficient Micro-Mobility", IEEE Journal on Selected Areas in Communications (JSAC), Special Issue on All-IP Wireless Networks, 1st Quarter 2004.

[57] A. O'Neill, G. Tsirtsis, and S. Corson, "Edge Mobility Architecture", Internet draft, March2000.

[58] I. D. Chakeres and E. M. Belding-Royer. "AODV Routing Protocol Implementation Design." Proceedings of the International Workshop on Wireless Ad Hoc Networking (WWAN), Tokyo, Japan, March 2004.

[59] P. Ferguson and D. Senie, "Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing", Internet Draft, January 1998.

[60] H. Haverinen and J. Malinen, "Mobile IP Regional Paging", Internet draft, June 2000.

[61] S. Thomson and T. Narten, "IPv6 stateless address autoconfiguration", Internet-Draft, December 1998.

[62] W. Feller, "An introduction to Probability Theory and its applications", volume II. Wiley, second edition, 1971.

[63] D.P Heyman, and M.J. Sobel, "Stochastic models in operations research", New York : McGraw-Hill, 1982.

APPENDIX A

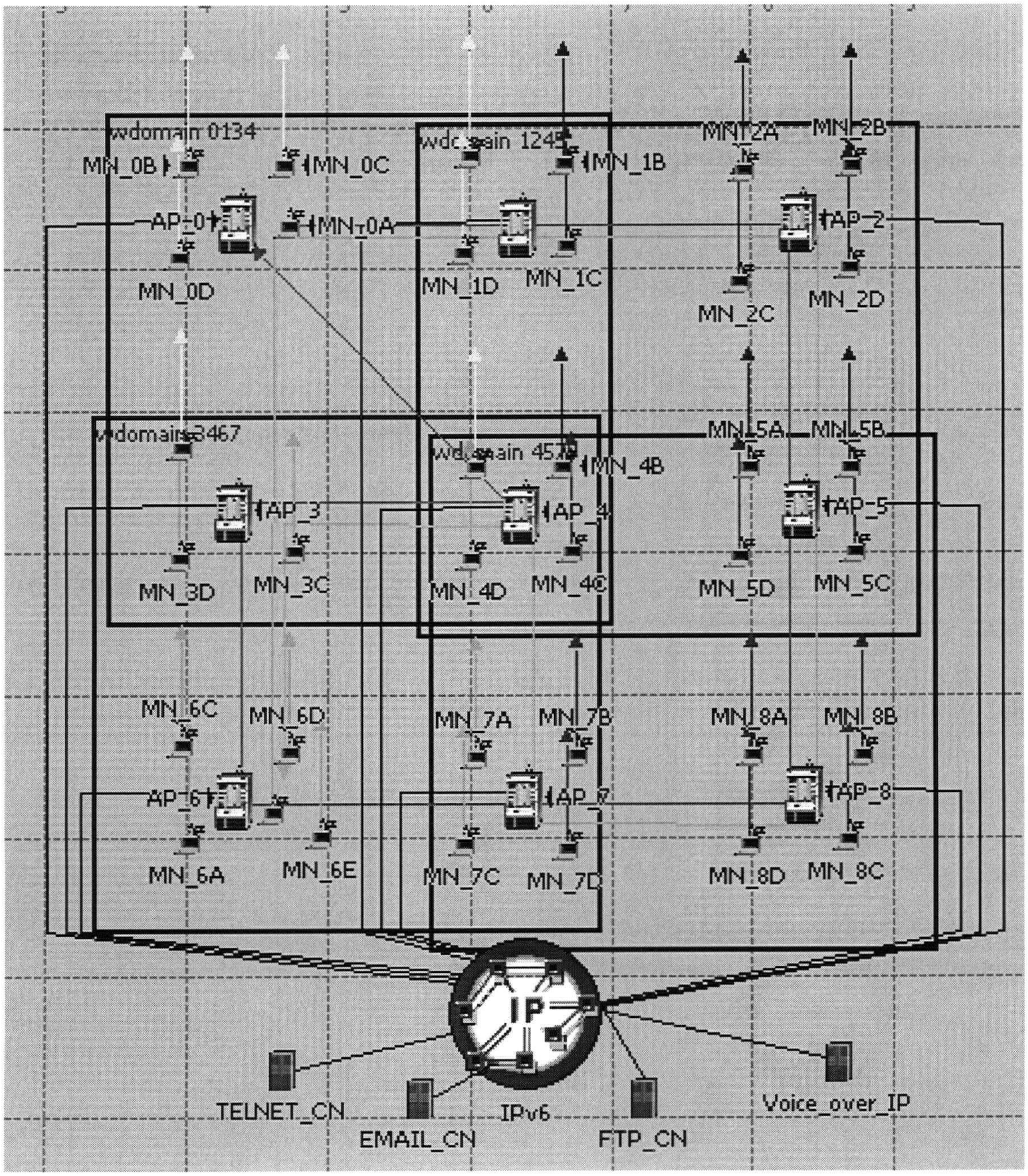


Figure 1: Network setup for simulating a handoff in layer-2 and layer-3

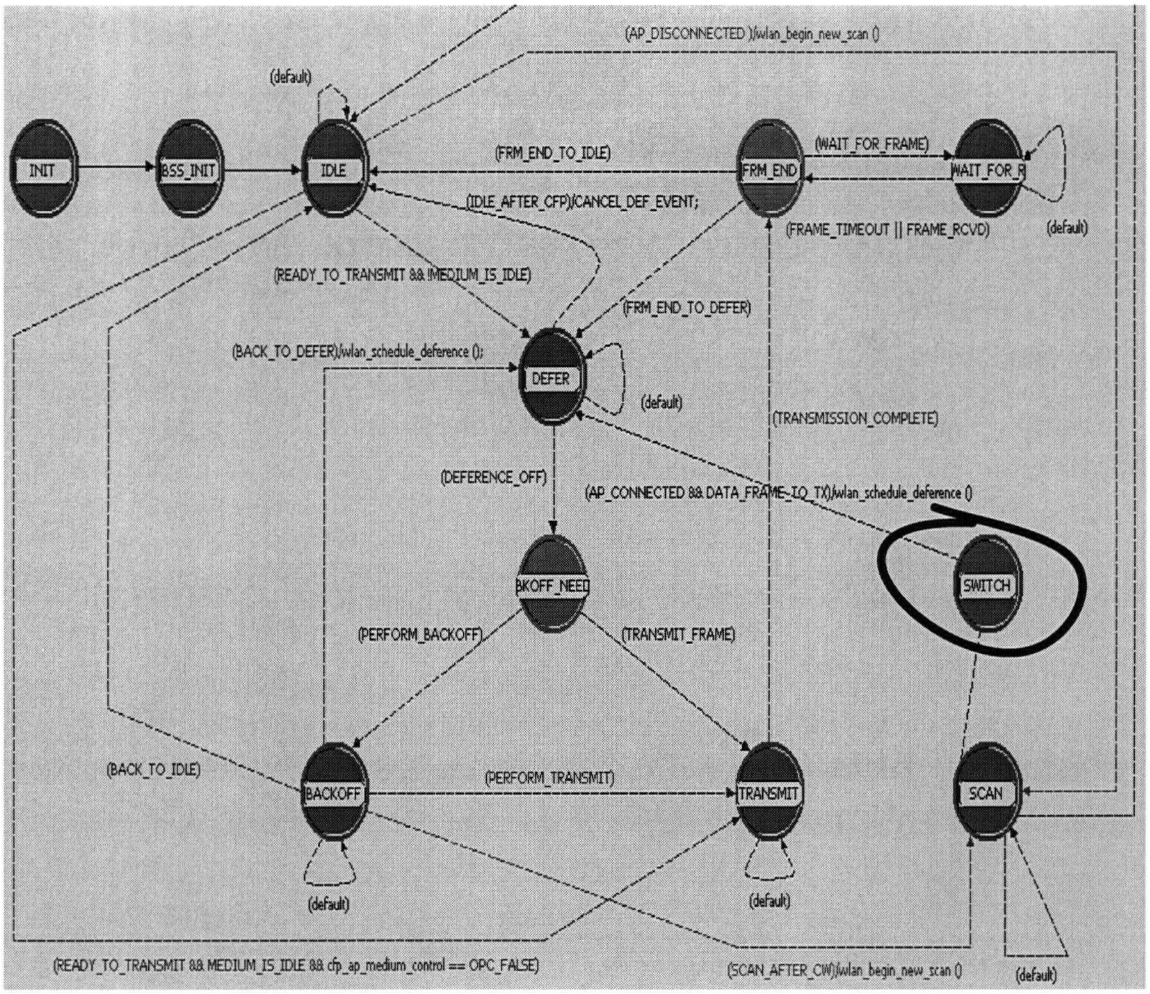


Figure 2: State machine used to simulate IEEE 802.11

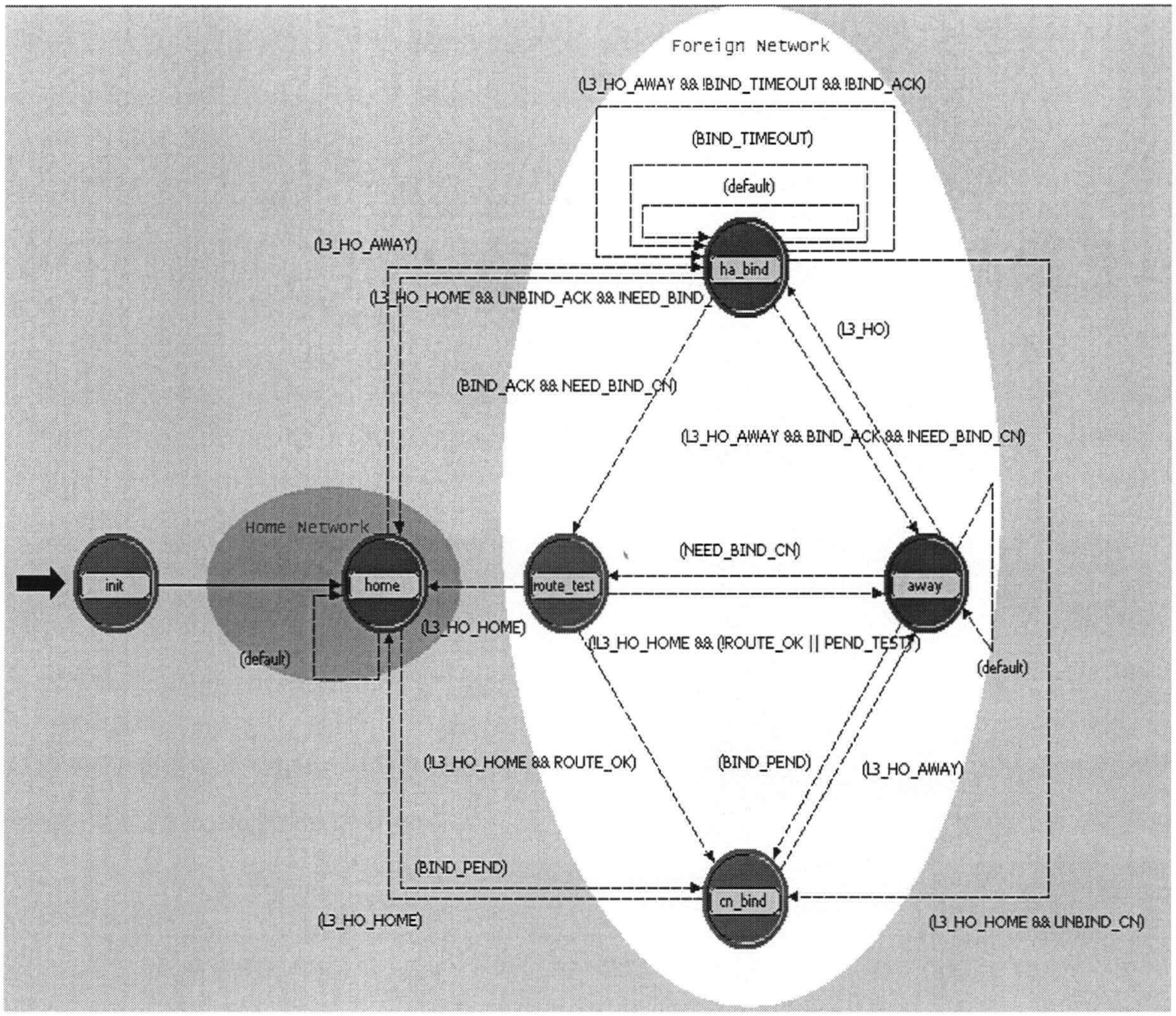


Figure 3: State machine used to simulate Mobile IPv6

APPENDIX B

The following is a pseudo code that provides a high level description of the simulator:

```
Import layer-2 configuration parameters from OPNET;
Generate pseudo random RA1;
Generate pseudo random RA2;
Initialize numberHandoffs;
Queue.empty();
Queue.add(RA1);
Queue.add(RA2);
T = 0;
While t < T
{
        newEvent = queue.pop();
        switch (newEvent.type)
        {
                case RA1 :
                {
                        T = newEvent.time;
                        Calculate (moveDetection);
                        Queue.add(moveDetection1);
                        RA1 = UpdateRA (RA1, numberHandoffs) ;
                }
                case handoffExec1 :
                {
                        T = newEvent.time;
                        Queue.add(handoffExec1);
                }
                case switchingInterval1:
                {
                        T = newEvent.time;
                        Queue.add(switchingInterval1);
                        numberHandoffs++;
                }
                case RA2 :
                {
                        T = newEvent.time;
                        CalculateMoveDetection (moveDetection2);
                        Queue.add(moveDetection2);
                        RA2 = UpdateRA (RA2, numberHandoffs) ;
                }
                case handoffExec2 :
                {
```

```
        T = newEvent.time;
        Queue.add(handoffExec2);
    }
    case switchingInterval2:
    {
        T = newEvent.time;
        Queue.add(switchingInterval2);
        numberHandoffs++;
    }
```