A Location-aided Decision Algorithm for Handoff Across Heterogeneous Wireless Overlay Networks

By
Areej Saleh

Thesis submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Computer Engineering

Dr. Nathaniel Davis, Chair
Dr. Scott Midkiff
Dr. Luiz DaSilva

July 8, 2004

Blacksburg, VA

Keywords: UMTS, WLAN, IEEE 802.11, 3G, Handoff, Heterogeneous, Location-aided

A Location-aided Decision Algorithm for Handoff Across Heterogeneous Wireless Overlay Networks

By
Areej Saleh

Abstract

Internetworking third generation (3G) technologies with wireless LAN (WLAN) technologies such as Universal Mobile Telecommunication Systems (UMTS) and IEEE 802.11, respectively, is an emerging trend in the wireless domain. Its development was aimed at increasing the UMTS network's capacity and optimizing performance. The increase in the number of wireless users requires an increase in the number of smaller WLAN cells in order to maintain an acceptable level of QoS. Deploying smaller cells in areas of higher mobility (e.g., campuses, subway stations, city blocks, malls, etc.) results in the user only spending a short period of time in each cell, which significantly increases the rate of handoff. If the user does not spend sufficient time in the discovered WLAN's coverage area, the application cannot benefit from the higher data rates. Therefore, the data interruption and performance degradation associated with the handoff cannot be compensated for. This counters the initial objective for integrating heterogeneous technologies, thus only handoffs that are followed by a sufficient visit to the discovered WLAN should be triggered. The conventional RF-based handoff decision method does not have the necessary means for making an accurate decision in the type of environments described above. Therefore, a location-aided handoff decision algorithm was developed to prevent the triggering of handoffs that result from short visits to discovered WLANs' coverage area. The algorithm includes a location-based evaluation that runs on the network side and utilizes a user's location, speed, and direction as well as handoff-delay values to compute the minimum required visit duration and the user's trajectory. A WLAN coverage database is queried to determine whether the trajectory's end point falls within the boundaries of the discovered WLAN's coverage area. If so, the mobile node is notified by the UMTS network to trigger the handoff. Otherwise, the location-based evaluation reiterates until the estimated trajectory falls within the boundaries of the discovered WLAN's coverage area, or the user exits the coverage area. By taking into consideration more then merely RF-measurements, the proposed algorithm is able to predict whether the user's visit to the WLAN will exceed the minimum requirements and make the decision accordingly. This allows the algorithm to prevent the performance degradation and cost associated with unbeneficial/unnecessary handoffs.

Dedication

To my wonderful mother, Mariam, for her unconditional love, encouragement, and support. Also to my loving fiancé, Christian, for always believing in me and supporting my goals.

Acknowledgements

My experience as a graduate student at Virginia Tech was greatly enriched by the people that I had the privilege of meeting and working with. At the top of the list is my advisor, Dr. Nathaniel Davis. His guidance and support were essential to my academic growth as an undergraduate and graduate student. His input and feedback were critical during the research process and the writing of this document. I would also like to thank my committee members, Dr. Scott Midkiff and Dr. Luiz DaSilva for taking the time to discuss my research goals and providing me with helpful feedback.

This thesis and its underlying details were greatly influenced by the mentorship and assistance of my friend, Dr. Scot Ransbottom. I learned a great deal about research from Scot during our many lengthy meetings. I extend my thanks to him for having so much patience with me and for always giving me his honest opinion.

From the first semester of my graduate program until I completed my thesis, Dr. Grant Jacoby acted as both a great friend and mentor. Grant's encouragement, support, and invaluable advice contributed to a number of wise academic and career–related decisions. I am very grateful for his kindness and will always cherish his friendship.

One person that has positively influenced my life every step of the way is my mother, Mariam. Her firm belief in the power of education was passed on to me at a very young age. Her vision of my future has provided me with a priceless blueprint to guide me through life. Her kindness continues to reach me across thousands of miles without fail. For that and for the many gifts that she has given me, I extend my gratitude and love to her.

I would also like to extend my thanks to the rest of my family in Kuwait. Their moral and financial support of my ambition to earn a graduate degree was a major factor in my ability to do so. I will always be grateful for their incredible generosity.

Gratitude is also due to my loving fiancé, Christian. His encouragement and patience, particularly during the demanding research and writing periods of this thesis, were much needed and appreciated. His strong belief in the importance of higher education complemented my own and resulted in a supportive environment where I could follow my dreams and pursue my goals.

My graduate experience would not have been complete without my wonderful friends in Blacksburg. While the list of friends is long, I will limit my specific thanks to those that have had to listen to me complain about the pains of research. To Joshua Edmison, Michael Thompson, and Alexandra Poetter, thank you for listening and for all the fun times.

Table of Contents

List of Figures

List of Tables

Glossary of Acronyms

3G	Third Generation
3GPP	Third Generation Partnership Project
A-GPS	Assisted Global Positioning System
AAA	Authentication, Authorization, and Accounting
AP	Access Point
AN	Access Network
BAAA	AAA Broker
BCCH	Broadcast Control Channel
BSS	Base Service Station
C-plane	Control Plane
CN	Core Network (for UMTS) or Correspondent Node (for Mobile IP)
CS	Circuit Switched
DAD	Duplicate Address Detection
DGPS	Differential GPS
DS	Distribution System
ESS	Extended Service Set
ETSI	European Telecommunications Standards Institute
FA	Foreign Agent
FAAA	Foreign AAA server
FDD	Frequency Division Duplex
GPS	Global Position System
GSM	Global System for Mobile Communications
HA	Home Agent
HAAA	Home AAA server
ITU	International Telecommunications Union
L3	Layer-3
LCS	Location Services (or location-based services)
LMU	Location-management unit
ME	Mobile Equipment
MN	Mobile Node
PCF	Position Calculation Function
PHY	Physical Layer
PLMN	Public Land Mobile Network
PPL	Predicted Path Length
PRCF	Position Radio Coordination Function
PRRM	Position Radio Resource Management
PS	Packet Switched
PSMF	Position Signal Measurement Function
QOS	Quality of Service
RAB	Radio Access Bearer
RNC	Radio Network Controller
RNS	Radio Network Subsystem
RPOA	Recognized Private Operating Agency
RRC	Radio Resource Control

RSS	Received Signal Strength
RVD	Required Visit Duration
SMLC	Serving Mobile Location Center
SNR	Signal to Noise Ratio
SRNC	Service Radio Network Controller
SQL	Structured Query Language
TEP	Trajectory End Point
TDD	Time Division Duplex
U-plane	User Plane
U-TDOA	Uplink Time Difference of Arrival
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
UTRAN	UMTS Terrestrial Radio Access Network
UVD	User Visit Duration
VHDC	Vertical Handoff Decision Controller
WCDS	WLAN Coverage Database Server
WLAN	Wireless Location Area Network

Chapter 1: Introduction

Until recently, the concept of wireless communications consisted of providing voice services over radio channels. However, an increasing number of wireless telecommunication providers are now offering both voice and data services to their users. These services and the emerging multimedia services demand higher data rates to achieve a better quality of service (QoS). Therefore, technologies such as the Universal Mobile Telecommunication System (UMTS) were introduced to provide more bandwidth in order to satisfy the QoS requirements [1]. UMTS is a Third Generation (3G) [4] mobile radio access technology and is seen as the successor to 2G and 2.5G systems such as GSM [1] and GPRS [3], respectively. In addition to offering a larger bandwidth, UMTS was designed to allow for global mobility, while maintaining the level of service offered to the user at his/her home network.

Certain characteristics of the 3G/UMTS technology have made it less suitable for small, indoor, and densely populated areas. Therefore, researchers and standardization bodies have considered other types of technologies that could be utilized to extend 3G/UMTS network services in such areas. The characteristics of Wireless LAN (WLAN) technology allow it to provide high data rates in indoor, small, and higher population areas. This made it the primary candidate for extending 3G/UMTS in a complementing manner. Since the IEEE 802.11 standard family [2] is the most common type of WLAN technology in the United States, the majority of the research efforts for integrating 3G/UMTS with WLAN technology have been geared towards using it in their efforts. However, several other research efforts have considered integrating other flavors of 3G and WLAN standards, as further addressed in Chapter 2. For this research effort, UMTS and IEEE 802.11 were used as the main representative technologies of 3G and WLAN, respectively. Note that "WLAN" and "802.11" are used interchangeably in this document and "3G/UMTS" is used to refer to UMTS or 3G.

As expected, integrating two very different technologies, such as 3G/UMTS and IEEE 802.11, introduced a number of technical and logistical issues that must be resolved in order to maximize the benefits reaped from such integration [5]. As described in

Chapter 2, 3G/UMTS and IEEE 802.11 are heterogeneous technologies that have different attributes (i.e., data rates, structure, etc.), mechanisms (i.e., coverage discovery, handoff decision, etc.), and administrative domains. Handoff between access points or base stations has long been an issue within the wireless telecommunication field. However, a higher level of handoff complexity, and thus issues, is introduced due to the differences between inter-networked heterogeneous wireless networks. This added complexity causes additional delays to the handoff process and utilizes more resources, from the mobile node's (MN) perspective and the network's perspective, as further addressed in Chapter 3. These handoff-related delays typically result in a period of instability caused by the interruption in data transfer as the handoff procedures take place and the MN's traffic gets re-routed to its new location (i.e., over the new connection). The longer the handoff-related delays the more severe the performance degradation, particularly with respect to the application-level throughput and response time.

The handoff from 3G/UMTS to WLAN is not done out of necessity (i.e., the need for coverage) but rather for performance optimization purposes. Therefore, ideally such handoff should only be triggered if it will result in enhanced performance. Typically, the data rates offered by WLANs are significantly higher than those offered by 3G/UMTS networks. Therefore, the degradation in performance that occurs as a result of the handoff-related delays is compensated for through the higher data rates, and thus does not affect the overall level of performance observed by the user. This is only possible if the user's visit to the WLAN coverage area is long enough to:

a) Complete the reconfiguration and mobility management procedures, the security procedures, and the accounting procedures [6] (as shown in Figure 3-1)
b) Allow the recovery of the upper layer protocols and the application(s) after the handoff, in terms of getting used to the new data rates and returning to a stable performance level (i.e., take appropriate advantage of the higher data rates)
c) Transfer sufficient user data packets to compensate for the interruption or throttling in data transfer, which took place during the handoff procedures

1.1 Problem Overview

It is expected that the number of deployed WLANs will increase in the observable future to accommodate the plans for extending 3G/UMTS networks with IEEE 802.11 cells, in areas of high user density and relatively high mobility (e.g., airports, subway stations, campuses, city blocks, parks, convention centers, etc.). Moreover, the number of wireless users has increased drastically in the previous years and continues to grow at a significant rate. Therefore, in order to maintain an acceptable level of QoS (in terms of bandwidth) the size of the WLAN coverage areas will become smaller. Therefore, a 3G/UMTS user that is traveling at a non-negligible speed will cross through a series of small WLAN coverage areas in a very short time period. Due to the benefits achieved from transferring users from 3G/UMTS to WLAN, the default procedure is expected to trigger a handoff from 3G/UMTS to (the higher data rate) WLANs upon discovering coverage availability. Therefore, the rate of handoff will increase significantly due to the user's short visit duration to the discovered WLAN(s) coverage areas.

In most technologies, the conventional approach for coverage discovery and handoff decisions is based entirely on RF measurements (i.e., received signal strength, and signal to noise ratio, etc.). A few technologies such as UMTS, take into consideration other factors in addition to the RF measurements. For instance, the base stations in UMTS are aware of the status of their neighboring base stations and can transfer users to neighboring cells merely for load balancing purposes [7]. However, the IEEE 802.11 standard does not have such capabilities, and thus relies primarily on physical layer (PHY) measurements and the MAC layer evaluation of such measurements to determine the appropriate time for triggering the handoff.

According to the current state of technology, a dual-mode (UMTS-WLAN) mobile node (MN) would rely entirely on RF measurements [2] for discovering WLAN coverage availability and for deciding on the suitability of the handoff to the discovered WLAN. Therefore, in environments that have smaller WLAN cells and high user mobility, the conventional RF-based handoff decision method could trigger unnecessary handoffs to WLANs. These handoffs are unnecessary and inefficient since the user does not remain in the WLAN's coverage area long enough to benefit from its higher data rates.

Recall that handoff procedures, and particularly those needed for handoff across heterogeneous networks, have significant delay. Therefore, by the time the handoff procedures have completed and the stabilization period was over, the user could already be headed out of such coverage area and requiring a handoff back to UMTS or to another WLAN. As with any handoff, this degrades performance in terms of interrupting data transfer and increasing application response time. Moreover, due to the short visit duration the application does not have the chance to benefit from the higher data rates, which compensate for the handoff-related delays. Therefore, the performance degradation will have a noticeable effect from the user's perspective. This defeats the purpose and counters the objective of inter-networking 3G/UMTS and IEEE 802.11 and performing handoffs to WLANs.

The problem described here is a result of the limitations of the conventional handoff decision method, which lacks awareness of the user's position, speed, and direction. This knowledge is necessary to prevent inefficient handoffs to discovered WLANs, which only result in performance degradation and unnecessary costs. Therefore, a different approach that combines the RF measurements with the location and spatial parameters of the user and the WLAN coverage area is needed to mitigate this issue and prevent its performance-degrading effects.

1.2 Thesis Objective

The objective of this research effort is to analyze the feasibility and benefits of augmenting the conventional RF-based handoff decision mechanism with the proposed location-based evaluation to develop a robust location-aided handoff decision algorithm for making handoff decisions in heterogeneous wireless overlay networks. The underlying objective is to eliminate or at least reduce the triggering of handoffs that result in performance degradation and wasting of resources due to the user's short visit to the discovered network's coverage area.

1.3 Solution Overview

The proposed solution was based on what is becoming commonly known as "location-based information"[8]. Location-based information has received a great deal of

attention, particularly in recent years. The emergence of the E-911 service, to locate users calling the U.S. emergency line from a wireless node, as well as the emerging commercial applications utilizing location-aware information, have prompted further research to support such services. The relevant standardization bodies have developed architecture and protocol plans to support such services in technologies such as 3G/UMTS [9]. In addition to utilizing location-based information in user applications or emergency situations, the standardization efforts have researched the use of location-based services for network optimizations purposes [10][11]. Therefore, these new developments were utilized in this research effort to study the feasibility of a location-aided solution, which consists of combining the RF-based conventional handoff decision mechanism with a location-based evaluation mechanism.

A location-aided handoff decision algorithm was developed to run a location-based evaluation at a node in the 3G/UMTS control network. The RF measurements at the MN's WLAN adapter were used to discover WLAN coverage and to determine the suitability of the handoff from an RF-perspective (i.e., based on Received Signal Strength, Signal to Noise Ratio, etc). Upon discovery of WLAN coverage, the MN triggers the location-based handoff evaluation at the UTRAN and waits for the response before proceeding with the Layer-3 handoff procedures to the discovered WLAN. The different aspects of Layer-1 to Layer-3 handoffs are described in Chapter 3 and further addressed in later chapters.

The location-based evaluation relies on a positioning method for obtaining user location fixes. Once the user's position, speed, and direction are computed, the evaluation takes place to predict the user's trajectory with respect to the WLAN's spatial characteristics. The evaluation node will accordingly predict the user's trajectory and whether he/she will remain in the same WLAN coverage area for the necessary duration to result in benefit (i.e., allow the application to transfer sufficient data over the higher data rate connection). If so, then the MN receives a notification from the UMTS network recommending the handoff. Otherwise, the evaluation cycle will periodically trigger to re-assess the user's trajectory. To avoid infinite triggering of the evaluation cycle, the MN notifies the UMTS network if the user exits the WLAN coverage area, which is determined by the PHY layer measurements obtained through the WLAN adapter.

The underlying mechanism for estimating the user's trajectory and for predicting the user's visit duration, are described in Chapter 4 along with the designs requirements, assumptions.

1.4 Validation & Analysis Overview

The performance of the algorithm was demonstrated and analyzed by comparing it to the performance of the conventional handoff decision method for making decisions in heterogeneous wireless overlay networks. Two primary scenarios were used to demonstrate the weaknesses of the current conventional handoff decision method and demonstrate the capabilities of the proposed algorithm, which was designed to tackle the source of such weaknesses. The limitations of the algorithm were also addressed in the context of the worst-case scenario. Moreover, the algorithm's benefits were demonstrated by comparing the outcome of the conventional handoff decision method to that of the algorithm's outcome, in a quantitative manner. The metrics for comparing the outcome of the two methods were the application-level throughput and application response time. Finally, based on such evaluation and analysis of the proposed algorithm's capabilities and limitations, specifications were established to point out the type of environments that would be most suitable for deploying the algorithm.

1.5 Thesis Outline

Chapter 2 provides further details regarding the technologies involved in this research effort. Chapter 3 provides a description of the factors that motivated this research effort and further details regarding the problem addressed in this chapter. Chapter 4 includes a description of the proposed location-aided handoff decision algorithm, and its associated requirements, assumption, and specifications. Chapter 5 validates the design through evaluating and analyzing the proposed algorithm's potential performance, limitations and benefits. Finally, Chapter 6 summarizes the conclusions reached at the end of this research effort and provides a description of the potential future work. It also briefly addresses some of the logistics and potential obstacles that maybe associated with deploying the proposed algorithm in a commercial setting.

Chapter 2: Background

2.1 Introduction

This chapter is intended to familiarize the reader with the concepts involved in this research effort. It provides an overview of the main technologies addressed in the problem space and investigated solution. As described in Chapter 1, this research effort was based on an environment where heterogeneous wireless overlay networks were assumed to be complementary and integrated at some level. This chapter further addresses this concept and describes the two technologies that were chosen to represent heterogeneous wireless overlay networks.

Section 2.2 provides an overview of Third Generation technology and focuses on UMTS as the representative standard of 3G for this research effort. Section 2.3 describes WLAN technology and the basic architecture of the IEEE 802.11 standard, which is the chosen standard to represent WLAN in this research effort. The integration benefits and relevant logistics are addressed in Section 2.4. Section 2.5 provides a description for a network handoff and categorizes handoff types according to a number of factors, particularly with respect to UMTS and WLAN. Section 2.6, describes the two main candidates considered for positioning in 3G networks. Finally, Section 2.7 summarizes the chapter and addressed topics.

2.2 Third-Generation (3G) Technology Overview

The standards for "Third Generation (3G)" technology were initially developed by the private sector and cellular community rather than by a formal standards organization. 3G-technology refers to a generation of wireless systems that offers higher data rates than previous generations and a variety of enhanced services. These services range from email and web-access to the more demanding multimedia applications. There are three primary standards that comprise 3G technology: W-CDMA, CDMA2000, and TD-CDMA [51]. The "Universal Mobile Telecommunication System" (UMTS) supports the W-CDMA standard and is currently deployed in Europe. UMTS is based on the Global System for Mobile Communications (GSM). GSM is the most widely used

2G standard and accounts for over 65 percent of the global wireless market [52]. Therefore, UMTS was chosen to represent the 3G technology in the context of heterogeneous wireless overlay networks throughout this research effort.

Despite the differences between UMTS and other 3G standard types, the specifications of this study can be applied to other 3G standards with minimal modifications. In fact, the study should apply to a variety of heterogeneous overlay networks that support handoffs across their boundaries and with other types of networks. The following subsections provide an overview of UMTS from a perspective that supports the focus of this research effort.

2.2.1 UMTS Standardization

The International Telecommunications Union (ITU) adopted the standardization of 3G technologies in order to homogenize the various flavors of 3G that were being developed. The standard was named International Mobile Telecommunications-2000 or IMT-2000. The ITU defined a list of requirements for the 3G systems [12] including a requirement to support:

- High data rates (i.e., 144Kbps to 2Mbps)
- Packet Switched (PS) and Circuit Switched (CS) data transfers
- High speech quality
- IP services
- Efficient use of the allocated spectrum
- Backward compatibility for smooth transition from 2G to 3G

A number of proposals were submitted to the ITU for the standardization of 3G. Third Generation Partnership Project (3GPP) is currently the main governing body for standardizing UMTS while (3GPP2) is the standardization body for cdma2000. The standardization of UMTS is a work in progress and is continuously updated and published by 3GPP in the form of "Releases."

2.2.2 UMTS Architecture

Global Systems for Mobile Communications (GSM) is the most widespread 2G technology currently used in cellular communities around the world. 2G-technology

refers to digital, circuit-based, narrowband systems that are suitable for voice and limited data communications. The architecture of UMTS networks is based on GSM architecture. The basic UMTS architecture can be divided into three main sections: The Core Network (CN), the Radio Network Subsystem (RNS), and the Mobile Node (MN). The MN may also be referred to as "User Equipment" (UE) or "Mobile Equipment" (ME) when associated with UMTS. Figure 2-1 shows the basic division of the UMTS architecture. Note that an RNS in UMTS is also commonly referred to as a UMTS Terrestrial Radio Access Network (UTRAN) and may be used interchangeably.

This type of network is also commonly known as a Public Land Mobile Network (PLMN), which is defined by the standard [13] as a network "established and operated by an administration or Recognized Private Operating Agency (RPOA) for the specific purpose of providing land mobile telecommunications service services to the public."

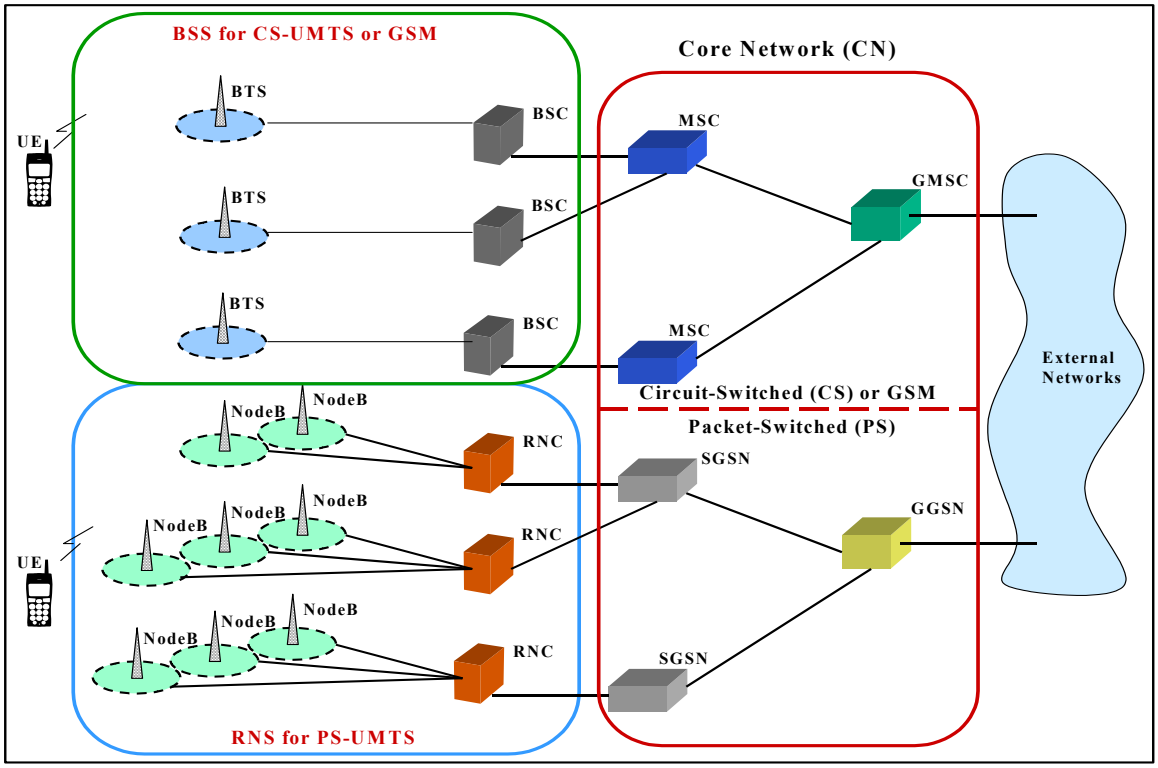

Figure 2-1. Basic UMTS Architecture based on information from 3GPP TS 23.002 [13]

Each section in the UMTS architecture is associated with a number of defined responsibilities. The core network (CN) is the structure responsible for transporting user data to its destination. Therefore, it involves the use of a number of switching entities

and gateways to external networks such as the Internet. It must also maintain information regarding the user's access authorizations, cell location, and status. Therefore, the CN also includes databases that store user profiles, mobility management information, and billing/accounting information.

From a different perspective, the architecture of the access plane of UMTS can be divided into two parts based on the supported switching method for exchanging information. The Circuit Switched (CS) method is based on reserving resources when establishing a connection between two communicating systems, whereas the Packet Switched (PS) method uses packets to carry data. These packets can take different routes to their destination and thus no reservation of resources is necessary.

The UTRAN [14] is the main UMTS Access Network (AN) unit that consists of fixed entities connected to nodes in the CN and provides an air interface to be used by the MNs to access services. Therefore, it is also responsible for managing radio resources and related functionalities. Each UTRAN consists of a number of Radio Network Subsystems (RNS). The RNS is the PS domain's access network, which is connected to the CN via a specific interface. It consists of a combination of Node-Bs and a Radio Network Controller (RNC). The RNC is the central node in a radio access network, whereas the Node-B is a base station that can serve more than one cell at a time. In addition to providing the air interface to the MN, the Node-B is responsible for taking measurements and providing the RNC with information about the cell's condition. This information is essential to the RNC for making appropriate decision about power management, load balancing, radio resource allocation, and handoff decisions.

The standardization efforts by 3GPP proposed an All-IP UMTS core network in [15]. This scheme would introduce some changes to the architecture described above, including the elimination of CS nodes. The All-IP approach will support CS and PS functions using PS entities and will provide all services over-IP.

2.2.3 UTRAN Protocol Stack

The protocol stack in UMTS between the MN and CN can be divided into two categories: "Control Plane" (C-plane) protocols and "User Plane" (U-plane) protocols. The C-plane protocols are responsible for exchanging control information between parties, prior to their communication, in order to reserve the required resource (Figure 2-2). On the other hand, the U-plane is responsible for transmitting user data (Figure 2-3).

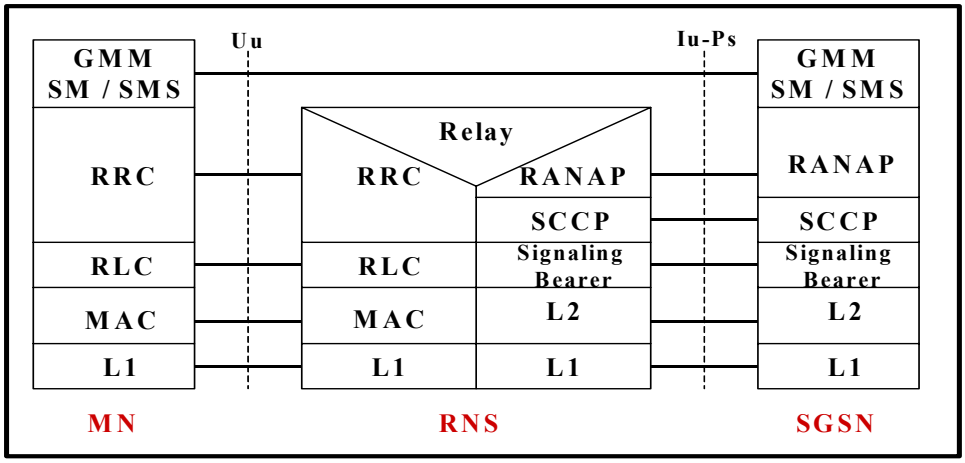

Figure 2-2. Control Plane (C-plane), based on information from 3GPP TS 23.060 [16]

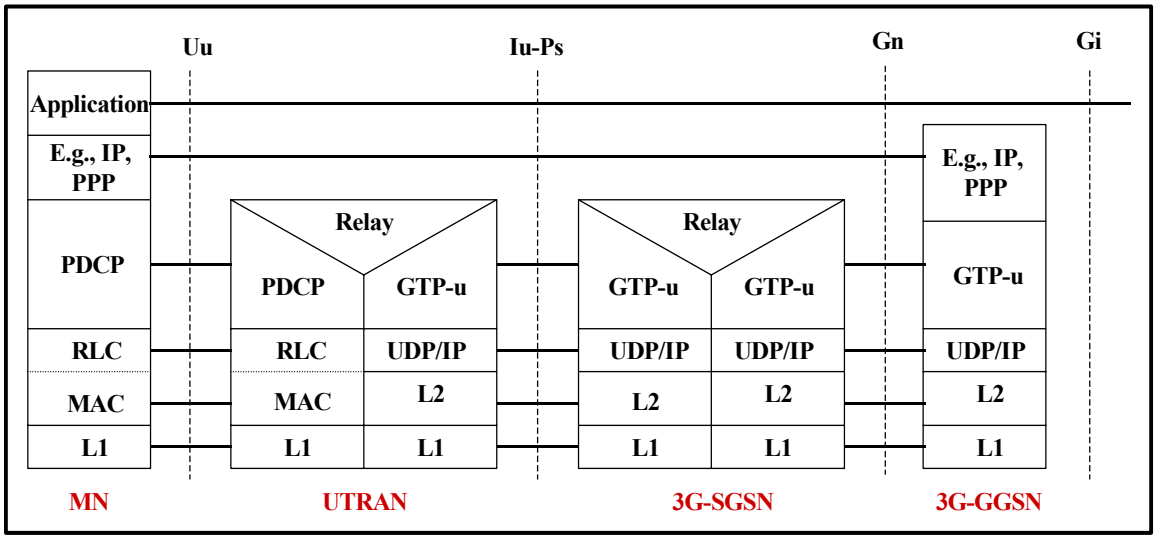

Figure 2-3. User Plane (U-plane), based on information from 3GPP TS 23.060 [16]

RRC Protocol

The Radio Resource Control (RRC) protocol is the most relevant to this research effort. It is mainly responsible for handling the layer-3 signaling in the C-plane. This signaling and the interaction between RRC and the layers below it is shown in Figure 2-4.

It is also responsible for satisfying requests from higher layers to establish, reconfigure, and release radio- bearers in the U-plane (i.e., the access stratum for transferring user data). Furthermore, it evaluates measurements relayed to it by lower layers (L1 & L2) then executes the necessary functions to support mobility procedures, such as paging and handoffs [17].

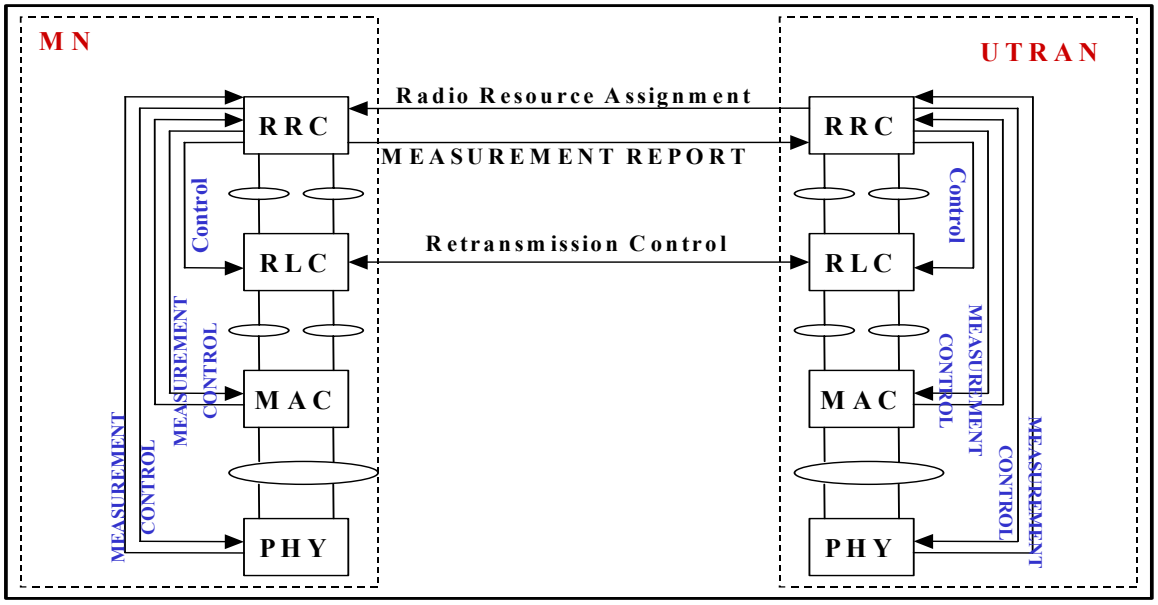

Figure 2-4. Interaction between RRC and the lower layers (Based on information from 3GPP, TS 25.301) [69]

2.3 Wireless LAN Technology Overview

Wireless Local Area Network (WLAN) is a technology that provides high-capacity IP connectivity to MNs in local (limited range) areas. WLAN is not a technology standard, instead it encompasses a number of standards that have been developed to meet certain needs and requirements. Although its initial deployment was in the corporate environment, its current usage has been extended to public areas (e.g., airports, coffee shops, and limited outdoor areas) and individual residential areas. A wireless hotspot consists of a group of MN-users accessing the Internet through wireless connectivity provided locally by a business or organization. Hotspots are currently based on the capabilities of WLAN technologies such as the IEEE 802.11 standard family [2]. The most widespread standard used in public hotspots is the IEEE 802.11b standard (Wi-

Fi). This research effort uses the IEEE 802.11 standard family to represent WLAN technology within the context of heterogeneous wireless overlay networks.

High Performance LAN type 2 (HiperLAN/2) is another WLAN standard proposed by the European Telecommunications Standards Institute (ETSI). It is claimed to meet the challenges currently facing 802.11 standard such as security, scalability, QoS, and ease-of-use issues [19].

2.3.1 WLAN Architecture

The architecture of a WLAN system is described here according to the IEEE 802.11 standard. The main building block of such architecture is the Base Service Set (BSS). There are two modes of configuration that can be used within the standard: the "Infrastructure" and the "Ad Hoc" mode. In the "Infrastructure" mode (Figure 2-5), which is the mode of configuration assumed for this research effort, the different BSS are interconnected with each other via a component called a Distribution System (DS).

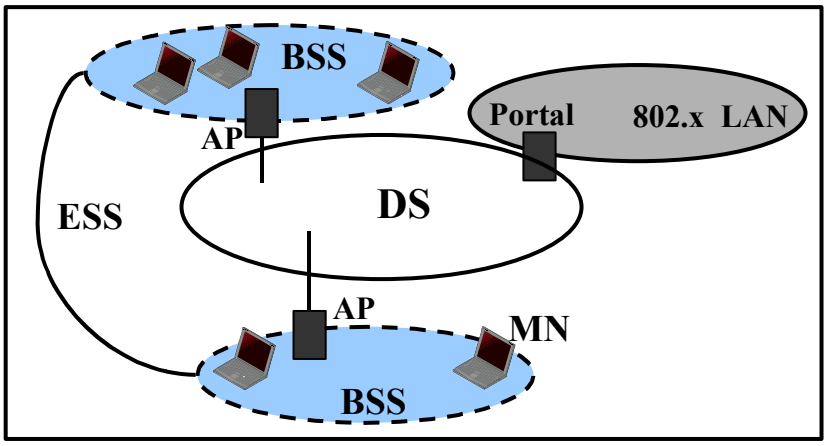

Figure 2-5. WLAN architecture (Infrastructure mode) based on information from the IEEE 802.11 standard [2]

Each BSS has a single Access Point (AP), which provides the MNs with access to the DS. These interconnected components form the Extended Service Set (ESS). The ESS is a large coverage area where MNs can move from one BSS to another without changes or notification to higher layers in the protocol stack. Finally, a "Portal" is required to integrate the WLAN architecture into the wired network (e.g., Ethernet) and may be integrated with an AP in a single device attached directly to the DS. In the "Ad

Hoc" mode, each MN can directly reach any MN in the BSS without going through an intermediate node (i.e., AP). The BSS is also not connected to the wired network.

2.4 UMTS-WLAN Integration

Academic researchers and industry leaders are currently studying the concept of integrating UMTS and WLAN technologies [20] in a manner that will enhance the overall performance of the network and provide an enhanced level of service. The common approach for this integration consists of extending 3G-UMTS networks with available or new WLAN public hotspots. The following is an overview of the integration.

2.4.1 Integration Benefits

The benefits of integrating UMTS and WLAN technologies are becoming increasingly appreciated in the wireless community. The most noticeable advantage stems from the significantly higher data rates offered by WLAN hotspots, particularly in comparison to data rates offered by UMTS. Table 2-1 shows the variation in data rates offered by UMTS networks, depending on factors such as the type of mobility and the range of coverage. Theoretically, the best-case scenario allows UMTS to offer up to a 2Mbps data rate. In practice, the achieved data rates are significantly below that figure. On the other hand, WLAN data rates offered by the IEEE 802.11 family are as high as 11 and 54Mbps. However, as with UMTS, the actual data rates achieved in practice are typically less. Regardless of the exact data rates achieved, WLAN data rates typically exceed those offered by UMTS networks under any normal conditions.

Table 2-1. UMTS Data Rates according to Cell Size, Speed, and Applications Type

Type of Service	Data Rate
Pico-cell and Micro-cell coverage, Pedestrian speed (<10km/h)	2.048 Mbps
Macro and small Macro-cell coverage, Medium vehicle speed	384 kbps
Large Macro-cell coverage, High vehicle speed	144 or 64 kbps
Very large cells and continuous low speed data application	14.4 kbps
Very large cells and Speech (real-time) applications	12.2 kbps
Global coverage through satellite	9.6 kbps

Another potential benefit is WLAN's suitability for indoor, low mobility, and small areas coverage environments. Therefore, they provide for better overall performance in areas where UMTS networks are particularly less suited. A study conducted under the framework of the IST ROMANTIK project funded by the European Union in 2003 demonstrated another benefit of the UMTS-WLAN integration. The project utilized "novel ray tracing," software-simulated physical layer performance results, and optimal base station deployment analysis. This information demonstrated, in a quantified manner, the enhancement in capacity of 3G networks when extended with WLAN hotspots. The study showed that in a 1km x 1km area, the number of supported connections increased by over a 1000 when 15 hotspots were deployed to extend the 3G-network [21].

From a different perspective, another attractive characteristic of WLAN technology is its use of the less costly unlicensed spectrum. Finally, the relatively low cost and simple setup for deploying a WLAN hotspot is a significant reason for making it an attractive approach to extend UMTS networks. In fact, an average WLAN hotspot can be set up for as little as 1% of the cost required to setup a single 3G base station.

2.4.2 Integration Logistics & Requirements

The integration of UMTS and WLAN technologies demands a significant amount of cooperation between these two wireless communities. There are a number of requirements and logistical issues involved in the integration process, from a technical and non-technical perspective. However, much of the non-technical issues such as billing and access control rely on technical capabilities. This includes issues related to architecture integration, billing issues, security-related differences, QoS policies, etc [5][20][26][53]. The aggregate affect of these integration issues has a direct impact on the seamless mobility between the two heterogeneous technologies. Therefore, a common protocol is needed to harmonize the technologies and separate upper layers from the heterogeneous lower layer technologies and standards of UMTS and WLAN. Since IP is a protocol common to WLAN and UMTS networks, it has been considered as the main underlying protocol for the integration. In addition, using the IP protocol in this case has the advantage of facilitating the use of the IETF proposed protocols (i.e., Mobile

IP and AAA [22][23][24][25]). 3GPP has considered such protocols in their feasibility and planning stages of the internetworking process between 3G and WLAN technologies. Further details will be presented in the next chapter with regards to the impact of using such IETF protocols on the handoff process between UMTS and WLAN networks.

2.5 Handoff Overview

2.5.1 Brief History

For first generation mobile radio systems, the concept of a cell consisted of a very large coverage area (150km radius) that was supported by a single transmitter. The infrastructure costs were consequently very low. However, due to the limited number of available frequency channels, the number of users per cell was also very limited. Due to the longer distance between the transmitter and the mobile stations, both had to transmit at higher powers in order to communicate. Therefore, mobile handsets were not a feasible solution; instead, a large terminal had to be installed into a user's vehicle [1].

The concept of "frequency reuse" was developed in order to accommodate a greater number of subscribers in a given coverage area. Its basic idea consisted of reducing the size of a cell's coverage area by decreasing the transmitting power of the cell's base station. Therefore, the same frequencies could be reused within relatively small distances of each other without a significant amount of interference (Figure 2-6).

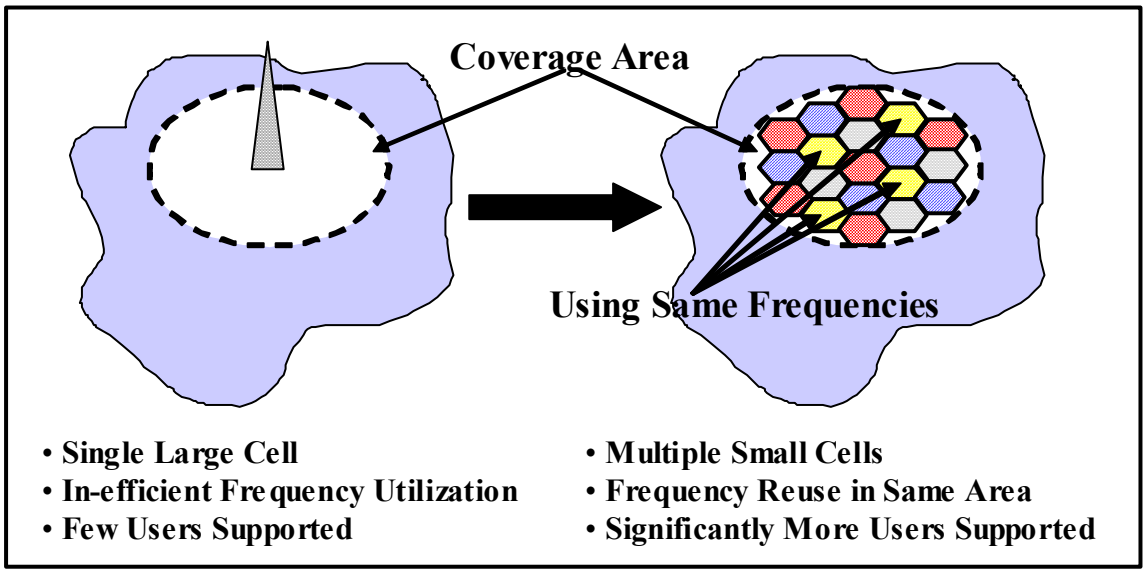

Figure 2-6. A comparison between First Generation coverage concepts and the cellular concept allowing for Frequency Reuse

The number of allocated channels per cell would vary according to the number of cells in the cluster and the load within each cell. Today, the term "cell" is used in the wireless domain to refer to oval shaped, square shaped, circular, and hexagonal shaped coverage areas, all employing frequency reuse [1].

Frequency Reuse and small cellular coverage areas created a need for a mechanism to quickly switch a subscriber's connection from one cell to a neighboring cell, as the subscriber moved from one location to another. This mechanism became known as a "handoff" or "handover" in the wireless community. The handoff mechanism introduced a new level of complexity to the system. The intention has been to avoid involving the user in the handoff process and to conduct it without the user's awareness (i.e., seamless handoff). The following sub-sections address the common types of handoff, particularly with respect to WLAN and 3G systems.

2.5.2 Handoff Types

The categorization of handoff types depends on a number of factors, which are further divided based on the type of technology and network architecture where the handoff takes place. The handoff process is no longer limited to the boundaries of a single technology. In fact, the emerging concept of "Always-Best-Connected" [27] has blurred the line between different access networks. The following are the categorization factors along with the handoff types that are based on them.

Factor A: Technologies Involved

a. **Horizontal Handoff:** a term used to describe the handoff process of a MN between base stations supporting the same technology, also referred to as intra-technology handoff.

b. Vertical Handoff: a term used to describe the handoff process of a MN between base stations supporting different technologies, also referred to as inter-technology handoff. Examples of such handoff include but are not limited to a handoff between 2G and 3G base stations, a handoff between 3G and WLAN base stations, or a handoff between base stations supporting

different 3G standards (i.e., UMTS and cdma2000). There are two types of vertical handoffs:

- **Downward Handoff:** A handoff from a large network cell with low data rates to a smaller network cell with higher data rates. Discovering the availability of a new network is not delay sensitive in the context of a downward handoff, since the MN is typically not at the risk of exiting the large cell it is currently connected to. Instead, the handoff is typically initiated for performance optimization reasons (e.g., UMTS to WLAN).
- **Upward Handoff:** A handoff from a small network cell with high data rates to a larger network cell with lower data rates. Since the user experiences better performance with the higher data rates, it is typically desirable to remain connected to such a network for as long as possible. However, a handoff to a network with less data rates becomes necessary when the user is exiting the current small coverage area. Therefore, discovering the upper network is delay sensitive in the context of an upward handoff.

Factor B: Administrative Domains Involved

The term "Administrative Domain" was defined in [28] as "A collection of End Systems, Intermediate Systems, and Subnetworks, operated by a single organization of administrative authority. The components which make up the domain are assumed to interoperate with a significant degree of mutual trust among themselves, but interoperate with other Administrative Domains in a mutually suspicious manner." Based on this definition of Administrative Domains, handoffs can be categorized as:

a. **Intra-Administrative Domain Handoff:** a handoff process where the MN switches between base stations supporting the same or different technologies, managed by the same administrative domain.
b. **Inter-Administrative Domain Handoff:** a handoff process where the MN switches between base stations supporting the same or different technologies, managed by different administrative domains.

Factor C: Number of Connections Involved

This categorization mainly applies to the handoff process within cellular networks [49]:

a. **Hand Handoff:** A term used to describe a handoff that involves the MN maintaining a connection with only one base station at any given time (Figure 2-7). This process is sometimes referred to as "Break before you make." Hard handoffs may be seamless or non-seamless depending on their severity and whether they are noticed by the user in the form of an interruption in service (i.e., dropped call, terminated session, or significant delays).

b. **Soft Handoff:** A term used to describe a handoff that involves the MN always being connected to at least one base station when moving between cells (Figure 2-7). This process is sometimes referred to as "Make before you break." Soft handoffs are possible in situations where the MN is moving between cells that operate on the same frequency. It has been realized as an option in 3G systems.

a. **Softer Handoff:** A term used to describe a type of Soft Handoff that involves the MN switching connections over radio links that belong to the same base station (Figure 2-7). This type of handoff is possible in such networks where a base station serves several individual sectors of a cell (i.e., Node-B in UMTS systems).

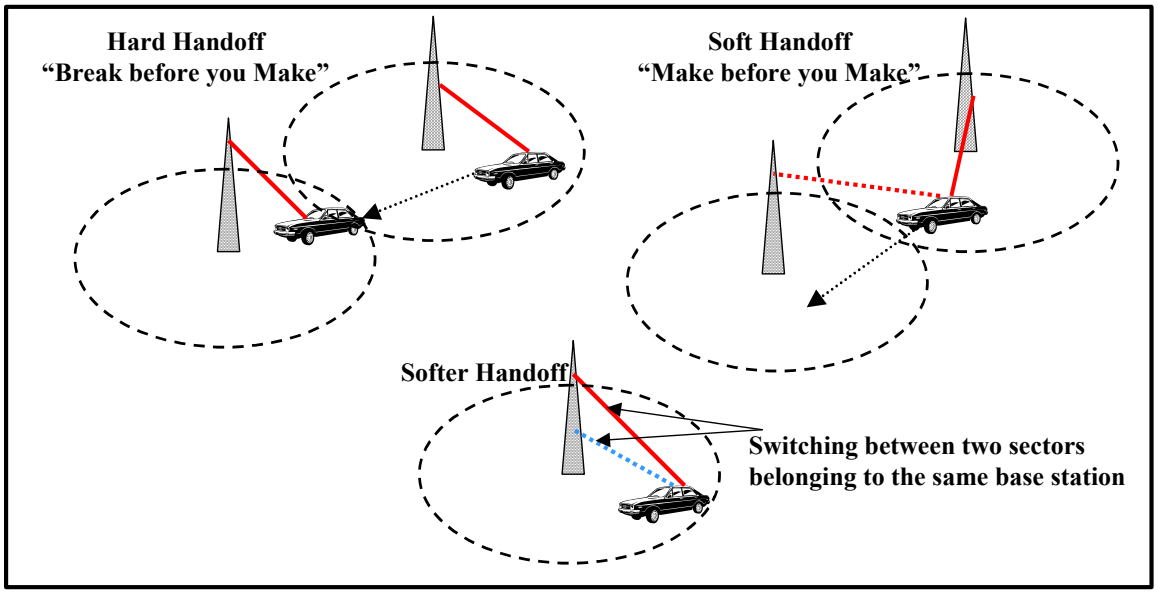

Figure 2-7. Three types of handoffs categorized according to the number of connections maintained by the MN as it changes its location

Factor D: Frequencies Involved

a. **Intra-Frequency Handoff:** a term used to describe the handoff process of a MN switching between stations that operate at the same frequency. This type of handoff may occur in 3G systems that use CDMA with Frequency Division Duplex (FDD[1]) allowing for a Soft or Softer handoff.
b. **Inter-Frequency Handoff:** a term used to describe the handoff process of a MN switching between stations that operate at different frequencies. This type of handoff is the only one supported in 2G systems (i.e., GSM) due to their reliance on TDMA[2] and FDMA[3] multiple access schemes, which require the change in carrier frequency. Another example is 3G systems that use CDMA with Time Division Duplex (TDD[4]). All inter-frequency handoffs are "Hard Handoffs."

Factor E: Layers Involved

a. **L2 Handoff:** An example to illustrate this type of handoff is the switching of a MN between two access points within the same ESS (Figure 2-8).
b. **L3 Handoff:** Following the example used in (a), an L3 handoff would take place when a MN switches between two APs in different ESSs (Figure 2-8). Therefore, handoff signaling will go through an intermediate router and Layer-3 signaling is required to manage the routing of data to the MN's new location.

[1] Frequency Division Duplex (FDD): a procedure to separate uplink and downlink transmission by using different frequency bands for transmitting and receiving
[2] Time Division Multiple Access TDMA separates the signals into consecutive time slots where each user can utilize the full bandwidth but for only short periods of time
[3] Frequency Division Multiple Access (FDMA) divides the spectrum into frequency channels, where individual users can utilize a separate channel for transmitting and receiving all the time
[4] Time Division Duplex (TDD): A duplex procedure where the MN and the base station take turns for transmission and reception.

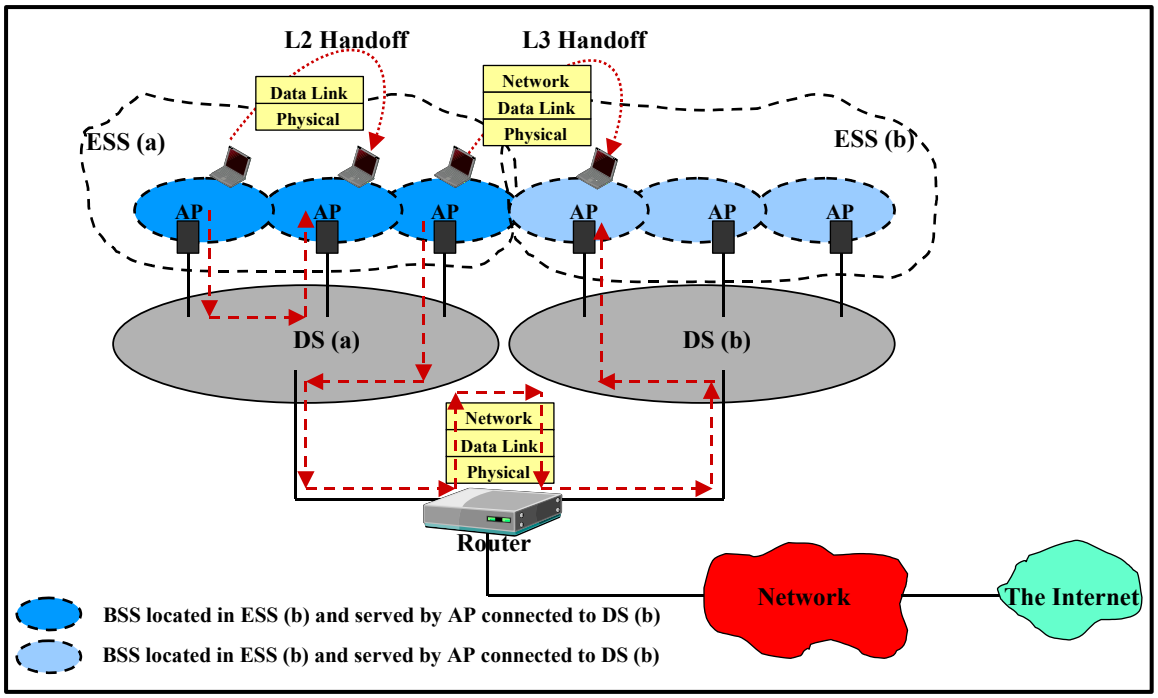

Figure 2-8. Layer-2 Handoff in ESS (a) vs. Layer-3 Handoff between ESSs (a) & (b)

Factor F: Network Elements Involved

This factor is dependent on the type of technology within which the handoff is occurring, due to the different elements/nodes and different hierarchy structures of each technology. For instance, an "inter-Node-B" handoff would refer to a handoff between different Node-Bs controlled by the same RNC (Figure 2-1), whereas an "inter-RNC" handoff would refer to a handoff between different cells served by different RNCs (Figure 2-1). On the other hand, WLAN uses a terminology that is centered on the BSS and ESS (e.g., "inter-BSS" would refer to a handoff between APs that belong to the same ESS and managed by the same DS).

2.6 Positioning Overview

As indicated in the title of this research effort, the approach proposed in Chapter 4 was dependent on the availability of location-based information and the positioning of the mobile node.

There are a variety of positioning methods for 2G, 3G, and other wireless technologies. This section provides an overview of the positioning methods that have been chosen by 3GPP to be integrated into the UTRAN. The standardization of

positioning methods follows 3GPP's effort to define the concept of location services (LCS) and develop an architecture and configuration to meet LCS needs. This includes the role of the positioning functions within the UTRAN itself.

2.6.1 Location Services (LCS)

The increased demand for location-based services has prompted 3GPP to standardize such services specifically for the UTRAN. In the standard, they are referred to as Location Services (LCS) and are described over three different stages in [9], [10], and [11]. Some LCS applications require the subscription to a provider while others can be utilized without any subscription (e.g., Emergency positioning situations). The following are the four types of LCS clients defined by the standard:

- **Emergency Services clients** are utilized when a user makes an emergency call
- **Lawful Intercept clients** are used for legal intercepts, depending on local regulations
- **PLMN Operator clients** are utilized to improve network performance and/or gather statistics for network optimization and enhancements
- **Value-Added Services clients**, which are typically applications running on the MN (e.g., navigation, weather, traffic updates, etc.)

The proposed algorithm in Chapter 4 falls under the "PLMN Operator Client" LCS client type. Therefore, the following overview mainly focuses on the aspects of the standard that support this type of client.

According to the standard, there are four main attributes that distinguish between LCS and have a significant impact on them. These four attributes are Accuracy, Coverage, Privacy, and Transaction-Rate. They are defined here with respect to LCS. Accuracy is the closeness of the estimated location to the actual location. Coverage is the area where sufficient QoS is observed at the MN. Privacy is the confidentiality level of the MN's location information. Transaction-Rate is the frequency of exchanging control messages.

2.6.2 Positioning Functions

The standard includes the categorization of the LCS functions into groups. The "Positioning Group" [29] includes four functions that are designed to operate in the UTRAN or in some cases at the MN itself. They are summarized as follows:

PRCF: The Position Radio Coordination Function (PRCF) manages the positioning task of a MN by coordinating resources (e.g., radio or GPS), communicating with other positioning functions, and handling positioning-failure and recovery.

PCF: The Position Calculation Function (PCF) is responsible for executing an algorithm that takes the previously obtained measurements as parameters and calculates the position of a target MN along with the achieved accuracy level.

PSMF: The Position Signal Measurement Function (PSMF) may be located at the MN, within a Node-B, or in a Stand Alone LMU node. It is responsible for collecting radio, and in some cases satellite, measurements to be used in positioning calculation functions. The positioning method(s) used will determine the type of measurements collected.

PRRM: The Position Radio Resource Management (PRRM) monitors the effect of LCS-related operations on performance of the network's radio resources. It also manages this effect by controlling frequency of measurement collection, signaling, and the acceptance or rejection of LCS requests according to their potential impact on overall performance.

2.6.3 LCS Architecture

To support LCS services in a PLMN, new LCS-specific entities [13] must be added in to the architecture. Figure 2-9 shows a UTRAN supporting LCS through additional entities and interfaces. The Serving Mobile Location Center (SMLC) can be implemented in the UTRAN within the serving RNC (SRNC) node or as a Stand Alone SMLC (SAS). The SAS can provide assistance data to a positioning unit (e.g., GPS unit) to enhance the performance of the positioning functions. Furthermore, unless the SMLC functionality in the SRNC is configured to calculate the final location estimate, the SAS

will act as a "location calculation server." The Location Measurement Unit (LMU) can be implemented either as a function of a Node-B or as a Stand Alone entity. It is responsible for taking measurements and providing them to the RNC periodically or upon request. These measurements (e.g., radio measurements) can be used to provide assistance to a positioning function in the UTRAN.

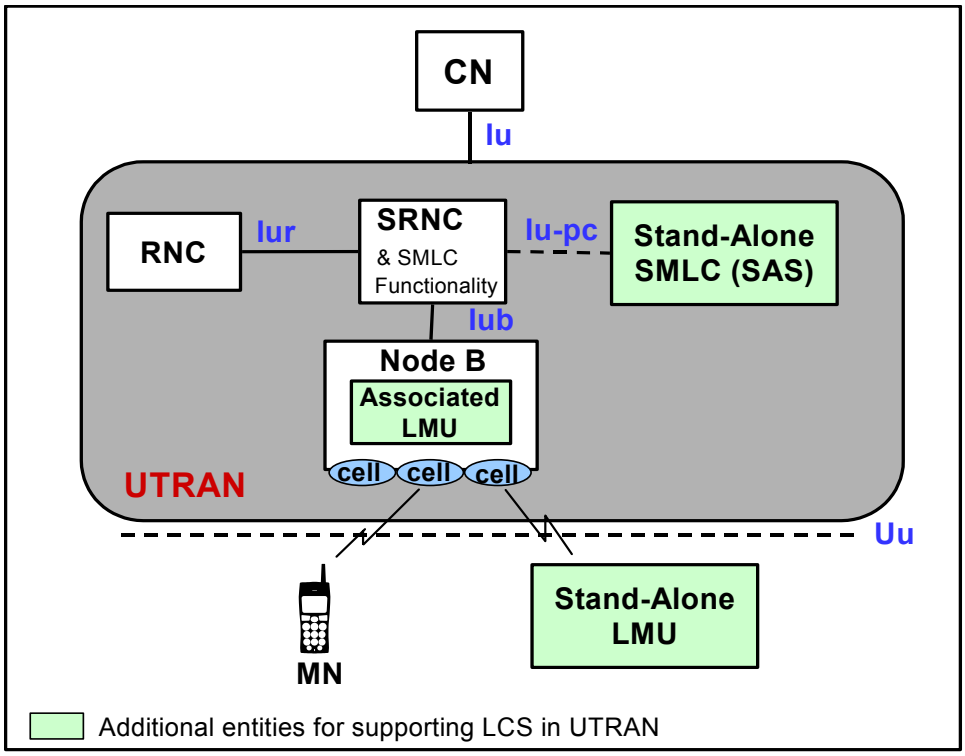

Figure 2-9. The architecture and configuration for supporting LCS in UTRAN, based on information in 3GPP TS 25.305, 2003

2.6.4 Positioning Methods

Some positioning methods have been chosen by the 3GPP to provide location estimates to LCS clients [29]. These methods use different approaches to obtain this information. Therefore, they each exhibit certain attributes that make them suitable for specific environments or specific positioning tasks. In some publications, a hybrid approach was suggested to enhance the positioning accuracy in different environments and to compensate for each method's shortcomings [30]. The following are the two promising methods for obtaining the position of 3G mobile systems.

2.6.4.1 A-GPS

The Global Positioning System (GPS) was developed by the U.S. Department of Defense in the early 1970s as a satellite-based navigation system for the military. Later on, it was extended and regulated for use by civilians as well. It provides positioning and information to a target anywhere in the world, independent of weather conditions. The stand-alone GPS method has been a useful tool in both military and civilian environments. However, some improvements were necessary in order to take advantage of GPS in mobile phones/nodes, which can only accommodate small modules with low power-consumption rates. In addition, a number of emerging location-based applications require enhanced GPS performance. This includes faster startup time, which is normally hindered by long acquisition times. It also includes signal detection capabilities, particularly in environments where GPS signals are weak or obstructed (i.e., inside buildings and "urban canyons"— The blocking of satellite signals due to the high density of tall buildings in urban areas).

Assisted GPS (A-GPS), as shown in Figure 2-10, consists of reference receivers that have a clear view of the sky and the constellation of GPS-satellites visible by the MN. Therefore, they can provide the mobile GPS receiver with assistance information for the measurement or calculation phase. The information can be provided through a broadcast and as a direct response to an "update request" from the MN (i.e., over a Point-to-point connection).

The A-GPS method was found by an extensive study [30] conducted at Motorola, Inc. to be the most accurate of all the standardized positioning methods in most outdoor areas and unobstructed areas. A-GPS is based on the concept of Differential GPS (DGPS), which is claimed to improve the location accuracy of the traditional GPS from 20m to 3-5m [46]. In fact, some of the literature claims an improvement from 20m to within 1m [54][55][67]. In addition, a few research efforts in "location-based handoff," which demand high accuracy, have been conducted based on an increasingly acceptable assumption. This assumption was based on the notion that the increasing public demand for location-based services and the emerging related-applications provide strong assurances that accurate positioning will become commercially available in the near future [46]

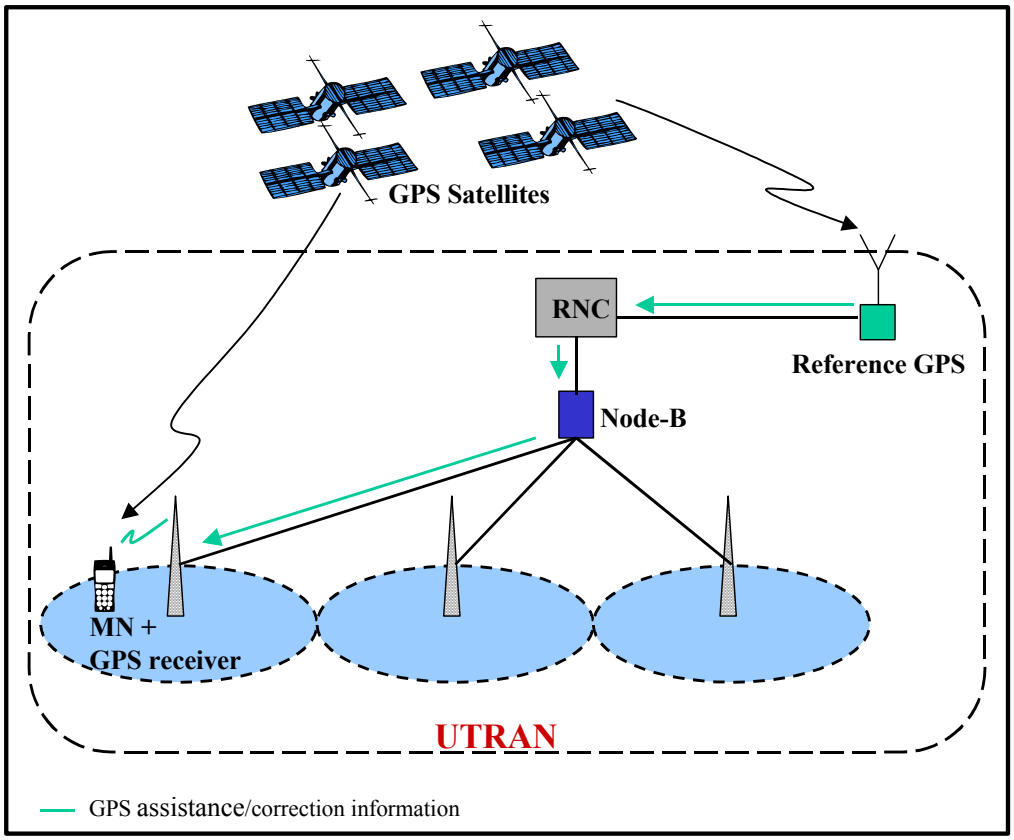

Figure 2-10. A-GPS positioning method in UTRAN

2.6.4.2 U-TDOA

The Uplink Time Difference of Arrival (U-TDOA) approach to positioning consists of locating a MN by comparing the time it takes a radio signal to reach several LMUs after leaving the MN. It does not require the handsets to be modified and thus can be used with legacy systems. It also does not require additional resources or add any noise to the channels during active positioning since it relies on the power of the signals transmitted by the MN during any uplink signaling.

TruePosition, Inc., a leading provider of wireless location-based technologies, announced on May 5th of 2003 [43] that U-TDOA has been formally standardized by the 3GPP—the official governing body for development and standardization of GSM and UMTS networks. According to the press release, this took place shortly after three US national GSM operators decided to use this method to support the FCC's E-911 requirement. Furthermore, 3GPP is in the process of considering it for UMTS systems. Proponents of the system claim that it will be particularly suitable for such technology

[44]. Their claim was based on the wide bandwidth and higher bit rates associated with UMTS, which have a lower spreading factors and thus higher power levels. Therefore, it would allow a greater number of LMUs to participate in the positioning process and thus provide better accuracy.

This particular positioning method has been found to perform exceptionally well in indoor and urban areas as well as suburban environments [44]. However, its performance is merely adequate in rural areas where the number of LMUs is lower and thus the accuracy level is not as high. The suitability of U-TDOA to UMTS systems as well as its suitability for use in indoor and urban areas make it a promising candidate for use in the implementation of the approach proposed in this research effort. This is mainly due to the fact that WLAN hotspots (indoor or outdoor) are typically located in indoor and/or urban environments. The initial predictive models of this approach claim that its current accuracy is within 25m but will improve as the number of LMUs increases.

2.7 Summary

This chapter provided an overview of a number of technologies, concepts, and issues that are relevant to the overall study conducted during this research effort. UMTS and IEEE 802.11 standards were described with respect to their characteristics, standardization, and basic architectures. The benefits of integrating the two technologies were then addressed to describe the interest in internetworking 3G and WLAN. The logistics of the integration were also visited, particularly with regards to mobility and the protocols involved. Since the core of this research effort was focused on improving the handoff process, a small section was dedicated to the description of the history and types of handoffs, particularly in 3G and WLAN. Moreover, the chapter addressed the 3GPP standardization of Location-based Services (LCS) and their proposed augmentation of the UTRAN architecture to support such services. Finally, two primary positioning methods standardized by 3GPP were described to address their core concepts and capabilities for providing accurate positioning of mobile nodes.

Chapter 3: Motivation

3.1 Introduction

The increasing demand for enhanced wireless access and the emerging developments in a variety of wireless data/multimedia services have prompted the current efforts to deliver such services over various types of networks and technologies. Each of these technologies has its unique set of characteristics and capabilities with respect to bandwidth, QoS, capacity, etc. Therefore, they can provide the user with a range of service levels in both adjacent and overlapping coverage areas. Therefore, a growing trend of internetworking these heterogeneous-wireless-overlay networks (e.g., UMTS and WLAN) has emerged to enable the user to take advantage of the emerging services without sacrificing transparent/seamless mobility.

The details regarding the motivation for internetworking 3G/UMTS and WLAN were addressed in Chapter 2. Recall that the primary motivation was due to the higher data rates offered by WLAN and the added capacity achieved by extending UMTS networks with WLAN hotspots [21]. Therefore, when WLAN coverage is available, a downward vertical handoff from 3G/UMTS to WLAN is usually desirable. However, there are certain situations where switching from UMTS to WLAN would not be efficient and/or would result in performance degradation, despite the better service level offered by the target network. This chapter describes such situations and provides the motivation for developing a solution that would predict the occurrence of these events and make the handoff decision accordingly.

Recall from Chapter 1 that handoff, particularly between heterogeneous wireless networks that belong to different administrative domains have significant latencies and thus cause significant delays that interrupt data transfer. Section 3.2 provides a detailed description of the handoff procedures and the cause of their associated latencies in order to provide a better understanding of how significant of an impact such handoffs would have on performance levels, which is addressed in Section 3.3. Section 3.4 defines the problem that motivated this research effort. Section 3.5 provided a brief overview of other research efforts, their approaches, and shortcomings to provide further motivation

for the proposed solution described in Chapter 4. Finally, Section 3.6 provides a chapter summary.

3.2 Handoff Procedures

The concept of using smaller inter-networked cells to provide wireless connectivity has the advantage of supporting lower transmission power and providing higher aggregated bandwidth. It has also provided a more efficient method for locating a mobile node when receiving incoming calls [56]. However, this approach inevitably results in higher handoff rates that vary depending on the size of the network cells and the type of user mobility (i.e., pedestrian or vehicular).

In the past, handoff was limited to base stations that supported the same technology and in many cases belonged to the same administrative domain. However, the current trend of internetworking cells that support different technologies introduces a more complex handoff process as addressed throughout this chapter. As described in Section 2.5.2, there are several types of handoffs and they each have a different level of complexity. The handoff complexity is usually the most significant contributor to the delay/latency associated with the handoff process.

Seamless mobility has been the source for a large number of research efforts and is now being studied from a different angle to accommodate mobility across heterogeneous networks. The following is the main list of requirements for achieving seamless mobility, as addressed in the majority of the literature [56][58]:

- Automatic handoff from one base station to another without any user triggering
- Transparent network transitions via a common underlying protocol (e.g., IP)
- Application session persistence during handoff procedures and temporary lower layer disconnections
- Minimal handoff-related delays and packet loss

Seamless mobility is heavily dependent on the ability to perform seamless handoff during the switching from one cell to another. A seamless handoff is one that performs a "fast handoff" and a "smooth handoff" [56]. A fast handoff involves no noticeable delays/latencies, whereas a smooth handoff involves no packet loss [56]. Many algorithms and approaches have been proposed in an effort to reduce handoff

latency and packet loss to lower the effect of mobility on performance [59][60]. However, despite the progress made in enhancing the handoff process in many technologies, seamless handoff still presents a challenge in wireless networks. Therefore, seamless vertical handoff between heterogeneous technologies presents an even bigger challenge, due to its higher complexity.

According to the 3GPP standard and other publications that have addressed the handoff process between 3G/UMTS and WLAN, certain procedures are necessary in order to support mobility across these heterogeneous networks, which typically belong to different administrative domains. Although a number of other procedures and approaches have been suggested by other research efforts [26], Mobile IP [25] and AAA [6] have been consistently included in the literature as the primary protocols for enabling mobility between 3G/UMTS and WLAN [6][25][39]. These protocols have latencies that contribute to the aggregated delay associated with a handoff from one network to another. Furthermore, according to a series of research efforts, other unavoidable latencies also result from handoff and contribute to the overall handoff latency as shown in Figure 3-1.

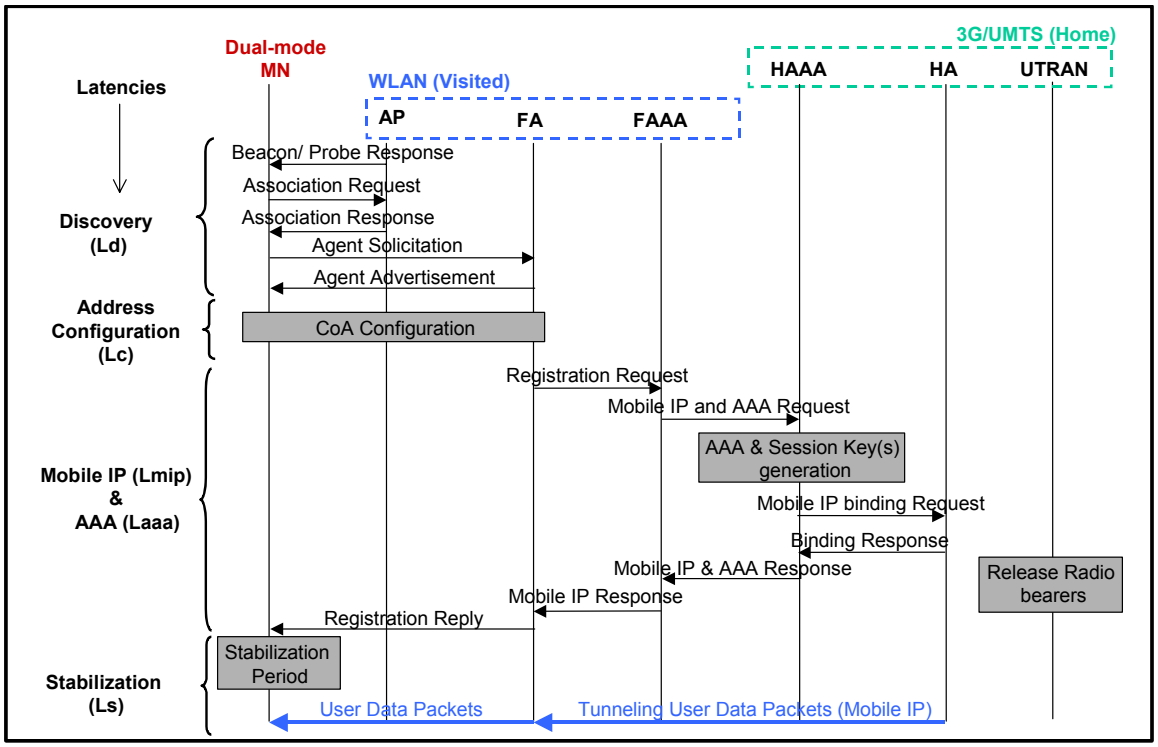

Figure 3-1. The downward vertical handoff procedure from 3G/UMTS to WLAN

3.2.1 Downward Vertical Handoff

Figure 3-1 shows the different sub-procedures that comprise the downward vertical handoff vertical handoff from a home 3G/UMTS to a visited WLAN 802.11 and the latency distribution among these procedures [2][6][26][31][57]. The downward handoff was shown in this diagram because the handoff from 3G/UMTS to WLAN (i.e., a downward vertical handoff) is the focus of this research effort. However, a user that belongs to a home 3G/UMTS network and is currently connected to a visited WLAN will utilize the same procedures when performing a handoff from one WLAN to another since the user's home 3G/UMTS network will need to be contacted to re-route traffic to the new WLAN. Therefore, such layer-3 (L3) horizontal handoff was also addressed in this research effort. The handoff process is divided into two main stages "Coverage Discovery," which is addressed in Section 3.2.1.1 and "Execution," which includes Address Configuration, AAA procedures, Mobile IP procedures, and the stabilization period needed to recover from the effect of handoff/mobility on the upper layer protocols and running applications.

3.2.1.1 Discovery

The current non-optimized method for detecting different types of wireless coverage in areas that have heterogeneous wireless overlay networks requires simultaneously active network adapters on the dual-mode MN. The IEEE 802.11 standard defines the method for detecting the presence of WLAN access points (APs) [2]. The MAC layer is responsible for triggering passive or (the optional) active scanning to discover WLAN coverage. Passive scanning consists of listening for beacons sent periodically by access points (APs) on specific channels, whereas active scanning consists of the MN sending probes on each channel and waiting for the AP(s) responses. The optional active scanning method offers a faster but less efficient approach—from a power and bandwidth consumption perspective.

When the WLAN adapter receives a beacon or probe response, the MAC layer checks its associated Received Signal Strength (RSS), then compares it with other received beacons' RSS values. The contents of the beacon are also checked to determine the supported data rates and other network-dependent information [2]. The MN then

sends an "association request" to the AP with the strongest beacon, which then authenticates the user and sends an "association response" to confirm the association. Upon receiving the response, the MN sends a router/agent solicitation to discover the access router (or foreign agent as described in Section 3.2.3) currently serving the AP. The router/agent returns an advertisement to the MN providing its relevant information.

Some literature distinguishes between "AP discovery" and "Router Discovery," whereas others combine the two and also include the Duplicate Address Detection (DAD) as part of the discovery [31][49]. In this research effort, the discovery latency was considered to be the time between activating the WLAN adapter and receiving the first router advertisement from the router associated with the selected/target AP. As for DAD, it was considered part of the "address configuration" stage addressed in Section 3.2.2. The discovery latency varies depending on the type of scanning, the scanning intervals, the beacon intervals, and the router advertisement interval as well as the frequency of these procedures.

Overlay networks have overlapping coverage areas that support different technologies. The "upper" network is typically larger and has lower data rates, whereas the "lower" network is smaller and offers higher data rates. Therefore, as stated in Section 2.5.2, the discovery step in the downward vertical handoff is not delay-sensitive since the MN is not at risk of exiting the large cell it is currently connected to (e.g., 3G/UMTS). Instead, the handoff is typically initiated for performance optimization reasons. On the other hand, upward vertical handoff takes place when the user exits the small coverage area (lower cell). Therefore, the discovery step in upward handoff is usually delay-sensitive and can cause disconnection and severe packet loss. This reaction would occur in cases where the user suddenly exits the coverage area, and at a high speed, thus not allowing the MN sufficient time to trigger the discovery stage to search for other 802.11 or UMTS coverage. Moreover, the network adapters' capabilities and configuration on the MN are also a contributing factor to the amount of delay associated with coverage discovery [63].

3.2.1.2 Address Configuration

When a MN visits a new (or previously visited) network, it usually needs to acquire a temporary Care-of IP address (CoA) in order to support seamless mobility, as described in the next subsection. There are a few alternatives for acquiring such CoA. For instance, in the context of Mobile IP as shown below, there is an option for using the router's IP address as the CoA without the need for allocating a unique IP address to the visiting node. In such cases, the latency could almost be negligible, but in some experiments was found to be as high as 1s [50].

According to the arguments made in the literature, other alternatives for acquiring a CoA are important to consider in terms of avoiding a bottleneck or in terms of supporting future protocols such as IPv6, which has a different addressing and mobility scheme [49]. The primary method for obtaining the so-called "co-located CoA" in a foreign network is based on the DHCP protocol [61]. According to the experiments conducted in [49] and [50], the delay associated with obtaining a CoA via DHCP can "involve random wait times that can lead to delays in the order of seconds, even when performing a handover between high speed wireless networks"[49].

In addition, when co-located addresses are allocated to users visiting in a foreign network, there is a risk that another user may have already manually assigned the same IP address to his/her MN's interface. Therefore, a Duplicate Address Detection (DAD) feature associated with DHCP was created to check for such situations, using methods such as ICMP Echo request/reply. The DHCP server checks for duplication prior to allocating an address to the visiting MN. It also recommends another check by the MN after it receives the CoA message (prior to finalizing the assignment to its interface) [49]. Therefore, these checks contribute to the latency associated with obtaining a CoA and were considered during the analysis in Chapters 4 & 5. Note that in some literature, the DAD-related delay was acknowledged but not distinguished as a separate latency value during experimentation. In such experiments it was either assumed to be part of the discovery latency or the address configuration latency for the sake of simplification [31].

3.2.1.3 AAA

The Authentication, Authorization, and Accounting (AAA) protocol is currently viewed as the most suitable solution for managing and exchanging subscriber information and credentials between different networks, particularly combined with Mobile IP. This protocol also enables the mobility and roaming across different networks, particularly ones that have different administrative domains. DIAMETER [23] is the next-generation of AAA and is currently being standardized by the IETF. DIAMETER has evolved from its predecessor RADIUS [42].

The basic AAA protocol, which is the basis for DIAMETER and RADIUS, is based on the communication between a Home AAA (HAAA) server and a Foreign AAA server (FAAA). The HAAA is responsible for authenticating and/or authorizing user access in a particular domain. The FAAA represents the authority in the foreign network and is responsible for negotiating client credentials with the HAAA. Furthermore, an intermediate node (proxy) between the FAAA and HAAA is referred to as the AAA Broker (BAAA). In many cases, a few proxies/agents are necessary to provide intermediate communication [6][35]. As expected, the higher the number of proxies that must be traversed, the longer the AAA latency.

3.2.1.4 Mobile IP

The Mobile IP protocol was initially proposed to provide a seamless solution to the mobility issue caused by the static characteristics of IP addressing [25]. Mobile IP was based on the IP protocol and was meant to mask the mobility from the upper layers of the protocol stack to offer seamless mobility. Its underlying concept consists of establishing a tunnel between the MN's home network and the current visited network. The tunnel has two end points: the Home Agent (HA) and the FA (FA). The HA is responsible for maintaining knowledge of the current location of the MN and forwarding incoming packets to it. The FA is responsible for communicating with the HA to ensure the delivery of packets to the MN while it is a visitor in its network.

The protocol's mechanisms are divided into three steps: agent discovery, registration, and tunneling. Agent discovery is necessary for notifying the MN of the existing FA serving the visited network and can be obtained through periodic agent advertisements or by solicitations from the MN. Registration consists of notifying the HA of the MN's new CoA in the visited network in order to bind its home IP address to its new CoA for forwarding packets. Finally, tunneling consists of the HA encapsulating packets received from the correspondent node (CN) at the user's home network, then forwarding them to its CoA, where the original packets would be extracted (Figure 3-2). Note that the CoA may either be the FA's IP address or may be an IP address specifically allocated to the MN when it entered the visited network.

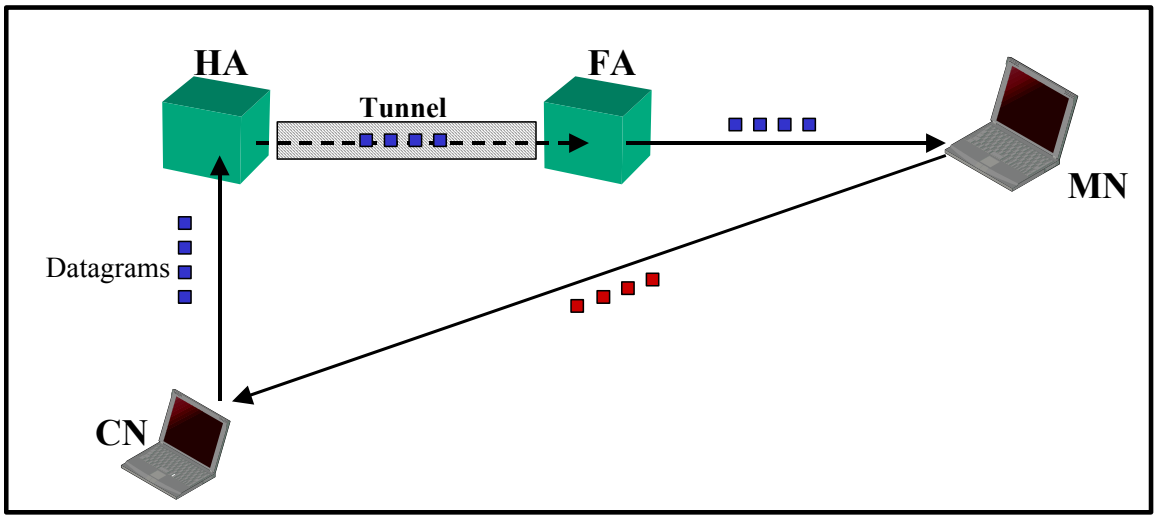

Figure 3-2. The Mobile IP tunneling mechanism, based on information from [25]

3.2.1.5 Stabilization

The stabilization period is a combination of the time required for the application(s) to recover after the handoff and adapt to the new data rate, as well as the residual TCP back-off time caused by the handoff [31][37][38]. TCP backoff is a mechanism that was originally introduced to handle packet loss due to network congestion in the wired network. However, such mechanism interprets the delays and packet losses, caused by handoff in wireless networks, as an indication of network congestion. Therefore, it reacts to such events in the same manner that it would network congestion.

Its reaction consists of first dropping the transmission window size to significantly reduce the amount of transmitted data to cope with the presumed congestion. It then activates what is known as the "slow-start" algorithm to slowly increase the size of the transmission window and allow the network to recover from the congestion. Finally, it resets the retransmission timer to an interval that would double with each timeout. Note that timeouts occur when acknowledgements, for sent data, have not been received within a specific period. In such cases, TCP would retransmit the packets believing that they had been lost [37][38].

Several research efforts investigated and addressed this issue and have quantified the impact of mobility and handoff on TCP and the resulting impact on throughput and delays during and after the IP-level handoff [37][38]. The primary contributor in this case was found to be the timeout (or consecutive timeouts). Recall that the IP-level (i.e., Layer-3) handoff procedures, which come after the coverage discovery step, cause a period of non-communication with the MN (i.e., no user data is received or transferred). Therefore, during this period the reliability mechanism in TCP causes it to perform the backoff procedure described in the previous paragraph. The longer the delay associated with the IP-level handoff procedures, the longer the timer intervals will become.

According to the experiments in [37] and [38], the long timeout could result in a pause that lasts from 0.8 seconds to a few seconds depending on the number of timeouts. Therefore, despite the completion of IP-level handoff, the data transmission will not resume until the timer expires and the first acknowledgement is received. At this point, the slow-start algorithm would also require some time to recover and allow the window size to return to its maximum value. The experiments in [37] indicate that the slow-start could take up to one second. However, for that duration the transmission is merely throttled, thus the impact of the slow-start on throughput is relatively moderate compared to timeout(s).

3.2.2 Upward Vertical Handoff

A MN currently attached to a WLAN AP monitors the current AP's beacons and evaluates the Signal to Noise Ratio (SNR) value to detect when a handoff is needed (i.e., when a user exits the current coverage area). Upon noticing degradation in the SNR, the

WLAN adapter starts scanning for beacons in search of another AP. After a certain period of scanning, if no other APs exist or have a strong beacon, the upward handoff to 3G/UMTS is triggered [63]. The MN then evaluates the RSS of the Broadcast Control Channel (BCCH) for the available 3G/UMTS cells and decides which cell to attach to accordingly. It then communicates with the nodes in the UTRAN to provide the appropriate nodes in the core network (CN) with its current cell position and requests that the Radio Access Bearers (RAB) be established. RAB are the access stratum for transferring user data [57] and are under the control of the RRC protocol, which is the protocol responsible for the L1 & L2 handoff management in UMTS, as described in Section 2.2.3.

The IP-level handoff steps (i.e., after the discovery step) are the same for upward and downward handoff, so long as the user is moving from a home network to a visited network. If the user is returning from a visited network to a home network, then the Mobile IP and AAA procedures would not be necessary and thus the overall latency is significantly lower, further details regarding the delay associated with upward handoff are provided in Chapter 5. Note that if the user is moving from one visited network to another, the latencies associated with Mobile IP and AAA can be longer due to the number of proxies/gateways that must be traversed to update the user's CoA and obtain the AAA credentials from the user's home network. Finally, the upper layers (i.e., transport and application) react in the same manner to the handoff-related delays, packet loss, and change in data rate, regardless of the handoff direction. Therefore, the stabilization stage's characteristics and impact are the same regardless of the direction of the handoff (upward or downward).

3.3 Aggregated Handoff Effect

The previous section provided a detailed description of the handoff stages and procedures for downward and upward vertical handoff in heterogeneous wireless overlay networks, particularly 3G/UMTS and WLAN (IEEE 802.11). The latencies of each stage of upward and downward handoff have been demonstrated and addressed by several research efforts to have a significant impact on delays and packet loss observed at the user level [31][37][38][62][63][64]. The aggregate delays and packet loss of an average

handoff cause interruptions in the data transfer during the handoff execution stage and the few seconds that follow. These interruptions result in throttling the application level throughput and increasing the application response time. However, such decrease in throughput would typically be mitigated by the higher data rates offered by the WLAN. Therefore, ultimately, the overall application response time would be significantly less, despite the handoff delay, which may have lasted a few seconds.

In order to benefit from the higher data rates, the user must remain within the WLAN coverage area for a period that exceeds that of the handoff latency. Note that this argument is limited to non-real-time and non-delay-sensitive applications since such applications would suffer from handoff-related performance degradation regardless of the time spent in WLAN and the data rates offered.

3.4 Problem Description

The number of wireless data services and users is rapidly increasing, thus prompting an increase in the number of WLAN coverage areas to support customer demand. Furthermore, the internetworking between 3G and WLAN overlay networks will also prompt the deployment of more WLAN hotspots in areas that have a high population density. This requires smaller cell in order to maintain an acceptable level of QoS, from a bandwidth perspective. Moreover, WLAN hotspots are being deployed in areas characterized by relatively high user mobility such as airports, subway stations, convention centers, outdoor urban areas, parks, and university campuses. As a result of all these factors, the rate of vertical handoff is expected to be significantly higher in such areas. Furthermore, the probability and rate of short WLAN visit durations is also expected to increase.

As a user enters and enter a WLAN coverage area a downward handoff would be triggered followed by an upward handoff as the user exits the coverage area, both triggered based on RF measurements (i.e., the conventional method). The delay associated with the downward than upward handoff results in a significant period of delay and packet loss. In essence, this situation would have a similar impact to that of the "ping-pong effect," which results from a MN traveling on the border of two or more cells thus causing the triggering of a series of unnecessary consecutive handoffs between such

cells. The negative impact of the "ping-pong effect" is demonstrated and addressed in [63].

The negative impact of the consecutive downward than upward handoff, without sufficient time to transfer data over the new WLAN connection, would result in data interruptions that throttle the application level throughput and reduce the application response time. Therefore, it would counter the purpose for which the handoff from 3G/UMTS to WLAN was triggered in the first place, which is to enhance performance. This represents a problem that has not yet been observed in practice, but that is expected to materialize as the internetworking of heterogeneous, wireless overlay, networks takes place and the number of small WLANs increases in areas of relatively high user density and mobility.

3.5 Potential Solutions

In light of the argument presented in the previous section, a mechanism for predicting the user's visit duration to a network and deciding on the handoff accordingly, would prevent the occurrence of inefficient/unbeneficial handoffs, thus preventing the degradation in application-level throughput and the wasting of resources. This applies primarily to downward handoff due to the upper cells (e.g. 3G/UMTS) being significantly larger than the lower cells, which means that the user's visit duration in the lower networks (e.g., WLAN) would likely be shorter. Other research efforts have called attention to this issue and noted its similarity to the "ping-pong effect" in homogeneous networks [62][63].

"Prediction" has also been the focus of research efforts attempting to achieve a more seamless handoff in homogeneous and heterogeneous networks, particularly based on a combination of the signal strength and the user's position [33][36][46][47]. However, the majority of the proposed approaches have relied on signal strength without sufficient assessment of the user's speed and direction. Furthermore, the research efforts that have considered the user's velocity have done so with regards to the distance between the MN and the AP only. Therefore, their estimates did not take into consideration the obstacles that could prevent the AP from having a full/continuous range (e.g., walls). They also assessed coverage in an AP-centric manner rather than treating

the hotspot coverage as a single entity. Treating multi-AP coverage area as a blanket of coverage would allow the prediction-based algorithm to avoid triggering when only a Layer-2 handoff is needed, due to its minimal associated latency.

Therefore, this research effort proposed augmenting the current conventional handoff decision method with a location-based evaluation that takes into consideration the user's actual position, speed, and direction as well as the spatial attributes of the discovered WLAN in an aggregated manner (blanket effect).

The main focus of this research effort was the downward handoff from 3G/UMTS to WLAN and the layer-3 horizontal handoff, of a 3G/UMTS user, between two foreign WLANs. While the handoff decision mechanism for triggering upward handoff could benefit from location-based information to identity the best time to trigger such handoff, this research effort did not support/modify it. This is due to the upward handoff's sensitivity to delay, which requires a mechanism that can make a quick decision to avoid lengthy disconnections, which could cause (the more severe) upper layer disconnections. Therefore, according to the literature, an approach based on RF measurements is more suitable; examples of such approach include the threshold approach, hysteresis approach, dwell timer approach, or a hybrid of these approaches [63].

3.6 Summary

This chapter provided a detailed description of the stages that comprise downward and upward handoff, particularly for 3G/UMTS and WLAN overlay networks. This description and qualitative analysis was meant to address the source of the handoff-related delays in order to show the potential delays associated with vertical handoffs and their impact. The problem that motivated the research effort was defined in Section 3.4 based on the understanding of the impact of handoff, which emphasizes the need for a better handoff decision mechanism. Finally, potential solutions found in the literature and their weaknesses were briefly addressed to bring to the reader's attention the weaknesses that had to be addressed in the proposed solution, described in Chapter 4.

Chapter 4: Design Description

4.1 Introduction

The benefits of integrating UMTS and WLAN networks were addressed in Chapter 2 along with the issues of integrating such heterogeneous technologies. These issues range from problems with internetworking the infrastructure, organizing the billing, ensuring security, and finally enabling seamless mobility. This research effort was focused on the improvement of mobility and the handoff process between heterogeneous wireless-overlay networks (3G/UMTS and WLAN). Recall from Chapter 3 that the handoff process between these hybrid networks involves a cost that is incurred by the network and observed at the user level (i.e., degradation in application response time). This cost would typically be overlooked due to the benefit gained from connecting to a network with a significantly higher data rate. However, as described in Chapter 3, this handoff cost would only be mitigated if the user remains in the lower network (i.e., WLAN) long enough to benefit from the higher data rate. However, a short visit to WLAN would resemble the negative impact caused by the ping-pong effect and defeat the purpose of the handoff [62]. The need for a location-aided approach was suggested in Chapter 3 to make more location-aware handoff decisions that result in the intended performance enhancement rather than degradation.

This research effort assumed an environment where the user's home 3G/UMTS and visited WLAN networks were loosely coupled—inter-networked without the need for the WLAN access network to have an Iu interface (see Figure 2-3) [65]. According to the literature, loose coupling is the primary internetworking approach due to its flexibility for allowing a network operator to offer 3G-WLAN subscriptions to the end user regardless of whether both networks are owned by the same operator. It was also assumed that the MN was equipped with the two corresponding adapters and was capable of automatically activating the WLAN adapter and triggering the WLAN discovery process when the user activated a non-real-time application, and triggering the UMTS discovery process when the user exits 802.11 coverage.

This chapter describes the location-aided handoff decision algorithm that was developed to handle the problem described in Chapter 3. Section 4.2 provides an overview of the basic conventional handoff decision for handoff between heterogeneous wireless-overlay networks. Section 4.3 describes the location-aided algorithm, which combined the RF-measurements with a location-based evaluation to evaluate the suitability of a handoff and make better handoff decisions than the pure RF-based conventional method. An important component of the proposed design, the WLAN coverage database, is addressed in Section 4.4. The algorithms specifications with respect to the assumptions, requirements, underlying mechanism, and calculations are addressed in Section 4.5. Section 4.6 provides a description of the proposed architecture for supporting the algorithm in the UMTS network. Section 4.7 provides a brief description of the simple communication protocol required to support the algorithm's functionality and the associated exchange of control messages between the MN and the UTRAN. Finally, Section 4.8 concludes the chapter with a summary of the proposed approach.

4.2 Conventional Approach (Discovery & Handoff)

The conventional method for discovering coverage and making the decision to trigger an upward or downward vertical handoff, as described in the literature, is shown in Figure 4-1. It has been based primarily on RF measurements [63]. The basic approach consists of a dual-mode MN currently connected to an upper network (e.g., 3G/UMTS, GPRS, cdma2000, etc.) and scanning for coverage using its other network adapter (e.g., 802.11b). Upon discovering WLAN coverage through beacons sent by Access Points (APs) and receiving a router advertisement from the associated router, a downward vertical handoff is triggered.

During the MN's visit to a WLAN network, it remains connected to the current AP until the RF measurement collected by the adapter indicate a degradation in RF parameters such as Signal to Noise Ration (SNR). The MAC layer triggers passive or active scanning to locate other APs. If another AP within the same hotspot is found, a L2 handoff takes place without notifying the home network. However, if the AP belongs to another hotspot, then a L3 handoff is required to maintain seamless mobility. A

horizontal Layer 3 handoff to a visited WLAN network includes notifying the MN's home UMTS network of the user's new location.

On the other hand, if the discovery stage fails to locate other APs with strong beacons, an upward handoff is triggered when the SNR value drops below the specified threshold. Upon such trigger, the UMTS adapter uses RF measurements to discover and evaluate UMTS cells then triggers the L3 handoff once the discovery stage successfully completes. According to the relevant literature, there have been several schemes for evaluating RF measurements to determine the best time to trigger a handoff in homogeneous and heterogeneous networks. Some of these schemes involve the use of a threshold, hysteresis, and/or dwell timer mechanism to reduce the errors that cause unnecessary handoffs (i.e., ping-pong effect) [63]. As mentioned in the previous chapter, the following design does not contribute nor hinder the upward vertical handoffs due to their sensitivity to delays and their need for a fast decision algorithm based on PHY layer measurements.

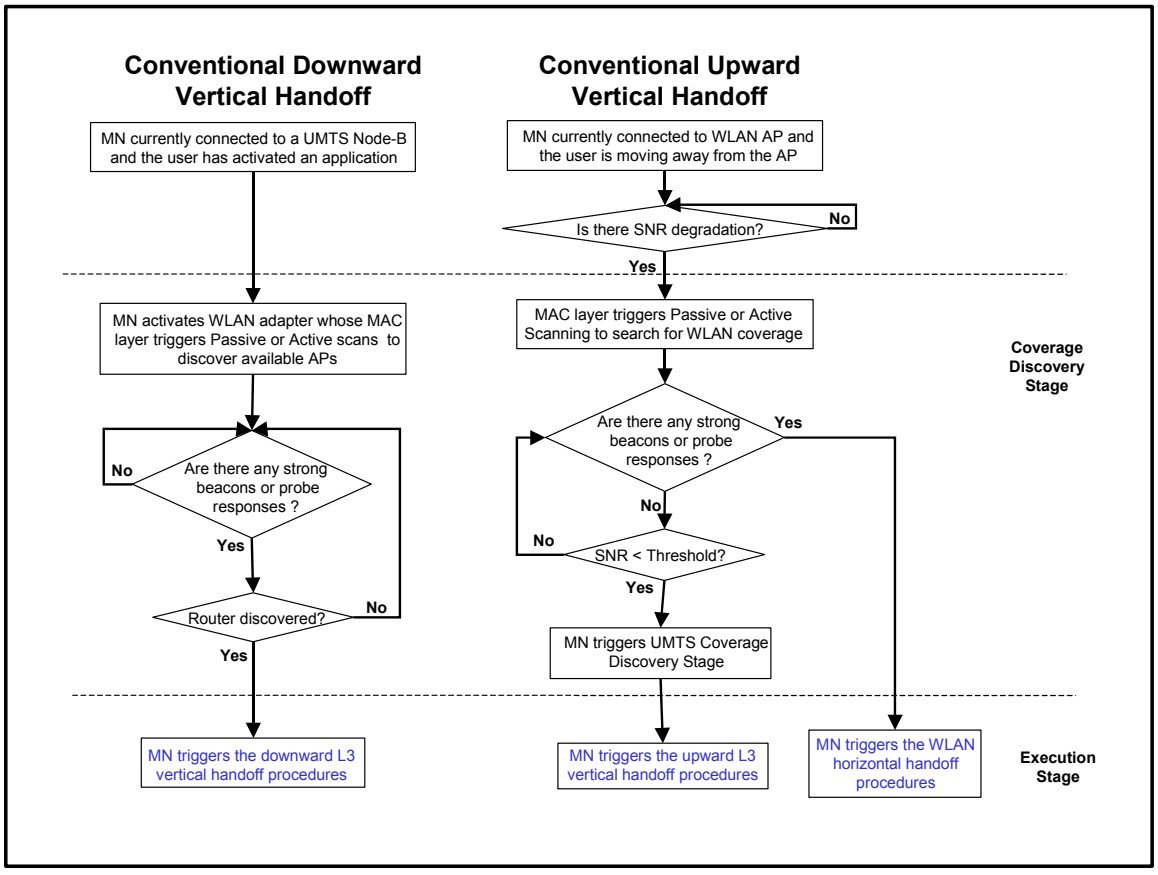

Figure 4-1. Upward and downward vertical handoff between UMTS and WLAN according to the conventional pure RF-based approach

4.3 The Proposed Location-aided Handoff Decision Algorithm

The main idea behind the proposed algorithm is to incorporate the user's position and velocity information into the handoff decision algorithm by predicting whether the user will be in the WLAN coverage area for a period longer than the required visit duration to justify the cost of the handoff. An algorithm was developed to coordinate the use of location-based information with the mechanisms of the conventional handoff procedures that decide on whether a L3 handoff should take place. This section provides a description of this algorithm for the downward handoff. It also provides the algorithm developed for coordinating the use of location-based information in horizontal L3 handoff to visited WLAN networks. The proposed approach takes into consideration the user's position, speed, and direction. It utilizes these parameters, in combination with other parameters, to determine the minimum required visit duration to WLAN coverage. This minimum time period is the period necessary for the handoff procedures to complete and for the MN to receive a specified number of user-data packets. The process for discovering WLAN coverage can also be optimized using location-based information to determine the distance between the user and the nearest WLAN hotspot. However, scanning or probing for beacons is assumed to be the mechanism for discovering WLAN coverage in the following proposed algorithm.

4.3.1 Downward Vertical Handoff Decision

Figure 4-2 provides the basic outline of the proposed algorithm for incorporating the location-based information into the conventional handoff model. The coverage-discovery stage would be entirely performed by the WLAN adapter at the MN. However, the proposed approach included a trigger that notifies the UMTS network when a strong beacon is received from an AP. Recall that in the conventional approach, the discovery of available WLAN coverage included receiving a strong beacon from an AP. Upon receiving the strong beacon, the MN would trigger the L2 and L3 handoff procedures to the target WLAN (Figure 4-1). In the context of the proposed approach, only the router/agent discovery procedure would take place and instead, a notification would be sent to the UTRAN to trigger the location-based evaluation mechanism. Only if the location-based decision recommends the handoff do the L3 procedures proceed. In order

to support such functionality at the MN, the "Location-based Evaluation and Decision Stage" was inserted into the conventional handoff paradigm at the MN (Figure 4-2). The following is a list of the proposed steps (and their descriptions) that comprise the mechanism for obtaining and utilizing the user's position and velocity and the hotspot's coverage footprint to advise the MN with regards to whether a handoff would be suitable at a given time. Note that the numbers associated with each step are merely for the purpose of labeling, not to indicate the order of occurrence.

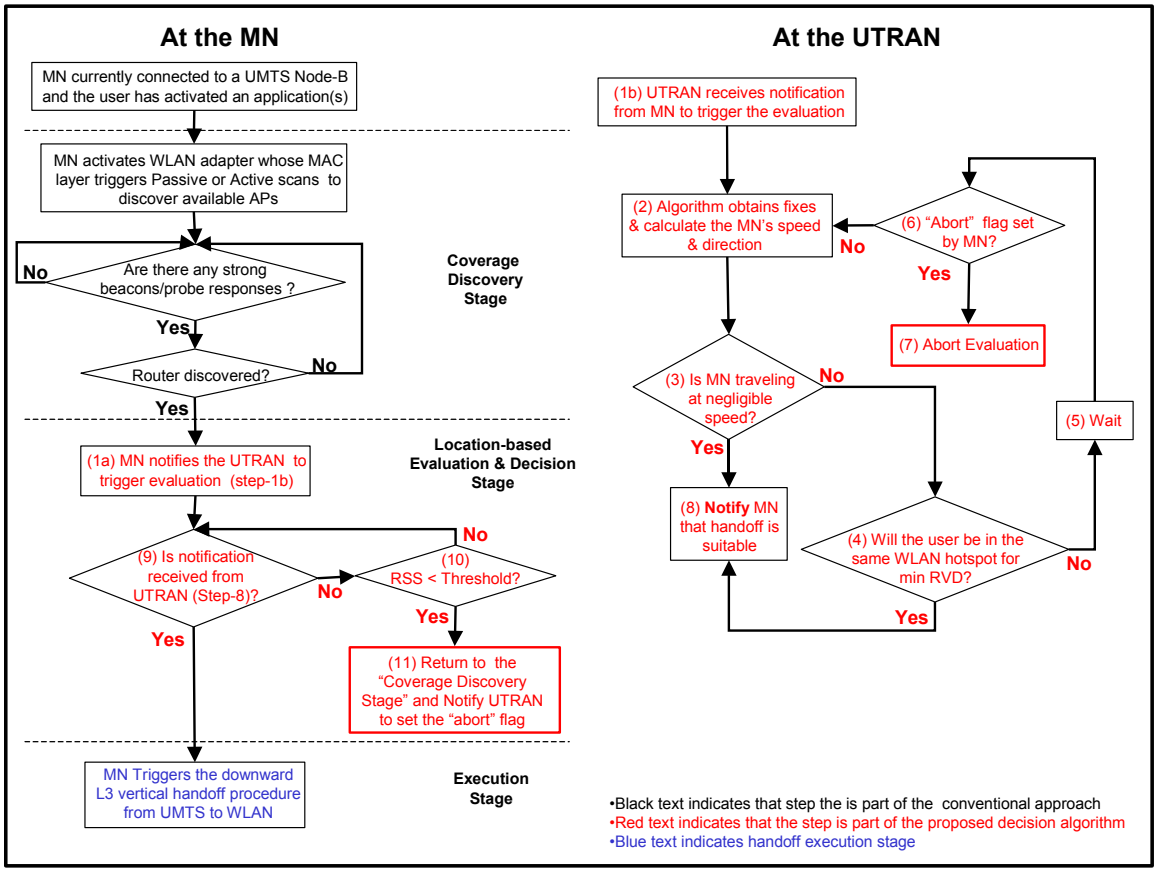

Figure 4-2. The steps of the proposed location-aided decision algorithm when the user is connected to the 3G/UMTS network and entering a WLAN coverage area

(1a) Upon receiving a router advertisement, the MN sends a notification to the UTRAN to trigger the location-based evaluation, which begins with Step (1b)

(1b) When the UTRAN receives the notification from the MN the appropriate node in the UTRAN is notified to trigger the location-based evaluation. The MN's notification to the UTRAN would include the hotspot's registration code (a parameter specifically

introduced to support the algorithm's mechanisms and maybe incorporated into the beacon information)

(2) The algorithm notifies the Location Measurement Unit (LMU) shown in Figure 2-10 [13] to obtain fixes for the user's location and calculate the MN's speed and direction of travel. Any positioning tool with a high degree of accuracy and reasonable response time can provide such fixes. However, a network-based approach (e.g., U-TDOA [43][44] described in Section 2.6.4.2) may be used to minimize control signaling and acquisition/response time.

(3) A check is conducted to determine if the MN is moving at a negligible speed, which would renders the evaluation of direction negligible a well. If so, the algorithm advises the MN that a handoff is suitable (i.e., Step (8)), otherwise Step (4) is invoked. The underlying purpose of this step is to allow the MN to:

a. Take advantage of WLAN coverage as soon as possible and take advantage of the fact that the user is not moving (i.e., the probability of existing WLAN coverage is low)
b. Avoid the waste of resources involved in the unnecessary execution of the later algorithm procedures.

(4) The minimum Required Visit Duration (RVD) is the time period that a user must remain within the same WLAN coverage area to ensure the successful completion of the handoff procedure and the transfer of a sufficient (configured) amount of data over the WLAN network. In other words, it is the amount of time that the user must remain within the same WLAN to allow the application to benefit from the higher data rates and compensate for the handoff-related delays. The underlying calculations for obtaining the minimum RVD are provided in Section 4.5.2. According to the user's speed and direction, the algorithm intelligently predicts whether the user will remain within the same WLAN coverage area for the minimum RVD. The details of the prediction mechanism are addressed in Section 4.5. If the algorithm predicts that the user will

remain in the same WLAN coverage area for the minimum RVD, Step (8) would be invoked otherwise Step (5) would be invoked.

(5) If Step (4) predicted that the user would not remain in the same WLAN coverage area for the minimum RVD, a timer is triggered to allow for a "wait" period between location-based evaluation iteration/cycles. This step helps reduce the number of unnecessary iterations and wasting resources. It also increases the probability of obtaining different positioning values for the MN (if the user was moving).

(6) After the "wait" period, a flag is checked to see if the MN had notified the UTRAN that it no longer received a strong beacon (i.e., it exited the coverage area). This "abort" flag was necessary to prevent the algorithm from iterating endlessly despite the user moving away from the WLAN coverage area. If a flag was set, then Step (7) would be invoked, otherwise Step (2) would be invoked to start a new location-based evaluation.

(7) If Step (6) confirmed that the "abort" flag was set, the location-based evaluation running at the UTRAN terminates.

(8) If the answer to queries (3) or (4) was a "yes," then a notification would be sent to the MN. This notification advises the MN to trigger the handoff "Execution Stage."

(9) The MN waits for a notification from the location-based evaluation at the UTRAN. If one is received then the handoff "Execution Stage" is triggered, which consists of the L3 handoff procedures described in Chapter 3 and shown in Figure 3-1. Otherwise, if a notification is not received, step (10) is invoked.

(10) This step represents the MN monitoring the RSS beacon to check if it had degraded below a certain threshold, thus indicating the user moving out of the coverage area discovered. If so, Step (11) is invoked, otherwise, the algorithm returns to Step (9).

(11) This step is invoked when the WLAN adapter notices degradation in the RSS value of the received beacon prior to receiving the algorithm's notification to trigger the handoff "Execution stage." Upon noticing the degradation in RSS, the WLAN "Coverage Stage" is invoked again to search for other WLAN coverage, and a notification is sent to the UTRAN to set the "abort" flag described earlier. This causes the location-based evaluation to be terminated in order to avoid infinite iterations for evaluating the user's position with regards to a WLAN, which he/she are no longer located at (i.e., avoids an infinite loop at the UTRAN).

The algorithm's iterations at the UTRAN will continue until one of the following events takes place:

- The user is no longer traveling at a noticeable speed
- The user is predicted to remain in the same WLAN hotspot for the minimum RVD value
- The user exits WLAN coverage and the MN notifies the UTRAN to abort the algorithm

Note that the algorithm at the UTRAN will only send a notification to the MN if the decision is positive (i.e., the outcome of Step (3) or (4) is a "yes"). Therefore, if the coverage discovery stage successfully completes before the MN receives a notification from the UTRAN, then either a prediction has not been reached or the user's predicted visit duration was found to be less than the minimum RVD (i.e., a new evaluation iteration begin after the "wait" timer expires).

4.3.2 WLAN Horizontal Handoff Decision

According to the IEEE 802.11 standard and the information found in the literature [2][63], a horizontal handoff from one AP to another or one WLAN to another is triggered when the MN notices degradation in the current SNR. The discovery stage, as shown in Figure 4-1, determines the availability of other APs and whether they belong to the same WLAN hotspot or a different hotspot. Since horizontal L2 handoff has very minimal latency and does not require contacting the user's home network, the proposed design does not trigger the location-based evaluation in such cases. However, if the

target AP belongs to a different WLAN, L3 procedures are required (as described in Chapter 3).

As stated earlier, the latency of a horizontal L3 handoff to a visited network could exceed that of a downward handoff from a home network to a visited network. Therefore, the proposed location-aided algorithm would be triggered to predict the user's visit duration to the new WLAN hotspot and assess the suitability of such handoff accordingly. As shown in Figure 4-3, if the target AP does not belong to the user's current hotspot, the MN notifies the UTRAN to trigger the same location-based evaluation described earlier in Section 4.3.1. Steps 1-11 are the as those described for the downward vertical handoff. However, in this case, while the MN waits for the notification from the UTRAN with regards to the decision, it must monitor the level of SNR. If the SNR drops below the configured threshold value, Step (11) would be invoked. However, in this case, the UMTS "Coverage Discovery Stage" is also invoked to avoid the MN being out of coverage for a long time, which could cause the more

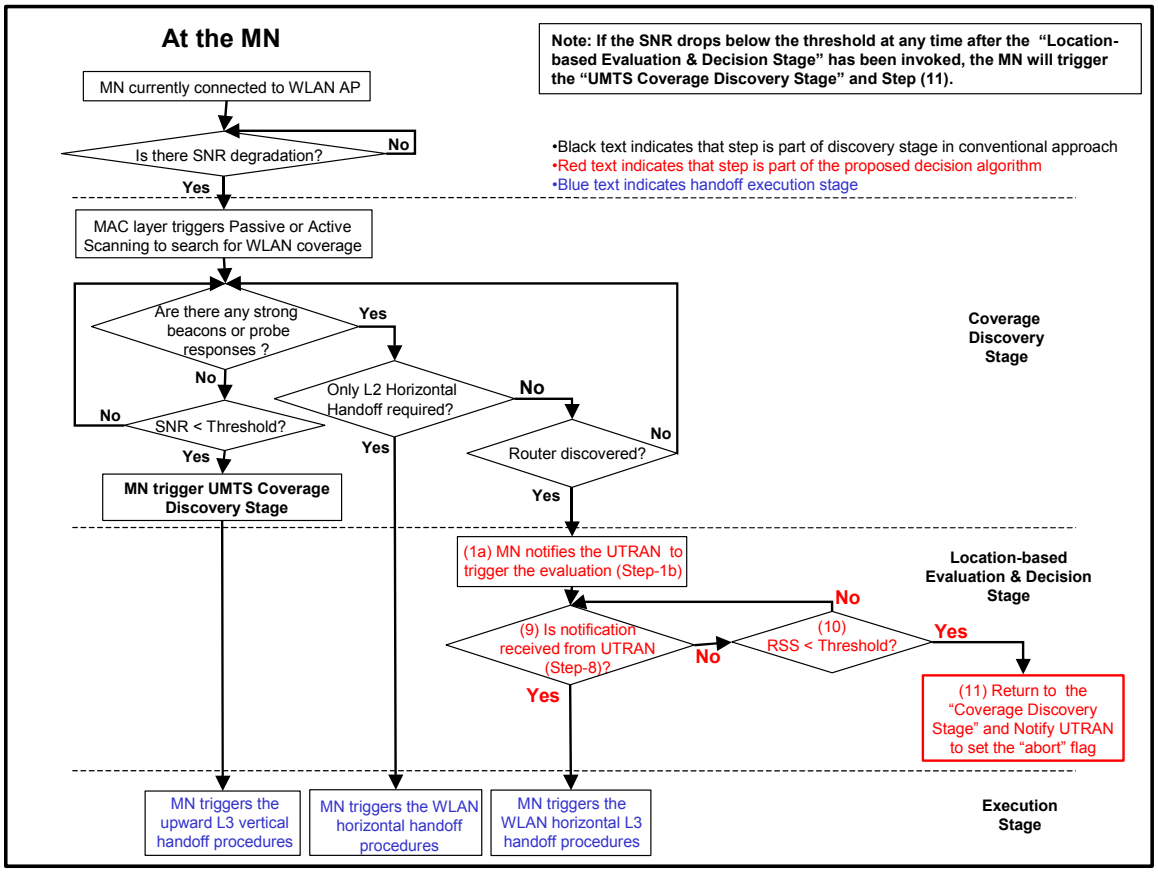

Figure 4-3. The steps of the proposed location-aided decision algorithm when the user is connected to a visited WLAN network in a heterogeneous wireless overlay environment

severe upper-layer disconnection (i.e., TCP and Application). This is due to the fact in the case of L3-Horizontal handoff situations, the user may move out of the current coverage area before receiving a decision from the location-based evaluation running at the UTRAN. Therefore, when the SNR drops below the threshold, an upward handoff is triggered to reestablish the connection with the 3G/UMTS network, since it is less risky than blindly triggering the L3 horizontal handoff (since it may result in only a brief visit and causes performs degradation). After the MN reestablishes its connection with 3G/UMTS, if the RSS of the discovered WLAN coverage continues to be strong, the "Location-based Evaluation Decision Stage" would be triggered again. However, in this case, the MN continues to transfer data over the 3G/UMTS connection (without the risk of going out of range) as it waits for the handoff decision to be sent in a notification from the UTRAN as described in Step (8) and Step (9) in Section 4.3.1.

4.3.3 Upward Vertical Handoff Decision

As addressed in Chapter 3, and for the reasons stated previously, the vertical handoff decision was found to be best served by a decision algorithm based on PHY layer measurements (Figure 4-1). Therefore, the proposed algorithm did not contribute nor hinder the conventional RF-based upward handoff decision mechanism.

4.4 WLAN Coverage Database

4.4.1 Database Structure

An essential component of the proposed approach is the WLAN coverage information in the form of a footprint representation encoded into a database. The location-based evaluation implemented at the UTRAN queries the database for spatial information regarding the target hotspots that a MN is trying to handoff to. An individual database for each RNS or for a small group of RNC is preferred in order to reduce the query response time. However, a distributed database can also be used to minimize the cost of implementing the proposed approach.

Refer to Section 2.6.3 for the UTRAN architecture proposed by 3GPP to support location-based services. Each cell in the UTRAN has a cell ID that the MN receives through beacons from the cell's Node-B. Therefore, when the cell ID received at the MN

does not match the stored ID, the MN sends an update to the network of its new location. Therefore, the network will always be aware of the user's cell location when the user is active. This cell ID can be used to narrow down the query and minimize the associated lookup response time.

In addition, for the purposes of the proposed approach, a hotspot (comprised of one or more APs' coverage areas) should be assigned a hotspot ID that distinguishes it from other hotspots within a UMTS cell. This would further narrow down the database search space and reduce the associated delay. However, more importantly, the concept of the hotspot ID was necessary for the correct functionality of the algorithm. Recall from Section 4.3 that Step (4) of the location-based evaluation at the UTRAN attempts to predict whether the user will remain within *the same* WLAN coverage area for at least the minimum RVD. Therefore, a tag (e.g., hotspot ID) was needed to differentiate between WLAN hotspots. This tag could be embedded into the AP beacon or probe response frame to be received by the MN, who would forward such value to the UMTS network to be included in the Location-based evaluation.

As for the overall querying, updating, or managing of the database, the Structured Query Language (SQL) was the recommended method for spatial databases as described in [66]. The actual queries may use any coordinate system for an index, so long as the indexing format is compatible with or derivable from the coordinates of the user's positioning tool (e.g., U-TDOA, GPS, etc.). The database structure is expected to be very simple since it merely consists of spatial records/entries stored at a certain resolution and encoded with a hotspot ID.

The concept of using a geo-location database to support and improve the handoff process between heterogeneous wireless overlay networks has been investigated by other research efforts [33][68]. The effort demonstrated in [33] described the use of a secure distributed geo-location database as well as fuzzy logic to determine when to trigger an L1 and L2 downward handoff (i.e., for discovery purposes) [33]. The database was queried to estimate (using fuzzy logic) whether the user was near or far from the WLAN coverage area in order to determine when the MN should activate the WLAN adapter to prepare for a downward handoff.

4.4.2 Footprint Accuracy

Ideally, the WLAN coverage representation or footprint should have high accuracy (e.g., within 10m). However, less accurate coverage footprints will not be catastrophic to the performance of the proposed algorithm. The information from a site-survey would be quite useful in enhancing the accuracy of the coverage representation. Furthermore, some of the more recently released software packages such as Ekahau[5]'s Site Survey™ 2.0 software or the like, could generate the necessary RF coverage map information in binary form. This type of information could potentially be ported directly into the WLAN coverage database to reduce the data entry efforts associated with encoding the spatial information. Figure 4-4 shows a snapshot of the user interface of Ekahau's Site Survey software.

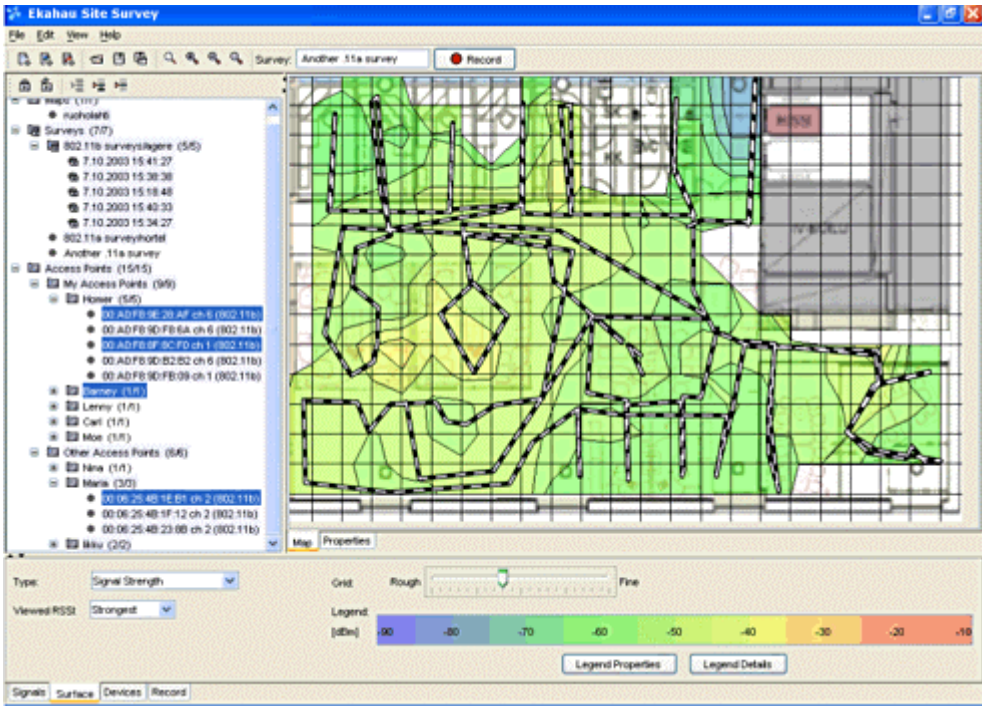

Figure 4-4. A snapshot of Ekahau's Site Survey™ 2.0 software interface showing the boundaries of a coverage area along with the signal strength obtained at each position [70].

[5] Ekahau's Site Survey™ 2.0 provides a map visualization service that records RF-data accurately then demonstrates such data in a visual format for analysis. RF-data includes SNR, RSS, Interference, etc. (www.ekahau.com)

4.4.3 Granularity

Although a highly granular coverage map would produce the most accurate result, a trade off between that and the query response time (which depends on the size of the database) is necessary in the context of the proposed approach. The proposed algorithm requires that a location-based decision be reached in real-time, thus it is time sensitive. This issue was addressed in the literature [68] by researchers whose effort included the use of a WLAN coverage database that provided coverage information at a 1.5m resolution, for the purposes of triggering location-based services depending on the user's location within a hotspot of considerable size (i.e., university building, airport, etc).

4.5 Algorithm Specifications & Calculations

The following are the calculations involved in obtaining the parameters necessary for estimating the user's trajectory, which would then be used to predict whether a user would remain in WLAN coverage for at least the RVD. If so, a notification would be sent to the MN to indicate the suitability of the L3 downward or horizontal handoff to a particular WLAN hotspot.

4.5.1 Assumptions

The following are the assumptions made during the design of the algorithm and it's underlying mechanisms and computations:

- A dual-mode MN equipped with both WLAN and UMTS adapters and the capability to exchange control messages with the UTRAN include the periods during which the MN is connected through WLAN.
- MN is active (i.e., user is running at least one non-real-time application)
- Positioning method provides positioning with a high accuracy (e.g., within 3-5m), which according to the literature is achievable through A-GPS based on the concept of Differential GPS (DGPS) or through some network-based positioning methods such as the recently standardized U-TDOA, which were addressed in Section 2.6.4 [46][54][55][67][68].
- The availability of a WLAN coverage footprint database [33][68], which was addressed in Section 4.4, to support the location-based evaluation proposed in

Section 4.3. The database would be queried for information regarding the user's visit duration to the target WLAN hotspot whose AP beacon(s) caused the MN to trigger the Location-based evaluation at the UTRAN (i.e., Step (1)). Ideally, the hotspot footprint/representation should have high accuracy (e.g., within 10m) to achieve the best performance. However, as mentioned previously, a less accurate footprint would not be catastrophic to the performance of the algorithm. The assumed resolution was 10m since the number of hotspots in a given a typical cell is reasonably small. Refer to Section 4.4 for further information regarding accuracy, resolution, and response time.

4.5.2 Required Visit Duration

This subsection describes the parameters that were taken into consideration when determining the minimum required length of time (i.e., visit duration) that a user must be in WLAN coverage in order for a handoff to be beneficial. Recall from Chapter 3 that a vertical or L3 horizontal handoff to a visited/foreign WLAN hotspot will involve some cost, particularly in terms throughput degradation and decrease in application response time. However, a handoff to a network with a higher data rate neutralizes the handoff cost by bringing significantly more benefit (i.e., higher throughput and lower application response time, etc). This is under the condition that the user must remain in the new network for a period long enough to allow for the following events to take place:

- The location-based evaluation to successfully complete and its decision notification to arrive at the MN. Based on the current state of technology, the algorithm's latency is heavily dependent on the latency of position acquisition delay, which depends on the method of positioning. However, in the observable future, the number of location-based services that require user tracking (i.e., periodic updates of the user's position) is expected to increase, as previously noted in Chapter 2. Therefore, the acquisition time, which is the primary contributor to the algorithm's latency, will no longer be an issue. Until then, the algorithm is expected to have a latency value that must be taken into consideration when calculating the minimum RVD value.
- The Address Configuration Stage to successfully complete (i.e., DHCP & DAD)

- The Mobile IP and AAA procedures to successfully complete
- The stabilization period to be over
- A sufficient amount of data to be transferred over the WLAN link to compensate for the increase in application response time due to handoff latency

The latencies associated with most of these events are network dependent. For instance, the discovery latency is dependent on the intervals between beacons and the frequency of router/agent advertisements. Furthermore, the registration latency, which encompasses both the Mobile IP and the AAA latencies, is a function of the network's MAC layer characteristics for the specific inter-networked UMTS-WLAN pair [31]. Moreover, a single cycle of the location-based evaluation has a latency that would depend primarily on the position acquisition time and the database query latency or response time, which are also network dependent assuming a network-based positioning method such as U-TDOA [44]. Furthermore, the amount of data that can be transferred after the stabilization stage depends on the throughput of both the current and target network. In addition to these network-dependent variation factors, which have the most impact on latency, there are some MN-dependent or adapter-dependent factors that may vary the latencies. An example of that is the address configuration latency, which depends on the MN's processing capability and the "state of the interface" [31].

In light of the argument made in the previous paragraph, it is recommended that the latency values be calculated periodically for each 3G/UMTS-WLAN pair. The throughput of the current and target networks is an important factor in calculating the RVD as described below. Therefore, the UMTS network should periodically probe the WLAN hotspots that are inter-networked with it, for the values of their current average throughput. The throughput in any network can vary depending on the network load, which varies based on the user density and the type of applications currently activated by the users. Therefore, periodic updates are necessary in order to maintain a reasonably accurate estimate of the current throughput in the WLAN networks within a UMTS cell.

Table 4-1 provides the list of acronyms and a brief description of their source. For a more detailed description of handoff procedures that cause these latencies, refer to

Chapter 3 and Figure 3-1. The acronyms are used in Equations (1) and (2) and in Chapter 5 during performance demonstration and analysis.

Table 4-1. Equation acronyms and their descriptions

Variable	Description
L_a	The latency of a single iteration of the location-based evaluation (at the UTRAN) in addition to the time required to trigger the evaluation and receive the decision notification after the decision is reached
L_c	The latency associated with the addresses configuration procedures, which consist of obtaining a temporary IP address from a DHCP server at the new network after it has been checked using the Duplicate Address Detection Mechanism (DAD) procedure [49][50][61]. (Section 3.2.1.2)
L_{aaa}	AAA protocol latency, which depends on the number of proxies that must be traversed in order for the foreign network's AAA server to communicate with the user's home AAA server (HAAA). This communication is important in order for the FAAA to negotiate the user's credentials before granting certain authorizations [6][35][39]. It also depends on the latency associated with the AAA sub-procedures, such as session key generation, as shown in Figure 3-1 [6] (Section 3.2.1.3)
L_{mip}	Mobile IP latency associated with the registration step, where the visited network's FA sends a request to the user's HA to update it with the user's new CoA and receive an acknowledgement from the HA (Section 3.2.1.4).
L_s	Time required for TCP to recover from the residual backoff time caused by the long communication pauses during the handoff procedure, even after the L3 handoff procedures have completed. The experiments in [31][37][38] have demonstrated the significant impact this has on throughput and thus application response time (Section 3.2.1.5).
τ	Time required for receiving data equivalent to the data that would have been received in UMTS during the pause during handoff procedures where no user data was transferred (or negligible user data was transferred due to TCP back-off) This is to avoid a situation where: WLAN visit duration = Handoff latency
$\mathbf{R_c}$	Throughput (in Mbps or Kbps) of the current network where the user is located, which could either be a UMTS network or a visited WLAN network
$\mathbf{R_t}$	Throughput (in Mbps or Kbps) of the target network whose strong WLAN beacon was responsible for triggering the L1 and L2 handoff procedures as well as the location-based decision mechanism at the UTRAN.

The minimum Required Visit Duration (RVD) represents the aggregate handoff latency, which is composed of the individual latencies for the handoff sub-procedures as well as the algorithm latency itself as shown in the following equation:

$$\text{Minimum RVD} = L_a + (L_c + L_{mip} + L_{aaa} + L_s) + \tau \qquad (1)$$

As mentioned earlier, the minimum RVD value depends on the throughput of both the current and target networks. The value of τ is where the two throughput values are used as follows:

$$\tau = (L_c + L_{mip} + L_{aaa} + L_s) \times (R_c / R_t) \quad (2)$$

In other words, τ = (handoff latency without algorithm) $\times$ (R_c / R_t)

Network operators can adjust the value of the minimum RVD to support any other protocols or issues that they anticipate would cause further latencies. Furthermore, the value maybe increased to improve the chances of achieving benefit from the handoff to a WLAN hotspot. However, it is recommended that the customized value of RVD remain reasonably low in order to avoid wasting opportunities for transferring a user to WLAN if the handoff is suitable (i.e., the user's visit duration is reasonable).

4.5.3 User Travel Speed

Recall from Section 4.3.1 that a positioning mechanism would be triggered at Step (2) in the location-aided decision algorithm (Figure 4-2). The positioning method would obtain one or more position fixes for the MN to be used for position, speed, and direction calculations. In the future, these fixes may be obtained periodically and not require specific triggering by the algorithm. In that case, the accuracy and response time would greatly improve. Until then, the proposed approach assumes that position fixes would be acquired when Step (2) of the algorithm is invoked (Section 4.3.1). Based on the assumption a tradeoff between the accuracy and response time is important and depends on the type of positioning method and capabilities available.

4.5.4 Predicted Path Length

Given the value of the minimum RVD and the user's average travel speed (M_s), the algorithm computes the Predicted Path Length (PPL). PPL is the estimated distance that would be traveled by the user in "RVD" second, at the user's average speed. This provides the distance of the user's trajectory for a period equal to the minimum RVD, assuming the user maintain fairly fixed speed. This parameter will be used in a later step of the algorithm, as described in Section 4.5.6. The PPL is simply calculated as follows:

$$\text{Predicted Path Length (PPL)} = \text{RVD} \times M_s \quad (3)$$

4.5.5 User Travel Direction

The proposed approach relies heavily on the algorithm's knowledge of a reasonably accurate estimation of the user's current travel direction. Further details regarding the use of the user's travel direction is described in the following subsection. A minimum of two fixes must be obtained in order to compute the user's velocity vector, which represents the user's speed and direction (Figure 4-5). The number of fixes obtained depends on the accuracy-response time tradeoff, which are dependent on the capability of the positioning method. As mentioned previously, if the user's position was tracked to support other location-based services, then depending on the frequency of position updates, such tracking information could reduce the latency of computing the user's location-based parameters and reduce the algorithm's overall latency.

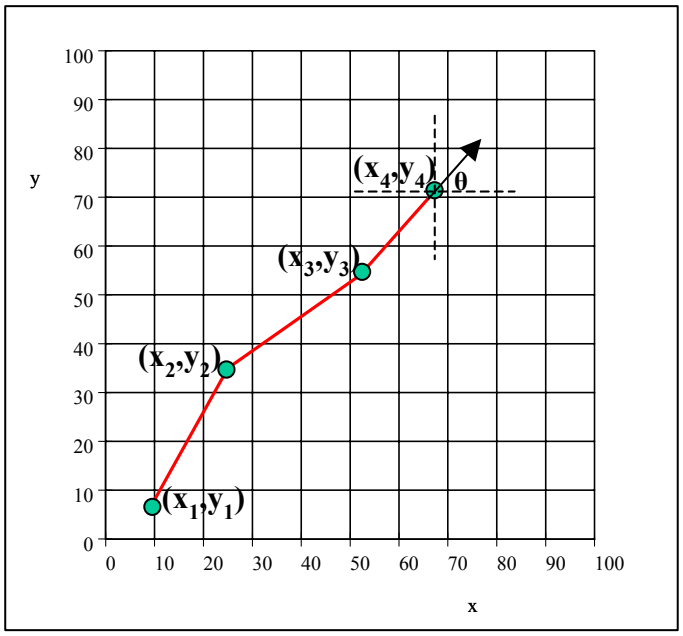

Figure 4-5. Visual representation of three velocity vectors obtained from 4 position fixes

4.5.6 Coverage Query

As described in Section 4.3.1, if the user's speed was not found to be negligible, the next step taken by the algorithm (Step (4)) would be to predict whether the user would remain within the same WLAN coverage area for at least the minimum RVD. This section describes the basic idea behind making this prediction. Note that at this point of the algorithm, the user's position, speed, and direction, have been computed. Furthermore, the minimum RVD and the user's average speed were used to calculate the

PPL, as described in Section 4.5.4. The algorithm then calculates the user's Trajectory End Point (TEP) coordinates (x_2, y_2), shown in Figure 4-6, based on the user's position coordinates (x_1, y_1), direction (θ), and the calculated PPL as follows:

$\text{Sin}^{-1}(\theta)$ = opposite/hypotenuse = a/PPL (4)

→ $a = PPL \times \text{Sin}^{-1}(\theta)$

$\text{Cos}^{-1}(\theta)$ = adjacent/hypotenuse = b/PPL (5)

→ $b = PPL \times \text{Cos}^{-1}(\theta)$

The coordinates of TEP (x2, y2) depend on the user's direction (θ):

If $0 < \theta < 90$ → $x_2 = x_1 + b$ and $y_2 = y_1 + a$

If $90 < \theta < 180$ → $x_2 = x_1 - b$ and $y_2 = y_1 + a$

If $180 < \theta < 270$ → $x_2 = x_1 - b$ and $y_2 = y_1 - a$

If $270 < \theta < 360$ → $x_2 = x_1 + b$ and $y_2 = y_1 - a$

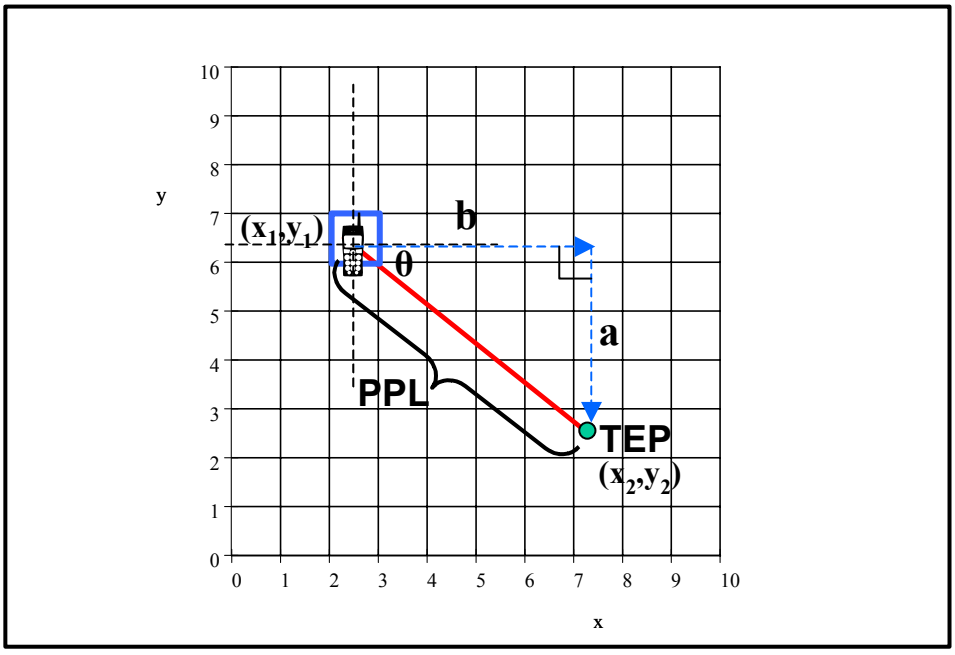

Figure 4-6. Obtaining TEP coordinates from the user's position, direction, and PPL

Upon completing the calculation of TEP coordinates, the algorithm would trigger a query to the WLAN coverage database. The trigger would include the UMTS cell ID where the user is located at that time, the WLAN hotspot ID of the target network, and the TEP coordinates. The cell ID and hotspot ID would then be used to retrieve the list of hotspot records that represent the target hotspot. This list varies in length depending on

the WLAN footprint representation resolution. A mapping mechanism would then link the TEP coordinates to the record/entry that encompasses the location represented by the coordinates. If the TEP coordinates do not fall within the WLAN hotspot footprint, then the algorithm does not recommend the handoff, and Step (5) (Figure 4-2 and 4-3) would be triggered. On the other hand, if the TEP coordinates were successfully linked to a record in the list (Figure 4-7), Step (8) would be invoked. Recall that this step would include sending a notification to the MN to recommend the triggering of handoff to the target hotspot.

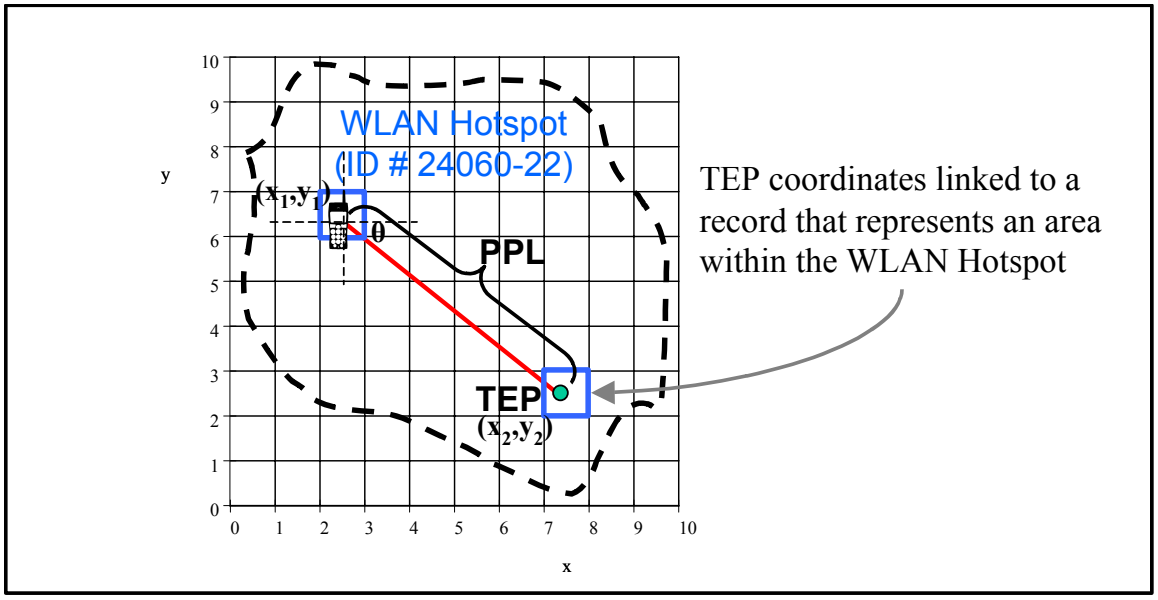

Figure 4-7. A visual representation of the coverage query mechanism based on TEP coordinates and the records representing the WLAN hotspot area at a certain resolution

The concept of using a WLAN coverage database to house records of WLAN footprints has been addressed by other research efforts to achieve different objectives. For instance, the research effort described in [68] proposed a positioning approach that was based on the idea of "RSS fingerprinting" rather than signal propagation (e.g., TDOA). The design involved collecting RSS fingerprints at a series of locations within WLAN hotspots and storing them in a database at a 1.5m resolution, to represent the WLAN hotspot's footprint. These databases were then indexed using RSS fingerprints, which were obtained through RF measurements at a WLAN adapter. Finally, a pattern recognition mechanism was used to link the RSS fingerprint, obtained at the MN, to a geographic location within the WLAN coverage area, thus obtaining the user's position.

The positioning concept proposed in [68] re-enforced the idea proposed in this section, which similarly involved the use of a WLAN coverage database. However, beyond the basic concepts associated with using such database, the two approaches were quite different, particularly with respect to objective, functionality, and underlying calculations.

4.6 Proposed Architecture

As mentioned in Chapter 2, Location Services (LCS) have been addressed by 3GPP and the architecture of the UTRAN was augmented to accommodate the entities/nodes that must be added to the infrastructure to support such services [9][10][11]. It was also determined in Chapter 2 that the proposed handoff decision algorithm acts as a "PLMN Operator client," a type of client that uses location-information to optimize performance of the network. Therefore, the LCS architecture in Figure 2-10 can be further augmented (Figure 4-8) to support the proposed approach by adding two main entities: a Vertical Handoff Decision Controller (VHDC) and a WLAN Coverage Database Server (WCDS).

The VHDC provides the functionality required to run the proposed algorithm in terms triggering the necessary actions in the UTRAN to obtain the algorithm's key input parameters (e.g., MN location fixes, speed, and direction). It also performs the computation for determining the duration that a user must be within WLAN coverage in order for the handoff to be beneficial. It then queries the WCDS over the defined Iu-w interface to determine whether the user's travel path, for the required visit duration, will have sufficient WLAN coverage.

The WCDS is a database server that holds the virtual 2-D maps of the areas served by the SRNC in the UTRAN. It communicates with the VHDC to respond to the queries regarding the existence of WLAN coverage in the user's path. The WCDS was placed in the UTRAN, rather than the UMTS core network or in the external network, to reduce the query response time. Another benefit of having a local WCDS in the UTRAN is to simplify the mechanism for updating the information regarding the WLAN coverage in the area served by the SRNC.

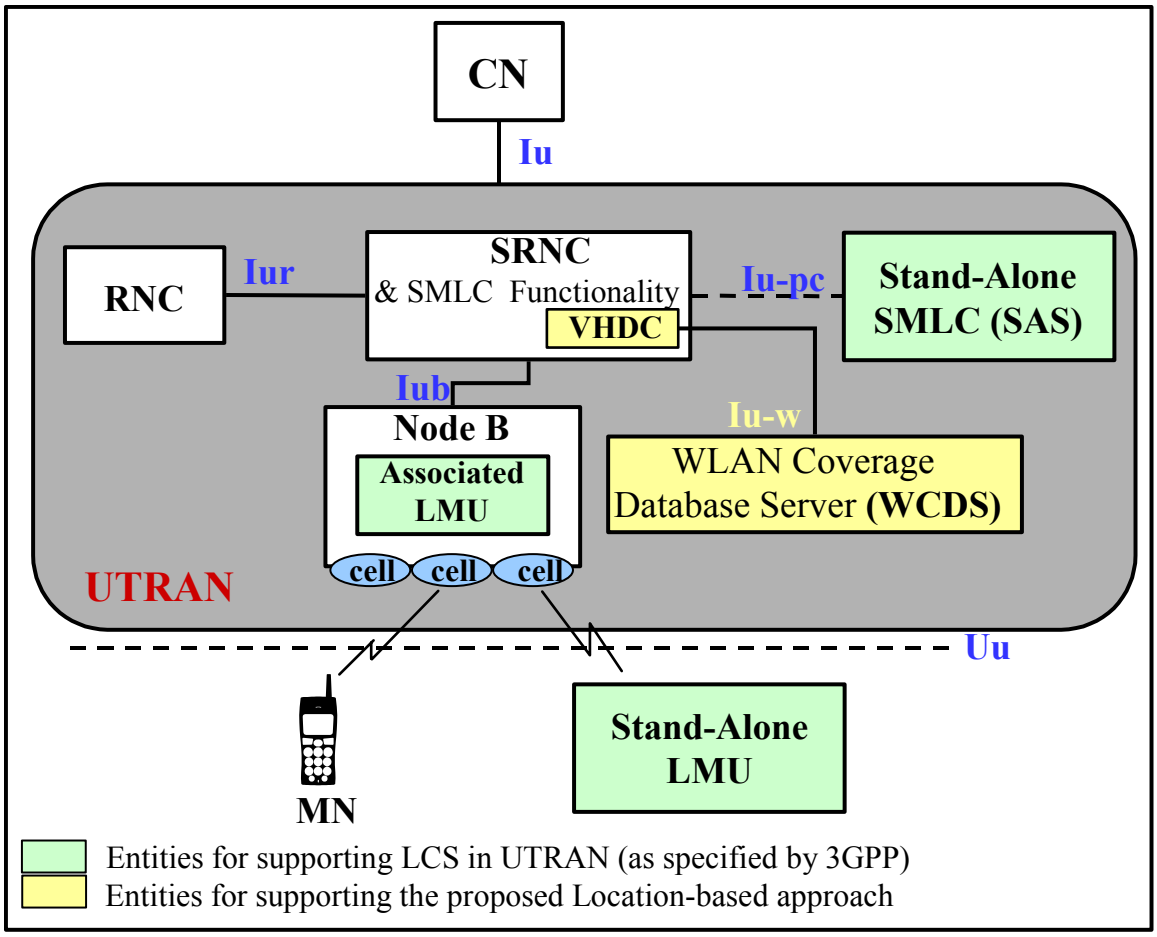

Figure 4-8. The architecture and configuration for supporting LCS in UTRAN (based on 3GPP TS 25.305, 2003) augmented to support the proposed handoff decision algorithm

Depending on the number of cells (i.e., Node-Bs) served by the SRNC, the size of such cells, the user-density and WLAN density in these cells, the WCDS may be able to accommodate more than one SRNC. On the other hand, in some cases where these attributes are high (e.g., urban areas with a large number of hotspots), additional WCDS units may be required to ensure a low query response time.

The VHDC functionality was placed in the SRNC for a number of reasons. The main reason is that the RNC, as described in chapter 2, is the central control unit in the UTRAN and is responsible for handoff control, among other responsibilities. Therefore, VHDC functionality will act as a natural extension to the handoff control mechanisms already established in the SRNC. Therefore, reducing the modifications necessary to support the proposed approach.

4.7 Communication Protocol

The proposed location-aided handoff decision algorithm requires the establishment of a protocol for exchanging commands and information between the MN and the UTRAN. Figure 4-9 shows the four different notification messages needed to support the proposed algorithm.

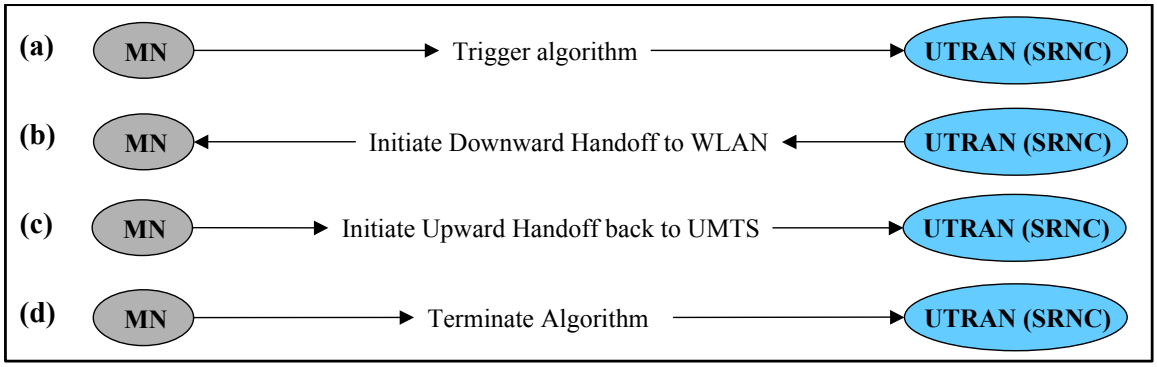

Figure 4-9. The protocol specifying the messages exchanged between the MN and the UMTS network to support the location-aided decision algorithm

The following is the description of each message type and the reasons that would cause it to be generated:

(a) This message would be sent to the UMTS network to trigger the location-based evaluation when the MN's WLAN adapter discovers WLAN coverage via a strong beacon/response. This message must carry the hotspot ID, described in Section 4.4.1.

(b) This message would be sent to the MN to recommend the triggering of the L3 handoff procedures (i.e., Mobile IP, AAA, etc).

(c) This message would be sent to the UMTS network to indicate the MN's desire to trigger an upward handoff due to the user exiting the WLAN coverage. Note that this message was not strictly used to support the proposed algorithm. It would be needed, however, with any algorithm that assumes an automatic handoff between UMTS and WLAN.

(d) This message would be sent to the UMTS network to terminate/abort the location-based evaluation running at the UTRAN if the user moves away from the WLAN coverage area. As previously described, this prevents infinite loops from occurring since the network would have no other way of knowing.

A potentially suitable approach for supporting these messages is to utilize the already existing protocols to carry such messages between the MN and the UTRAN. The RRC layer described in Section 2.2.3 is the most suitable for such a task due to its existing support for handoff procedures in UMTS [17].

4.8 Summary

This chapter provided the description of the location-aided algorithm design, which was proposed to meet the research objective stated in Chapter 1. The algorithm was based on the concept of combining the location-based decision mechanism with the basic conventional approach for triggering Layer-3 handoff procedures in heterogeneous wireless overlay networks. The algorithm mainly supported downward handoff from 3G/UMTS to WLAN and Layer-3 horizontal handoff to a visited WLAN. Upward handoff from WLAN to 3G/UMTS was not supported nor hindered by the proposed algorithm due to the reasons stated in Chapter 3.

The conventional approach was augmented without changing the basic flow of events that comprises the IEEE 802.11 specifications with regards to PHY and MAC layer handoff. The location-based evaluation was designed to wait for a trigger from the MN when WLAN coverage was discovered via RF measurements at the MN's WLAN adapter. The location-based evaluation would then run on the network side and notify the MN if the handoff is suitable. Otherwise, the evaluation would continue monitoring the user's position and computes the TEP if changes occur. This continues until the user exits WLAN coverage or the user's location-based information indicates that he/she will remain within WLAN coverage for the minimum RVD, thus triggering the handoff to the discovered WLAN.

The underlying assumption, specifications, requirements, and calculations were also addressed in this chapter. The main underlying mechanism for obtaining and computing the minimum RVD, PPL, and TEP-coordinates was described in detail. Furthermore, the concept of using a WLAN coverage database to predict whether the user would remain within the same WLAN coverage for the RVD was also described in this chapter. In addition, the modifications that would be imposed on the UTRAN architecture were addressed with respect to the 3GPP standard architecture for location-

based services (LCS). Finally, the basic control messages that would be required to support the proposed algorithm's functionality, in terms of a communication protocol between the MN and the UTRAN, were briefly described in an effort to provide a more comprehensive look at the proposed algorithm's requirements.

Chapter 5: Validation & Analysis

5.1 Introduction

The proposed location-aided algorithm described in Section 4.3 does not entirely replace the conventional RF-based handoff method. Instead, it encompasses the functionality of the RF-based method and uses such functionality to discover the existence of WLAN coverage during the WLAN "Coverage Discovery Stage." The successful completion of the "Coverage Discovery Stage" then triggers the "Location-based Evaluation and Decision Stage" at the MN. The latter stage begins by sending a notice to the UTRAN to invoke the location-based evaluation for assessing the user's velocity, and the average handoff latency between the two networks. It uses this information along with the spatial information of the discovered WLAN to predict whether the user's visit to such coverage area will be long enough to allow the application to benefit from the higher data rates. In other words, the location-aided algorithm pre-examines the suitability of the handoff based on more parameters than merely the received signal strength of a beacon, which is the basis for the conventional method.

This chapter addresses the algorithm's functionality, capabilities, limitations, and benefits. Section 5.2 demonstrates the algorithm's functionality and performance by comparing the outcome of the algorithm in two primary scenarios to the outcome of the conventional approach. It also addresses the algorithm's limitations by examining its potential performance under worst-case scenarios. Section 5.3 demonstrates the algorithm's performance in a quantitative manner by addressing the end result/benefit of applying the algorithm. Section 5.4 briefly addresses the environments where the algorithm would be most suitable and beneficial based on the evaluation of its capabilities and limitations in this chapter. Section 5.5 provides a comprehensive summary of the differences between the conventional handoff decision method and the proposed algorithm including functionality, performance, advantages/disadvantages, etc. Finally, Section 5.6 provides a chapter summary along with some concluding remarks.

5.2 Performance Demonstration & Analysis (Qualitative)

In order to demonstrate the performance and functionality of the location-aided handoff decision algorithm proposed in Section 4.3, two primary scenarios were established, as described in this section. They were used to show how the proposed algorithm differs from the conventional handoff decision method in terms of reaching a different decision under the same conditions. As described in Section 4.2, the conventional approach is essentially divided into a "Coverage Discovery Stage" and the handoff "Execution Stage."

Upon receiving a strong beacon or probe response at the WLAN adapter, the handoff "Execution Stage" begins. Recall that the "Execution Stage," as described in Chapter 4, consists of the handoff procedures described in Chapter 3. The conventional approach bases its decision on the received beacon's strength at the location where the WLAN adapter received such beacon. Therefore, for a user traveling at a negligible speed, the conventional approach and the proposed algorithm would both yield the same outcome, which is to trigger the handoff to the discovered WLAN. This is due to Step (3) of the location-based evaluation running at the UTRAN. Recall from Section 4.3 that Step (3) is responsible for checking the user's speed to determine whether it is negligible. If so, it would generate and send a notice to the MN to trigger the handoff to the discovered WLAN upon receiving the notice. Therefore, in such cases where the user is static or moving at negligible speed, no benefit is gained from the use of the algorithm instead of the conventional handoff method.

On the other hand, the algorithm's full functionality and main benefit would be observed when the user is traveling at non-negligible speed. This is due to the fact that the algorithm was developed to alleviate a handoff-related issue, which typically results from user mobility between different wireless cells. Furthermore, it was designed for handoff between 3G/UMTS and WLAN in areas where the probability of a short visit to WLAN is significant, such as city blocks, airports, parks, university campuses, malls, etc. These areas are the primary spots where the conventional handoff decision method could cause performance degradation at the application level as further addressed throughout this chapter.

5.2.1 Ideal Algorithm Performance

The following scenarios demonstrate the performance of the algorithm during user mobility. These scenarios were developed under an assumption of an error free environment (ideal conditions) to eliminate irrelevant factors and highlight the performance of the algorithm compared to the conventional method.

5.2.1.1 Scenario-1 (Downward Vertical Handoff)

(A) Only Conventional Method Applied

In this scenario (Figure 5-1), the MN is connected to the user's home 3G/UMTS network, which overlaps with the WLAN coverage area in Torgersen Hall (highlighted in red in Figure 5-1). The user activates a non-real-time application, which triggers the WLAN "Coverage Discovery Stage" (shown in Figure 4-1). In this scenario, the user is already within the coverage area of a WLAN in Torgersen Hall. Therefore, upon discovering the existence of the WLAN, the handoff "Execution Stage" is triggered, which includes the L3 handoff procedures described in Chapter 3.

The problem in this case is that the conventional method does not take into consideration that the user is headed out of the discovered WLAN coverage area. This is due to the fact that it is unable to examine the user's position, speed, and direction, which will result in a visit shorter in duration than the period required for completing the L3 handoff procedures. In other words, the user's visit duration (UVD) is less than the minimum Required Visit Duration (RVD) to compensate for the cost/impact of the handoff, which is further addressed in Section 5.3. The minimum RVD was described in Section 4.5.2 and is essentially the amount of time that the user must remain within the discovered WLAN's coverage area to complete the handoff procedures and transfer data over the new WLAN connection.

As the user moves towards the exit of the building and out of the coverage area, the MN's WLAN adapter notices degradation in the signal to noise ratio (SNR) value from the current AP. Therefore, it triggers the WLAN "Coverage Discovery Stage" to search for better WLAN coverage. However, in this scenario, no other APs are available since the user is exiting the building. The SNR value drops below the configured

threshold thus triggering the UMTS "Coverage Discovery Stage" at the MN in an attempt to perform an upward handoff back to UMTS, as shown in Figure 4-1.

In this case, the downward handoff execution was triggered inappropriately and did not result in any benefit since no data was transferred over the new WLAN connection. Furthermore, this handoff would in fact result in a negative impact on performance as further addressed in Section 5.3.

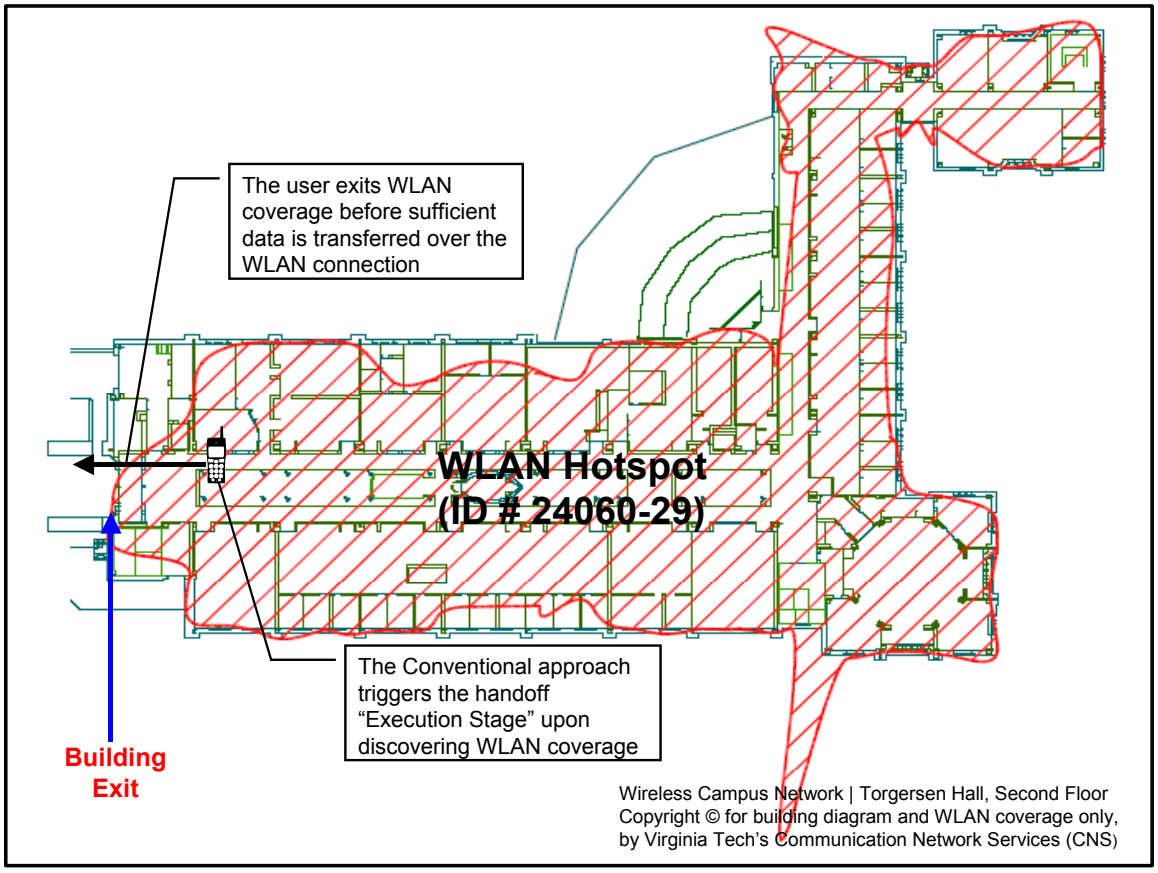

Figure 5-1. Scenario-1 showing an outcome caused by conventional approach limitations

(B) Proposed Location-aided Algorithm Applied

Recall from the algorithm description in Section 4.3 that upon receiving the first advertisement from the router associated with the discovered WLAN AP, the "Location-based Evaluation and Decision Stage" is triggered, thus triggering the location-based evaluation at the UTRAN. The algorithm would have then obtained the user's position, mean speed, and direction then utilized such information along with the minimum RVD value to calculate the predicted path length (PPL) for the user as follows:

τ = (Handoff Latency) x (R_c / R_t)

RVD = L_a + (Handoff Latency) + τ

PPL = RVD x M_s (Where M_s is the user's mean travel speed)

The PPL value and the user's direction angle would have then been utilized for computing the TEP coordinates as described in Section 4.5.6. The WLAN coverage database serving the user's current UMTS cell would have then been queried using the UMTS cell ID, the Hotspot ID for the WLAN in Torgersen Hall, and the TEP coordinates for the user.

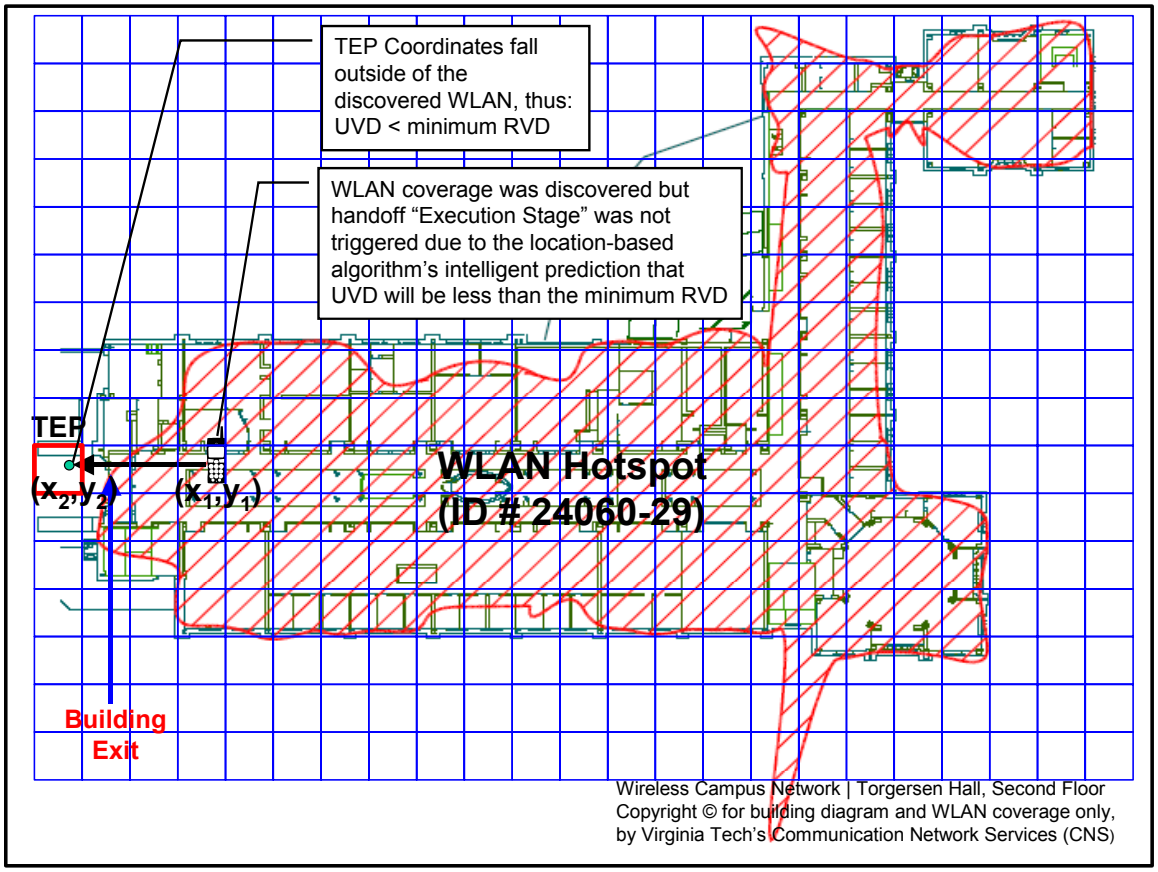

Figure 5-2. The functionality and outcome of the location-aided algorithm in Scenario-1

Note that the grid shown in Figure 5-2 aims to provide a visual representation of how the WLAN coverage is stored in the WLAN coverage database. As described in Chapter 4, the TEP coordinates are mapped to the appropriate grid block (in this case the block is highlighted in red). If the grid block belongs to the discovered WLAN coverage

area, which is not the case in this scenario, then the algorithm triggers the L3 handoff procedures to the discovered WLAN.

In this scenario, however, the algorithm running at the UTRAN determined that the user's current TEP is outside of the WLAN coverage area. Therefore, instead of triggering the handoff "Execution Stage," the algorithm sets a "wait" timer (Step (5) in Figure 4-2) and iterates after it expires to obtain new TEP coordinates and use them for a new database query. In the meantime, the MN refrains from triggering handoff "Execution State" and continues to transfer data over the UMTS network while waiting for a decision from the UTRAN.

According to Figure 5-2, the user is heading towards a building exit and thus is heading away from the WLAN coverage area. Therefore, eventually the value of the SNR degrades, thus triggering the WLAN "Coverage Discovery Stage" in an attempt to find better WLAN coverage. When the SNR drops below the configured threshold, which occurs here due to exiting the building, the algorithm at the MN sends a notification to the UTRAN to set the "Abort" flag, which causes the algorithm to terminate. It then returns to the "Coverage Discovery Stage" to search for other WLAN coverage.

Essentially, this scenario showed that if the location-aided handoff decision algorithm were applied to in place of the conventional method, the downward handoff from UMTS to the discovered WLAN (in Torgersen Hall) would not have been triggered, thus the handoff-related performance degradation would have been prevented. Further details regarding the benefit of preventing such handoff event are provided in Section 5.3.

5.2.1.2 Scenario-2 (Layer-3 Horizontal Handoff)

(A) Only Conventional Method Applied

Another scenario is shown in Figure 5-3. This scenario illustrates the same shortcoming of the conventional approach but in the context of a layer-3 horizontal handoff to a discovered WLAN. In this scenario, the user has already performed a handoff from its home 3G/UMTS network to a visited WLAN (Hotspot-1) and is currently receiving data over the established WLAN connection. As the user moves towards the border of the current WLAN coverage area, the SNR begins to degrade thus

triggering the WLAN "Coverage Discovery Stage" to search for other coverage. In this scenario, the MN's WLAN adapter receives a strong beacon from an AP that belongs to a different WLAN (Hotspot-2) as shown in Figure 5-3. Upon receiving the strong beacon and discovering the AP's associated router, the conventional approach triggers the handoff "Execution Stage" (i.e., L3 handoff procedures) to establish the connection through Hotspot-2. However, in this scenario, the user is moving at a speed and direction that results in him/her heading out of the newly discovered WLAN coverage area as well.

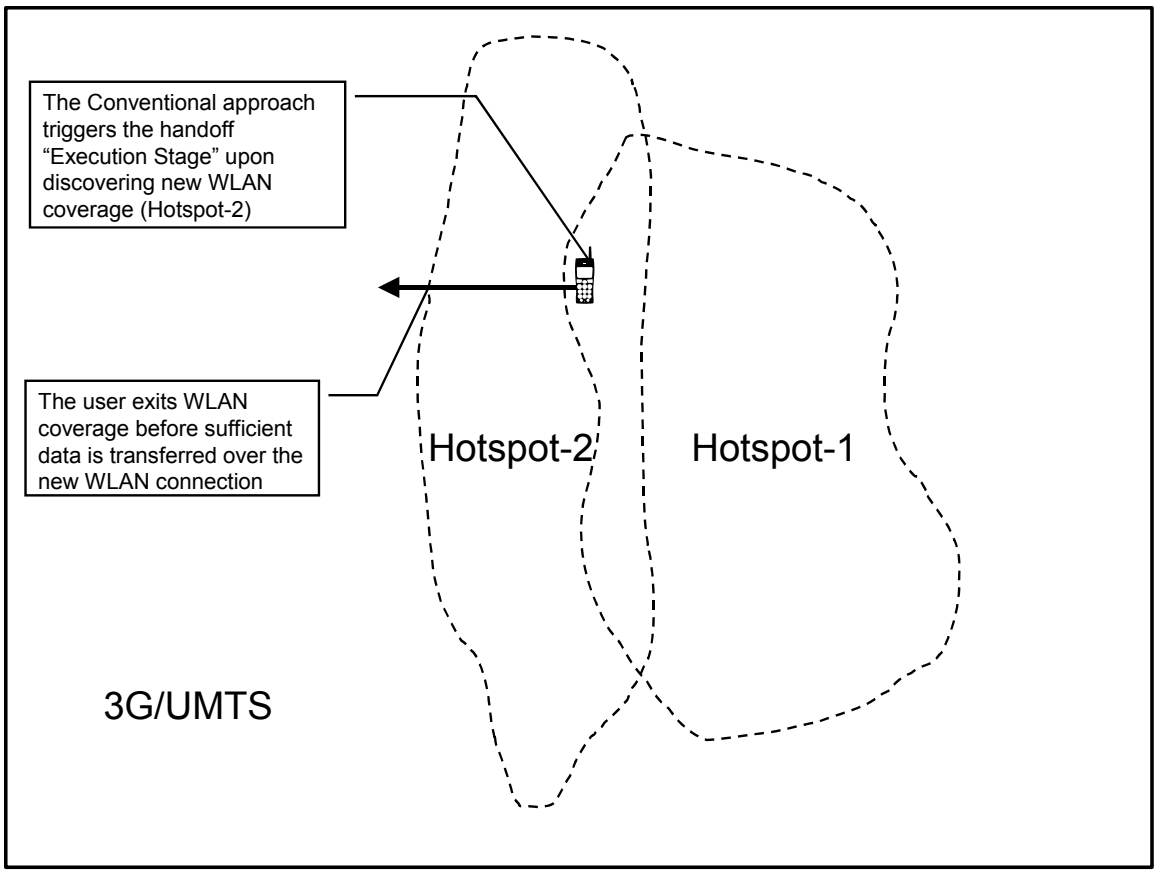

Figure 5-3. Scenario-2 showing an outcome caused by conventional approach limitations

As with the previous scenario, the MN's WLAN adapter noticed degradation in the SNR and triggered the WLAN "Coverage Discovery Stage," which failed to find any other coverage. When the SNR finally dropped below the configured threshold, the UMTS "Coverage Discovery Stage" was triggered to perform an upward handoff back to UMTS. Therefore, the handoff from Hotspot-1 to Hotspot-2 was wasteful of resources (i.e., bandwidth for handoff control messages, and processing power at the MN). It also

contributed to performance degradation rather than enhancement. Further details regarding the extent of performance degradation are provided in Section 5.3.

(B) Proposed Location-aided Algorithm Applied

Had the location-aided algorithm been applied in the scenario shown in Figure 5-3, it would have taken into consideration the user's position, speed, and direction as well as the minimum RVD value. Based on such parameters, the TEP coordinates would have been computed then used to query the WLAN coverage database. As shown in Figure 5-4, the grid block containing the TEP is outside of the coverage area of Hotspot-2. Therefore the algorithm would not have triggered the handoff "Execution Stage" (i.e., L3 handoff procedures).

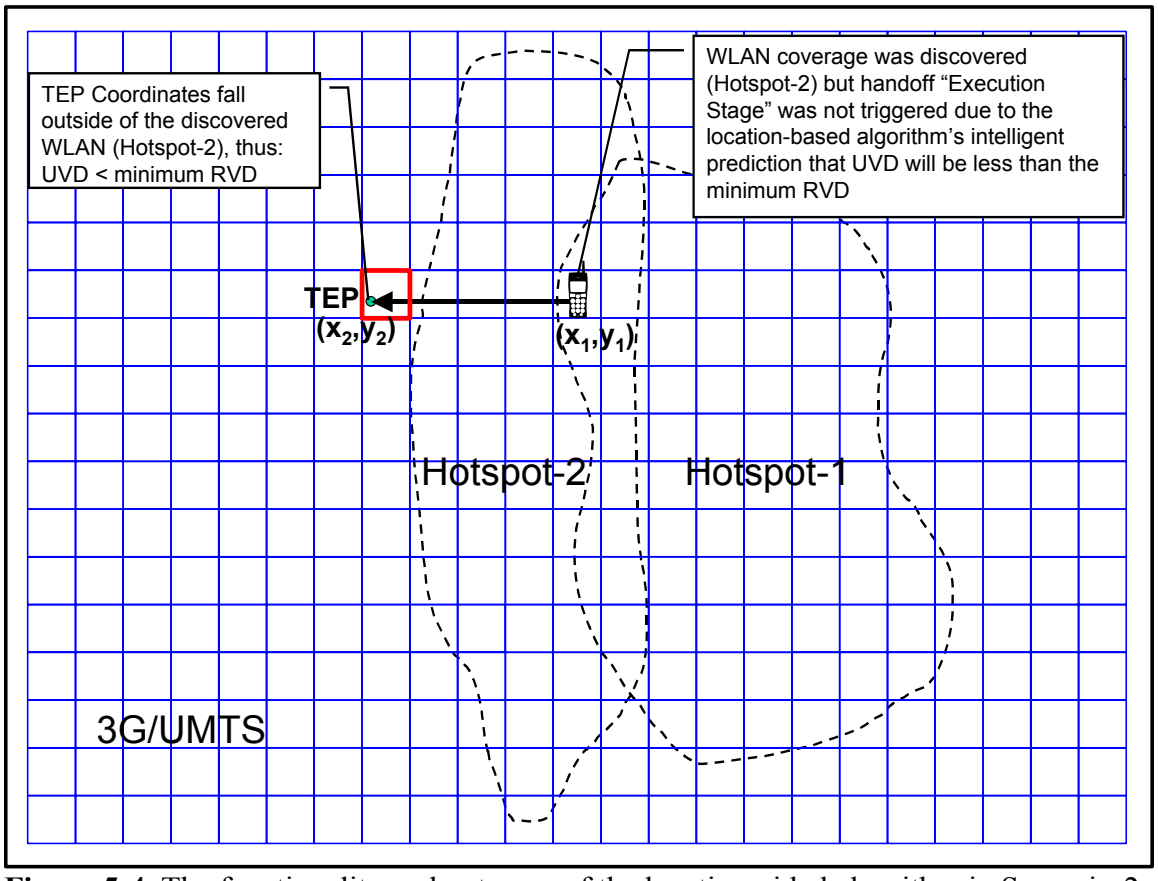

Figure 5-4. The functionality and outcome of the location-aided algorithm in Scenario-2

The MN in this case would have continued to transfer data over the WLAN connection established with Hotspot-1 until the SNR value of that WLAN degrades

below the configured threshold. Upon dropping below the threshold, the UMTS "Coverage Discovery Stage" would have been triggered to attempt an upward handoff from WLAN (Hotspot-1) directly to the 3G/UMTS network, without wasting resources in performing an inefficient and unnecessary handoff to Hotspot-2.

This scenario demonstrated the algorithm's performance during a situation where the user was already connected to another WLAN (i.e., Hotspot-1) when entering the boundaries of a new WLAN (i.e., Hotspot-2). However, the algorithm's outcome would have been the same had the user been connected to 3G/UMTS when entering the coverage area of Hotspot-2, as shown in Figure 5-5.

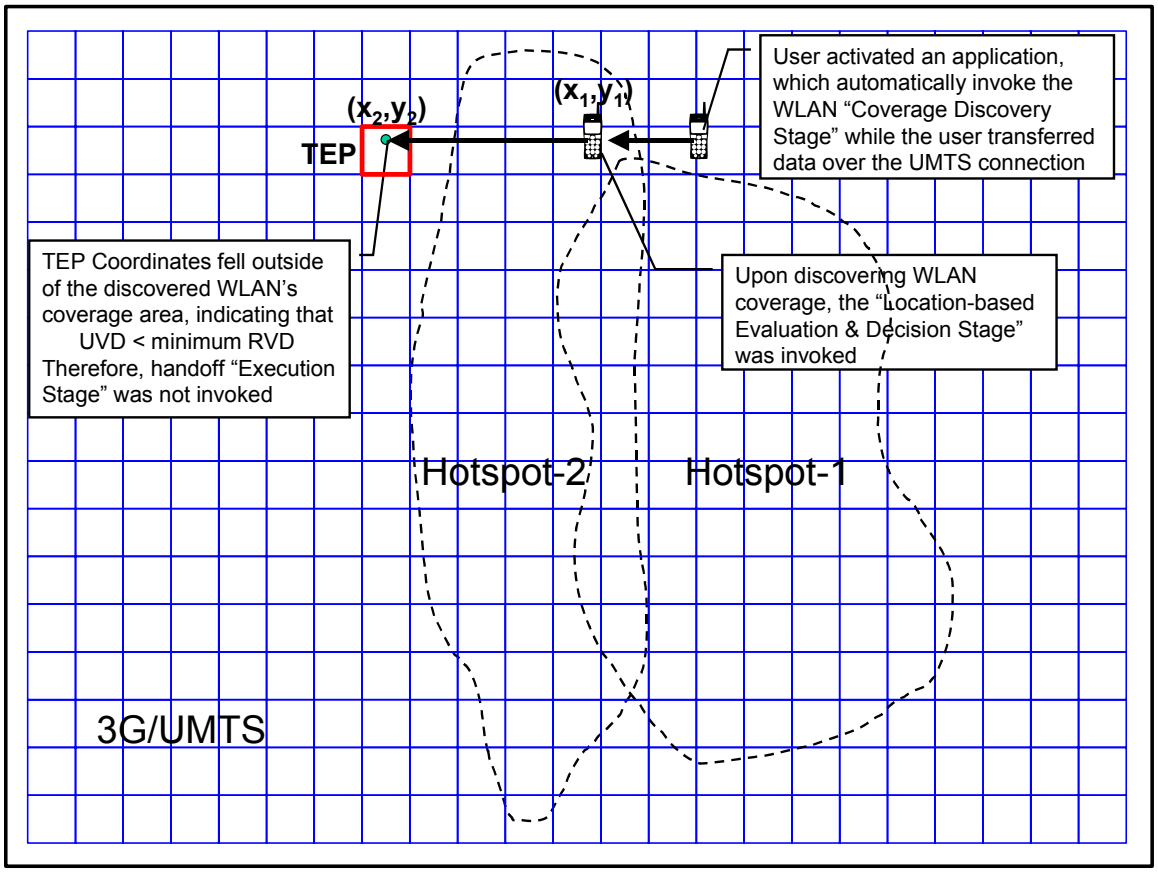

Figure 5-5. The algorithm's performance and outcome if the user enters Hotspot-2 while transferring data over the 3G/UMTS connection

5.2.2 Algorithm Limitations

This section provides a description of the algorithm's performance limitations and weaknesses observed in the worst-case scenario. Three main limitations are addressed in this section and illustrated using scenarios similar to those used in the previous section.

5.2.2.1 The Physical Boundary Limitation

As shown in Figure 5-6, the physical boundary limitation or the "no-exit" problem is mainly observed in indoor WLAN coverage areas. It occurs when the proposed algorithm is triggered as the user moves towards the boundaries of the building, which could also be the WLAN's coverage area.

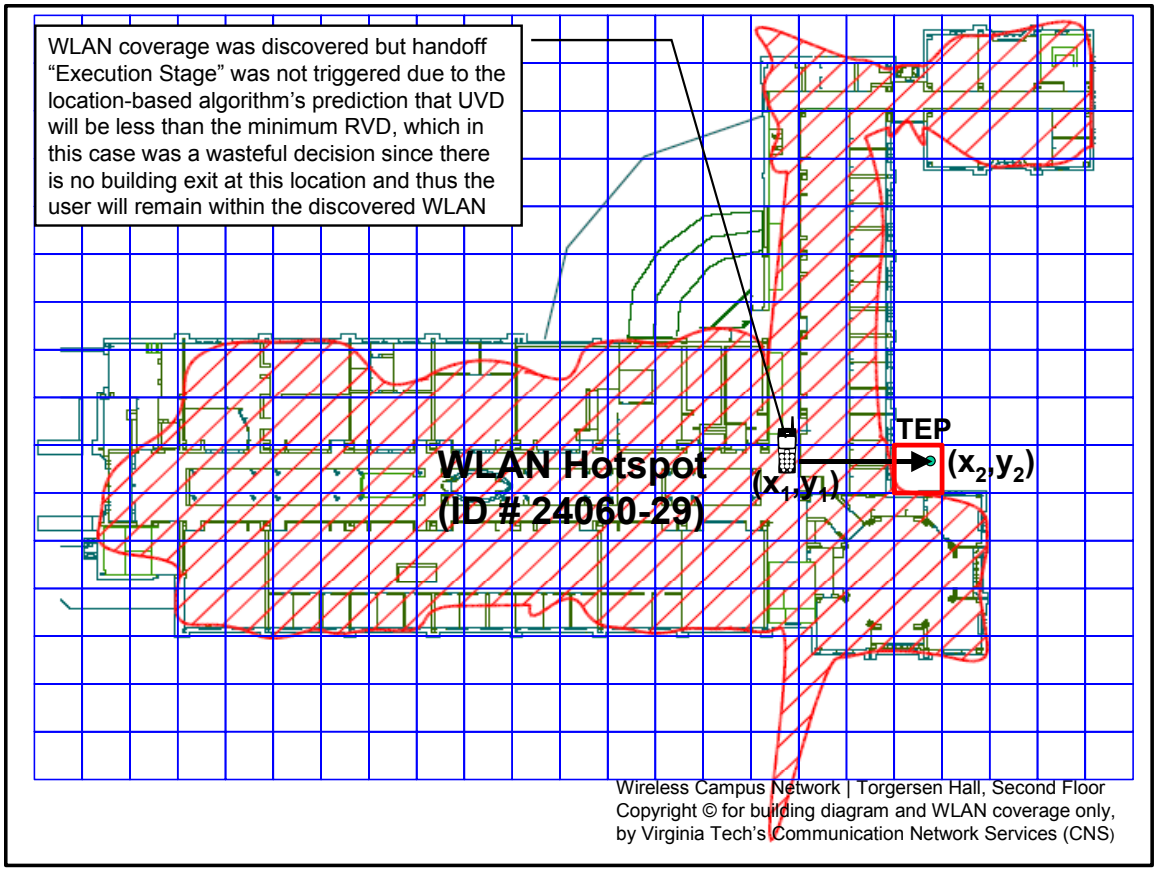

Figure 5-6. A scenario demonstrating the physical boundary limitation

In this scenario, the algorithm's computations determine that the user's TEP falls outside of the WLAN coverage area. Therefore, the location-based evaluation at the UTRAN does not notify the MN to trigger a handoff to the discovered WLAN. However, since this is an indoor WLAN coverage area whose boundaries are defined by the physical space (i.e., the building) in which it is deployed, the user cannot exit the WLAN coverage area unless a building exit is available. Therefore, in such situation, the algorithm's initial outcome is to set the "wait" timer and not trigger the handoff. Since there is no building exit in this case, refraining from the handoff is considered wasteful.

However, this decision does not persist since the user will either be forced to change his/her direction or reduce his/her speed as he/she approaches the boundaries of the building.

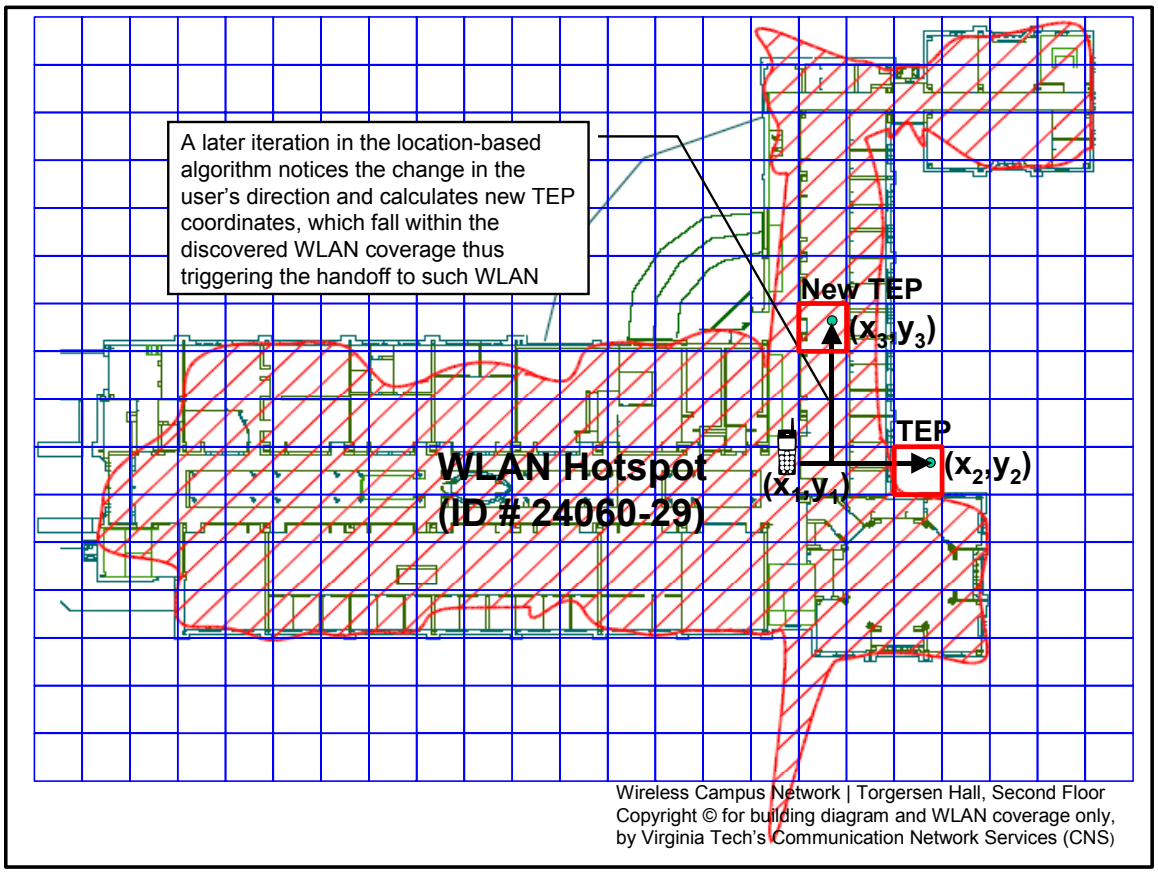

Figure 5-7. The algorithm's self-correction mechanism for handling the no-exit issue

In the former case, the algorithm's new iteration will compute a new TEP and submit a new query to the WLAN coverage database, which will determine that the new TEP falls within the discovered WLAN's coverage area (Figure 5-7). Depending on the spatial characteristics of the WLAN and the amount of change in the user's direction, a number of iterations maybe required before finally triggering the handoff. On the other hand, if the user does not change his/her direction but rather reduces his/her speed to a negligible level due to approaching the boundaries of the building, the algorithm's new iteration will immediately notify the MN to trigger the handoff without the need to query the WLAN coverage area, as described in Section 4.3.

5.2.2.2 The Sudden Drastic Change Limitation

The sudden drastic change limitation would occur if the user makes a significant change in his/her direction (Figure 5-8) or speed (Figure 5-9) after such parameters were processed by the location-based evaluation running at the UTRAN

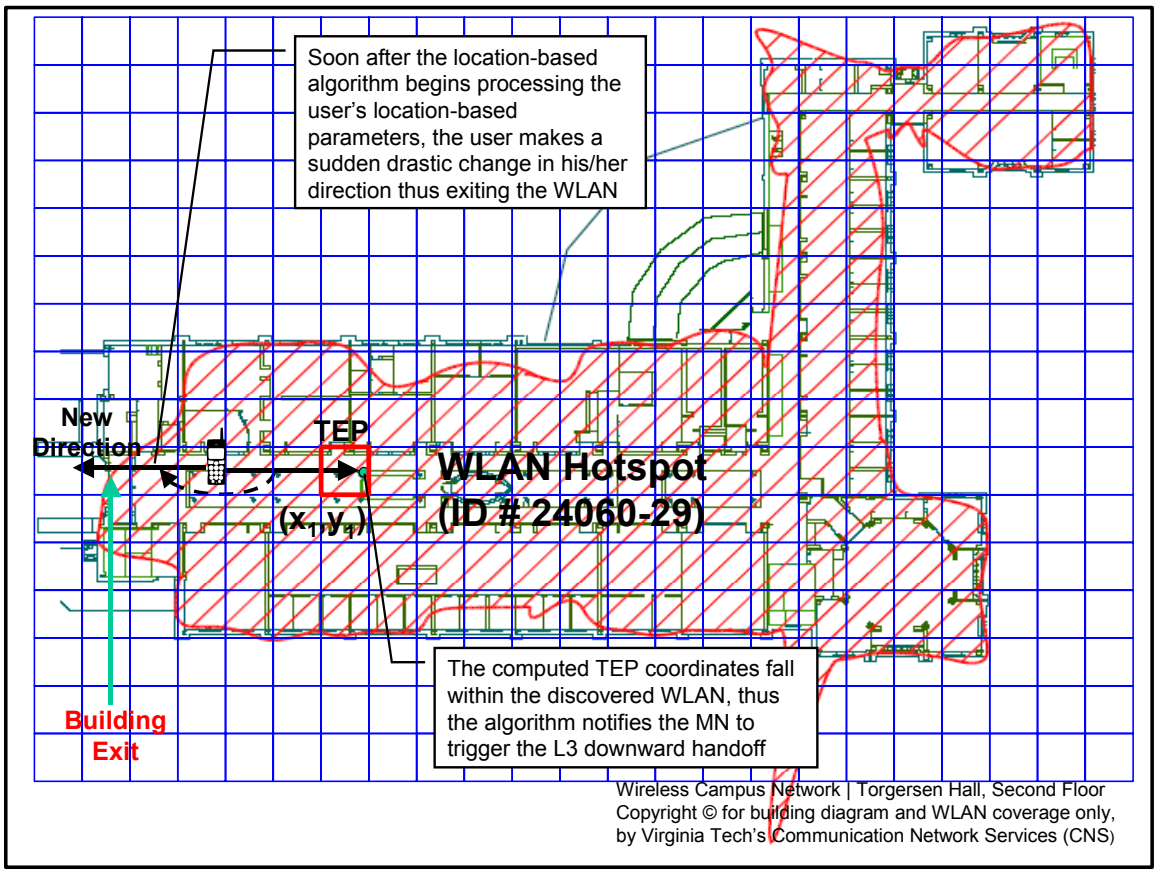

Figure 5-8. A scenario showing the impact of the sudden drastic change in direction

The worst-case scenario is shown in Figure 5-8 and Figure 5-9 to demonstrate the maximum negative outcome caused by such limitation. The user in Figure 5-8 changes his/her direction after the algorithm begins processing the location-based parameters at the UTRAN. On the other hand, the user in Figure 5-9 increases his/her speed after the algorithm begins processing such parameters. In both cases, the algorithm computes the TEP coordinates according to the initially obtained position, speed, and direction values then queries the WLAN coverage database with the TEP coordinates.

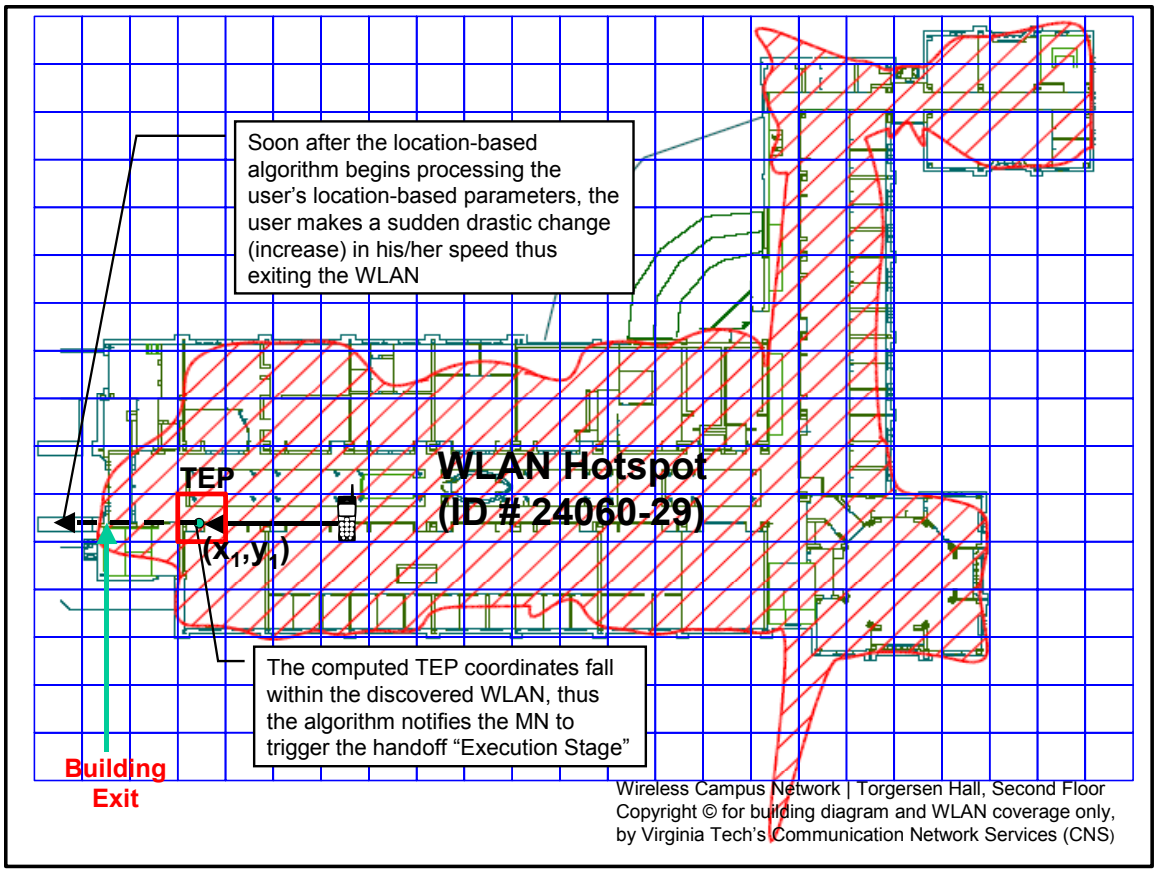

Figure 5-9. A scenario showing the impact of the sudden drastic change in speed

In both cases, the obtained parameters result in a TEP within the WLAN coverage area thus causing the algorithm to notify the MN to trigger the WLAN handoff "Execution Stage." However, due to the sudden drastic change in either user's case (speed or direction change), the actual UVD is less than the time required to complete the handoff procedures and transfer sufficient data over the WLAN connection. Therefore, despite the triggering of the WLAN handoff "Execution Stage," the user exits the WLAN coverage area before any benefit can be achieved from such handoff. Instead, the handoff-related service interruption results in performance degradation and wasting of resources as further addressed in Section 5.3.

5.2.2.3 The Algorithm Latency Limitation

The algorithm's latency (L_a) was not a concern in terms of adding delay to the period between WLAN discovery and downward handoff execution. This is due to the

fact that handoff to lower networks (i.e., WLAN) from an upper networks (i.e., 3G/UMTS) is typically triggered for performance optimization reasons, not lack of coverage. This argument was addressed in both Sections 2.5.2 and 3.2.1.1.

The value of L_a depends heavily on the delay associated with obtaining the user's position fixes, which depends on the acquisition method used to obtain such fixes. For instance, an unreasonably large L_a may result from an obstruction of the satellite signal (in the case of GPS) or longer processing time at the LMU due to network overload (in the case of U-TDOA), etc. Figure 5-10 shows the worst-case scenario where the value of L_a may result in a conservative decision that wastes an opportunity for a handoff to WLAN.

Recall that the value of PPL depends on the value of the minimum RVD, which includes the value of L_a. Therefore, the value of L_a has a direct impact on the computed TEP coordinates. Since the value of L_a is directly proportional to the value of the minimum RVD and PPL, then such value can have a misleading effect on the algorithm's decision. Therefore, the value of L_a should be calculated every time position fixes are obtained in order to provide the location-based evaluation mechanism with reasonably accurate values to include in its computations.

The worst-case scenario was addressed here in order to demonstrate the maximum negative impact of an unreasonably long L_a value. In this scenario, the location-based evaluation would have determined that the TEP falls within the WLAN coverage area had the value of L_a been reasonable. However, due to the high value of L_a in this particular scenario, the PPL was larger, and thus the TEP coordinates fell outside of the WLAN coverage area, as shown in Figure 5-10. As a result, the algorithm makes the conservative decision to not trigger the handoff during this iteration. Note that this decision does not persist, particularly since there is no exit in this location. As the user either slows down or changes his/her direction, the algorithm computes a TEP that falls within the coverage are, and thus triggers the handoff.

In the observable future, this issue should eventually become obsolete as the number of location-based services, which require user-tracking (i.e., periodic updates of the user's position) increases. In this case, the position acquisition time, which is the primary contributor to the algorithm's overall latency, would no longer fill such a role.

However, until then, the algorithm's decision may be affected in some cases by the algorithm's own latency value as shown in Figure 5-10.

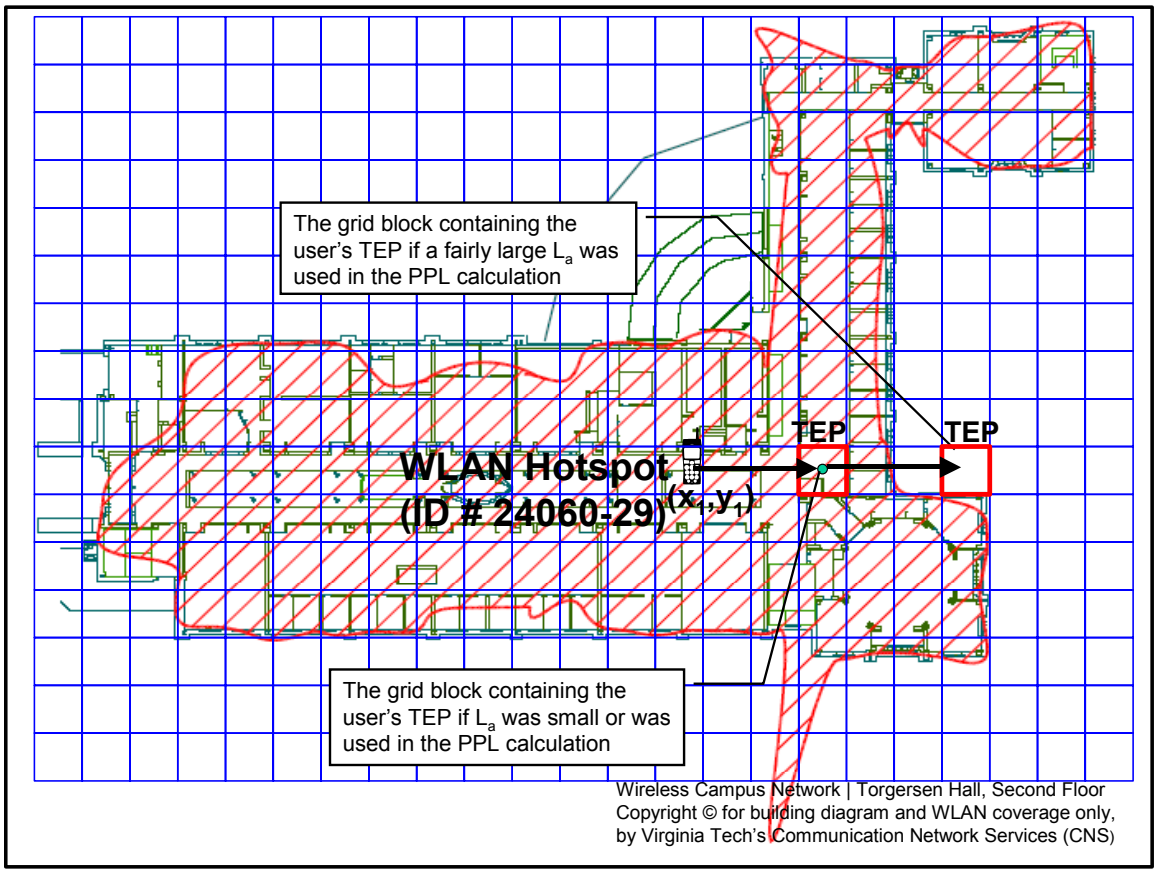

Figure 5-10. A scenario showing the potential worst-case scenario when L_a is significant

5.3 Algorithm Benefit (Quantitative)

As described throughout Chapters 2 and 3 of this document, there are significant benefits to transferring users from 3G/UMTS networks to 802.11 when coverage is available, mainly with respect to increasing the network capacity and providing the users with significantly higher data rates (in some cases an order of magnitude higher). However, transferring users from their current 3G/UMTS network to a WLAN network requires a number of handoff procedures to support mobility for the application(s) running on the MN. These handoff procedures were described in Chapter 3 along with details regarding their specifications and potential latencies.

According to the information found in the literature, the aggregate delay associated with the downward and upward handoff execution process is relatively

significant as shown in Tables 5-1, 5-2, and 5-3 [31][37][38][62][63][64]. These values are used in this chapter as nominal values to demonstrate the potential benefit of the location-aided handoff decision algorithm compared to the conventional handoff decision method, in a quantitative manner.

Table 5-1 shows the latency of the handoff procedures for the handoff of a MN from the user's home 3G/UMTS network to a discovered foreign WLAN. If the user was already in a visited 3G/UMTS network, the latency could potentially be greater. On the other hand, Table 5-2 shows the total latency for 3G/UMTS coverage discovery and upward handoff procedures for a MN handoff from the visited WLAN to the user's home 3G/UMTS network. If the user had exited the coverage area of his/her home 3G/UMTS network while moving through the visited WLAN coverage area, then the upward handoff latency would be as shown in Table 5-3. The difference between the values in Table 5-2 and Table 5-3 is a result of the less demanding handoff procedures required to reestablish a connection with the user's home network. In the upward handoff back to a home 3G/UMTS network, such procedures would mainly consist of re-establishing the radio bearers and resetting the user's original configuration (e.g., unbind the user's home address from its foreign care-of-address) to resume data transfer over the home network.

Table 5-1. The latency associated with the downward vertical handoff from the home 3G/UMTS to a discovered foreign WLAN

Latency Division	Min (ms)	Mean (ms)	Max (ms)
DHCP and DAD (L_c) [50]	2593.6	2621.1	2662.8
AAA (L_{aaa}) [40]	5785	5785	5785
Mobile IP and Stabilization ($L_{mip} + L_s$) [31]	2585	4654	7639
Total Execution Latency:	10963.6	13060.1	16086.8

Note that the AAA-related latency found in [40] and shown in Table 5-1 and 5-3 assumes a direct connection between the FAAA and the HAAA. However, in many cases, the AAA protocol messages must traverse a number of relay agents or proxies during the communication between the FAAA and the HAAA [39]. The simulation in [39] has shown that in addition to the AAA authentication delay shown in the tables below, a significant delay may be caused by the protocol due to the traversal of a few agents or proxies. This delay depends on the number of proxies traversed and thus is

specific to each pair of inter-networked 3G/UMTS and WLAN. Therefore, the average value for a specific 3G/UMTS-WLAN pair may be used in the calculation of the minimum RVD, which should be updated periodically as described in Section 4.5.2.

Table 5-2. The latency for discovering the home 3G/UMTS network and performing the upward vertical handoff procedures from the visited WLAN to the home 3G network [31]

Min (ms)	Mean (ms)	Max (ms)
5322	6896	8833

Table 5-3. The latency for discovering a foreign 3G/UMTS network and performing the upward vertical handoff procedures from the visited WLAN to foreign 3G network

Latency Division	Min (ms)	Mean (ms)	Max (ms)
3G Discovery [31]	200	808	1148
DHCP and DAD (L_c) [50]	2593.6	2621.1	2662.8
AAA (L_{aaa}) [40]	5785	5785	5785
Mobile IP and Stabilization ($L_{mip} + L_s$) [31]	2339	2997	3649
Total Execution Latency:	10817.6	12211.1	13244.8

During the handoff execution stage, data transfer is interrupted at a certain point and remains halted or significantly throttled until the end of the L3 handoff procedures are completed and the traffic is re-routed to the user's new location. This period includes the post-handoff stabilization period, which as indicated in the literature and in the above tables, could be up to a few seconds [31][37][38].

In case of a downward handoff from 3G/UMTS to WLAN, the delay associated with the handoff procedures would interrupt the data transfer for a certain period (e.g., 13.06s on average according to Table 5-1). However, after the handoff completes successfully, the application benefits significantly from the higher data rates, which are sometimes an order of magnitude higher than its current 3G/UMTS network's data rates. Therefore, a handoff from 3G/UMTS to 802.11 is typically desirable, as addressed in Chapter 2. In other words, the higher data rates in WLAN compensate for the interruption in data transfer thus resulting in better performance in the end, as observed by the user. However, as shown in Section 5.2.1, there are cases where the user's visit duration to the coverage area of a WLAN is not long enough to take advantage of the better data rates. The outcome of such situation is addressed in the following example.

Suppose that a user activated a TCP application on the MN to download a file of size 1MB. Approximately half way through the download (500KB) the WLAN adapter, which was activated automatically due to the activation of the application, discovers WLAN coverage. Two different outcomes could take place depending on the user's position within the discovered WLAN's coverage area, the user's speed, and direction. If these mobility attributes result in the user remaining in the discovered WLAN's coverage area for a duration longer than the time required to complete the handoff procedures (e.g., ~13.06s according to Table 5-1), the application response time would decrease due to the higher data rates available through the WLAN. This is shown through Scenario (A) in Figure 5-11.

If the algorithm were applied in this case, then the processing associated with it would not interrupt the transfer of data over the current 3G/UMTS connection. Instead, the algorithm's main location-based evaluation would take place at the UTRAN (i.e., in the background) as user data continues to be transferred at the current 3G/UMTS data rate. However, until the value of L_a is made obsolete by using continuous user tracking or faster and more reliable position-acquisition methods (Section 5.2.2.3), the value of L_a will briefly delay the triggering of the handoff to a discovered WLAN. Scenario (B) in Figure 5-11 demonstrates the potential impact of the algorithm on delaying the triggering of the handoff, while it runs at the UTRAN, assuming the decision is to invoke the L3 handoff procedures.

On the other hand, if the user's mobility attributes result in a short visit to the discovered WLAN, then the MN would be forced to trigger the upward handoff back to 3G/UMTS before the application had time to benefit from the higher data rates in the WLAN. In such case, the data interruption during the downward handoff as well as the data interruption during the upward handoff increases the overall application response time. This is shown through Scenario (C) in Figure 5-11, where the handoff latency, during which data transfer is interrupted, is 13.060s for downward handoff and 6.98s for upward handoff. Therefore, assuming that no other delaying factors are involved, the decision to trigger the handoff to the discovered WLAN increased the application response time by approximately 20s. In other words, this handoff to 802.11 caused

degradation in performance, contrary to the objective of inter-networking 3G/UMTS and IEEE 802.11.

The proposed algorithm, as demonstrated earlier, is able to intelligently predict such situations, and thus prevent their occurrence by refraining from triggering the handoff to the discovered WLAN. Since the majority of the algorithm-related processing is done at the UTRAN, and minimal control signals are needed, the algorithm does not hinder the transfer of data over the current 3G/UMTS as it evaluates the user's position, speed, and direction and queries the WLAN coverage database to make a decision. Therefore, the outcome in the example described above, would be as shown through Scenario (D) in Figure 5-11. Note that the algorithm's latency (L_a) does not affect the current performance of the application(s) running on the MN. Instead, the algorithm runs in the background and decides that the handoff for the user in the example was not suitable. Therefore, the application continues download the remaining 500KB over the 3G/UMTS connection without any interruptions.

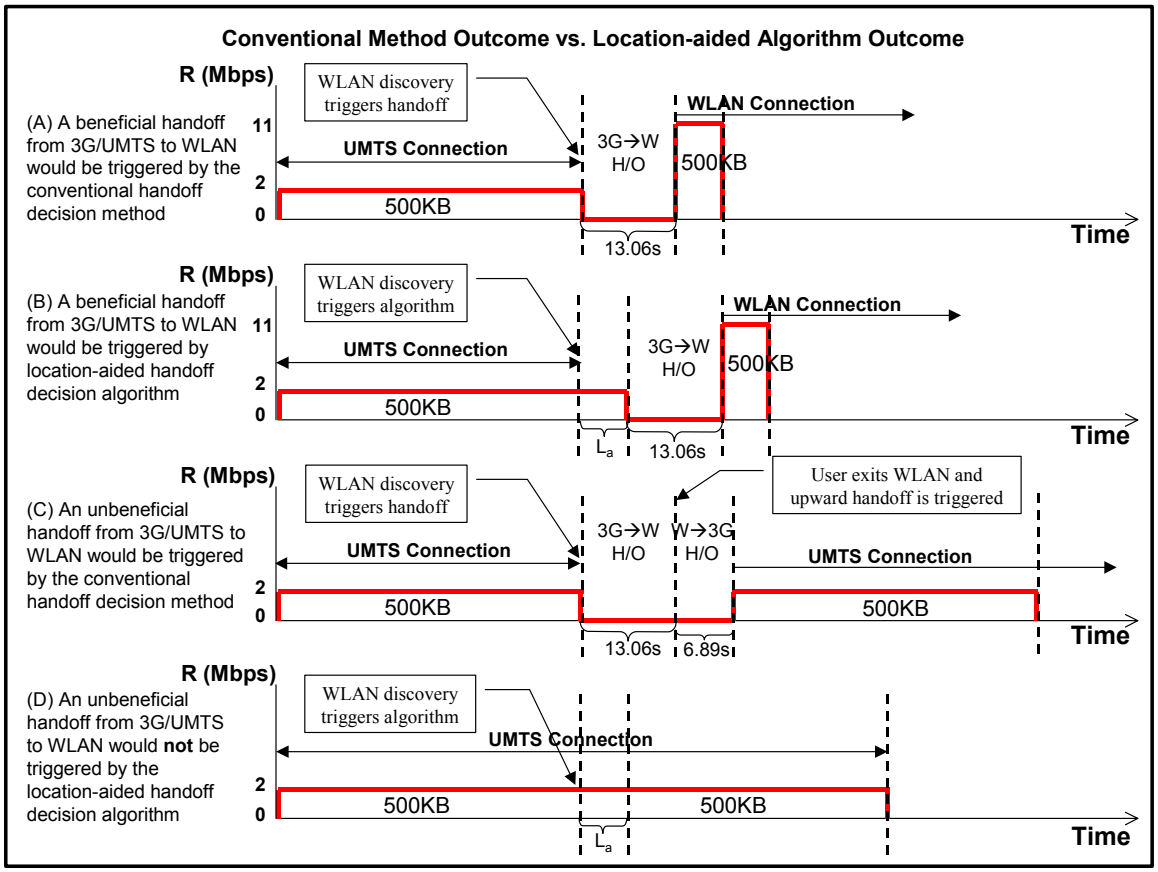

Figure 5-11. The outcome of discovering a foreign WLAN network during a 1MB file download over a 3G/UMTS connection in the user's home 3G/UMTS network

5.4 Suitable Environments

According to the evaluation of the algorithm's performance, capabilities, limitations, and benefits in the earlier sections of this chapter, the algorithm would perform best in areas which posses one or more of the following attributes:

- Outdoor WLAN hotspots (parks, city blocks) since exiting outdoor coverage areas is not restricted by physical boundaries as with indoor hotspots (Figure 5-12)
- High user density (urban areas, convention centers, university buildings) where the likelihood of scenarios similar to those in Section 5.2.1 is high
- Fast pace pedestrian mobility (airports, metro stations, urban areas, malls)
- High density of WLANs with overlapped coverage areas (as in Scenario-2)

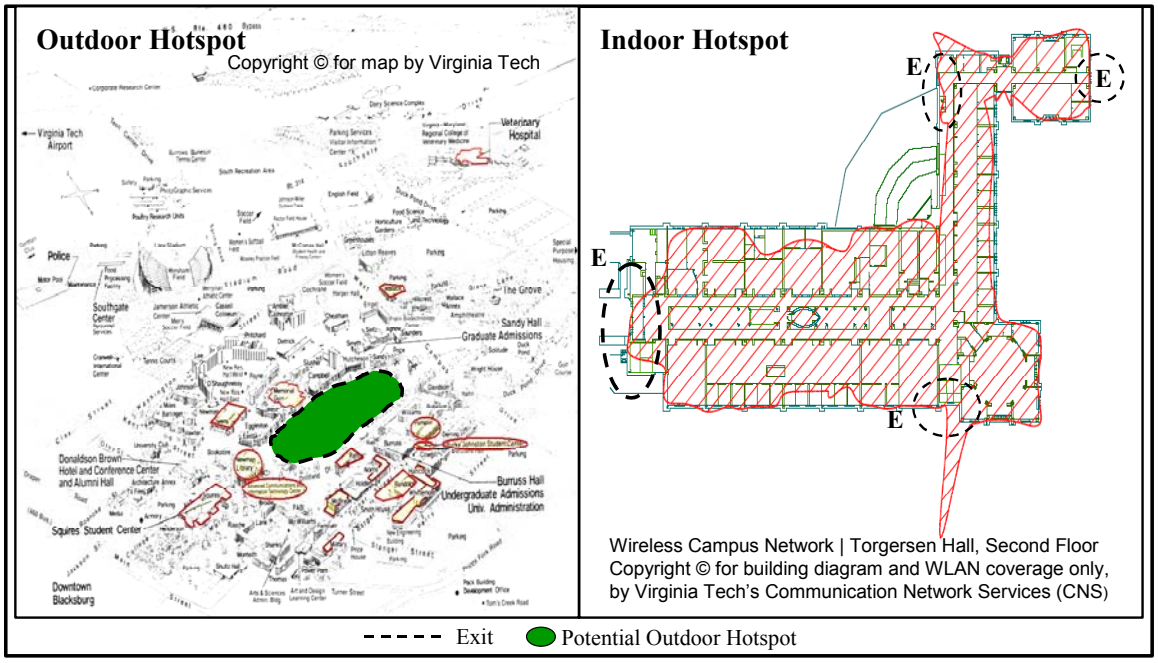

Figure 5-12. Entrance/Exit of outdoor vs. indoor WLAN hotspots

5.5 Comparison Summary

Table 5-4 summarizes the differences between the conventional handoff decision method and the proposed location-aided handoff decision algorithm with in the context of inter-networked heterogeneous wireless overlay networks.

Table 5-4. Comparison of Conventional vs. Location-aided handoff decision algorithm

Conventional Method	Location-aided Algorithm
Relies on RF attributes for handoff decision (High RSS → Initiate Handoff)	Relies on RF attributes to trigger the algorithm (High RSS →notify UTRAN)
Short delay between discovery and handoff initiation	Longer delay between discovery and handoff initiation due to position acquisition and database query lookup time
Unaware of the user's position & velocity	Aware of the user's position & velocity
Unaware of the extent of WLAN coverage area	Aware of the spatial characteristics and properties of the discovered WLAN due to the stored WLAN in the database
Access Point (AP)-Oriented	APs belonging to the same WLAN are treated as a single hotspot (i.e., blanket coverage)
Less efficient when users are moving at a non-negligible speed	Most beneficial in areas where users are traveling at a non-negligible speed
Sufficient if user are stationary or moving at negligible speed	Less useful if a user remains stationary (thus algorithm triggers handoff then terminates upon such discovery)
Prone to problems that waste resources and degrade performance (i.e., longer application response time due to decreased application-level throughput)	Intelligently predict handoff situations that would cause performance degradation due to the user's short visit to WLAN, thus preventing handoff in such situations
Immediately initiates a handoff to any AP that has a high beacon RSS, thus the probability of wasting a potentially beneficial handoff is very low	May delay or block beneficial handoffs due to the "Physical Boundary" limitation
Does not require the addition of any hardware or software to the infrastructure or handset	The main addition associated with the proposed algorithm is the WLAN coverage database described in Chapter 4
The handoff from UMTS to WLAN should only include the exchange of messages between the MN and network during: ▪ AP or Node-B attachment Request/Reply ▪ Mobile IP Registration Request/Reply	The handoff from UMTS to WLAN introduces at least 2-3 new messages that are required to deliver notifications between the UTRAN and the MN to: ▪ Trigger the location-based evaluation ▪ Abort the location-based evaluation ▪ Trigger the "Execution Stage" for 3G/UMTS to WLAN handoff

5.6 Conclusion

This chapter provided a demonstration of the differences between the capabilities of the conventional handoff decision method and the proposed location-aided handoff decision algorithm. Two primary scenarios were used in Section 5.2.1 to demonstrate

such differences. The limitations of the conventional handoff decision method were shown to stem from its lack of awareness of the user's speed and direction. Therefore, the scenarios highlighted the algorithm's ability to take advantage of these parameters and also utilize a WLAN coverage database concept for making better handoff decisions than the conventional method. The scenarios were chosen because they highlighted the algorithm's ability to intelligently predict that the user will not remain in the discovered coverage area long enough to reap any benefit (i.e., sufficiently take advantage of the higher data rates). Therefore, in both scenarios the algorithm's decision was to not trigger the handoff, contrary to the conventional handoff decision, which was to trigger the handoff in both cases.

After demonstrating how the algorithm's functionalities and capabilities allowed it to reach a different decision than the conventional handoff decision method, the benefit of such capability was addressed in Section 5.3. The benefit of the algorithm, in terms of preventing the triggering of handoffs associated with a short visits to the WLAN coverage area, were analyzed in terms of application-level throughput and application response time (as performance metrics). Handoff procedures, particularly for vertical handoff events, have significant delays and thus cause significant interruption in data transfer, as described in Chapter 3. The values provided in Tables 5-1 to 5-3, aimed to provide examples from the literature to emphasize the extent of the delay and data interruption that could be experienced during handoff.

The algorithm was shown to prevent an unnecessary and unbeneficial handoff to a WLAN, which would have resulted in the urgent need for another handoff as the user rapidly exits the coverage area. Therefore, it was found to prevent the data interruption that would have been caused by a downward handoff followed by either a L3 horizontal handoff to another WLAN (if one existed) or an upward handoff back to the 3G/UMTS network. According to the values presented for the handoff procedures, the total interruption could be as high as twenty seconds, which would either be experienced as one long interruption or a few consecutive interruptions. This depended on how quickly the user exits the newly discovered WLAN coverage area after the L3 handoff procedures were invoked.

Any vertical handoff is expected to cause some interruptions in data transfer, unless other measures are taken. The current measures described in the literature required inefficient use of bandwidth and other resources. Even in such cases, the handoff is still expected to cause some interruption. As described previously, handoff to a WLAN is desirable as long as the visit period is sufficient to take advantage of the higher data rates and compensate for the interruption in data transfer. However, if interruptions (or consecutive interruptions) take place and are not followed by an opportunity to utilize the higher data rates (Section 5.2.1) the application-level throughput becomes significantly throttled. This would ultimately result in an increase in application response time, which could be noticed by the user, depending on the length and frequency of interruptions. Therefore, by preventing such handoff events, the algorithm prevents degradation in performance.

In addition to the benefit from the user's perspective, preventing these inefficient handoff events and their associated interruptions would allow the current network to complete the task (e.g., file transfer) quicker and thus free up more bandwidth for other users' tasks. Moreover, by preventing these inefficient handoff events, the algorithm prevents the wasteful use of bandwidth and transmission power for exchanging unnecessary handoff control signals. It would also prevent inefficient loading of network resources such as DHCP servers, AAA servers, Mobile IP agents, etc.

Finally, in addition to demonstrating the algorithm's capabilities and benefits, this chapter addressed the algorithm's limitations and their potential outcome. The extent of these limitations' effect on the handoff decision accuracy and timing was also addressed in Section 5.3. In all three limitations' cases, unless the absolute worst-case scenario takes place, the algorithm has the ability to correct itself. In other words, since the algorithm's decision to refrain from triggering the handoff is not final, the outcome caused by these limitations should not persist.

Chapter 6: Conclusion

6.1 Thesis Summary

The internetworking of heterogeneous wireless overlay networks such as 3G/UMTS and WLAN/IEEE 802.11 has been the focus of a great deal of research efforts due to the variety of emerging applications that require different levels of QoS. Furthermore, the complementing nature of these technologies makes them suitable for integration to enhance the overall capacity of the 3G/UMTS network and provide better performance to users via the higher WLAN data rates. In an environment where such networks are loosely coupled and their coverage areas overlap, a dual-mode MN is expected to handoff between such networks depending on WLAN coverage availability. The conventional method for discovering WLAN coverage and making the handoff decision is based primarily on RF measurements such as Received Signal Strength (RSS). This approach has been the source of a number of problems (e.g., ping pong effect, near-far effect, etc.) that have been addressed in many publications. This research effort addressed another handoff-related issue that stems from relying entirely on an RF-based method for making the handoff decision between heterogeneous networks.

A number of factors were expected to contribute to the increase in the rate of handoff between heterogeneous networks in the observable future. The increase in the number of wireless users required that wireless cells become smaller in order to maintain an acceptable level of QoS. Furthermore, the increase in the number of WLAN deployed in areas of higher user mobility (e.g., airports, city blocks, parks, malls, campuses, etc.) were expected to result in users spending only a short period of time in any one WLAN's coverage area. In such cases, the RF-based handoff decision method was expected to trigger a downward handoff upon discovering WLAN coverage then trigger a handoff to another WLAN or back to 3G/UMTS upon exiting the current coverage area. The occurrence of such handoff events depends on the size, shape, and number of WLAN cells in a given area as well as each user's speed and direction of travel in such area.

The nature of handoff, particularly between heterogeneous networks that typically belong to different administrative domains, was found to cause substantial interruptions in data transfer. However, handoff to WLAN was expected to compensate for such interruptions by completing the interrupted task at a significantly higher data rate. However, due to the expected increase in short visits to a WLAN's coverage area, the user could exit the discovered coverage area before the completion of the handoff procedures and the transfer of sufficient data to compensate for the interruption. In such cases, the interruptions caused by the handoff to and from the discovered WLAN merely caused degradation in performance and wasted resources. Recall that the handoff from a 3G/UMTS network to a WLAN is mainly triggered for performance optimization purposes, not out of necessity. Therefore, performance degradation caused by one or more inefficient and unnecessary handoffs would essentially defeat the original purpose for internetworking 3G/UMTS networks with WLANs.

The technologies involved in this research effort (i.e., IEEE 802.11, UMTS, etc.) were studied during the initial research period in order to determine the capabilities and limitations of such technologies, as described in Chapter 2. Upon establishing a better understanding of these technologies' specifications, a clearer view of the problem and the potential solution was achieved. The RF-based conventional handoff decision method was sufficient for low user mobility and relatively large WLAN cells (e.g., corporate use). However, it was clearly lacking the means to accurately assess the available WLAN coverage for users traveling through areas characterized by small cells and higher user mobility, as addressed in Chapter 3. This was mainly due to the conventional method's lack of awareness of the user's position, speed, and direction. Therefore, there was a need for a solution that took into consideration the user's position and velocity, the handoff delays between the inter-networked 3G-WLAN pair, and the spatial properties/attributes of the discovered WLAN.

An algorithm was developed to support an environment where short visits to WLAN coverage areas were likely to occur and the handoff rate was high. The algorithm was not designed to entirely replace the current RF-based method. Instead, it utilized the RF-based measurements to discover coverage and trigger the proposed evaluation stage, rather than triggering the handoff procedures automatically. As described in Chapter 4,

the majority of the location-based evaluation ran at the UTRAN and took into account the user's position, speed, and direction. These parameters were obtained via one of the positioning methods that were included in the 3GPP standard for Location-based Services (LCS) such as A-GPS or U-TDOA. The architecture introduced by 3GPP to support LCS consisted of specific nodes for supporting the calculation of the user's position, speed, and direction as described in Chapter 2.

The algorithm also took into account the values of the average delays associated with a handoff between the inter-networked 3G/UMTS and WLAN pair. These delays were used to compute the minimum period that a user must remain in the coverage area of the discovered WLAN in order for the application to benefit from the higher data rates. The value obtained from this calculation as well as the user's average speed and direction were used to estimate the user's potential Trajectory End Point (TEP) coordinates.

The TEP coordinates along with the discovered WLAN's ID and the UMTS cell's ID were used to query a WLAN coverage database. The database used a mechanism to map the TEP to a particular grid block or entry. If the TEP was found to be within the boundaries of the discovered WLAN's coverage area, then the UTRAN sent a notification to the MN to trigger the handoff to the discovered WLAN. Otherwise, the location-based evaluation was designed to continue iterating until the TEP coordinates fell within the boundaries of the discovered WLAN's coverage area, or the user exited the coverage area and notifies the UTRAN to abort the evaluation.

As described in Chapter 4, the location-based evaluation running at the UTRAN was the same regardless of whether the user was transferring data over a 3G/UMTS connection or another WLAN connection. In either case, the MN communicated with the 3G/UMTS network over the control plane (C-plane) to perform the location-based evaluation. On the other hand, the procedures running at the MN were different depending on whether the MN was transferring data over a 3G/UMTS connection or another WLAN's connection when discovering new WLAN coverage.

Two primary scenarios were provided in Chapter 5 to demonstrate the capabilities and performance of the algorithm in cases where the MN was connected to a 3G/UMTS network or when the MN was connected to another WLAN, upon discovering new WLAN coverage. The algorithm does not provide any improvements over the RF-based

handoff decision method in cases where the decision is to trigger the handoff. Therefore, the two primary scenarios consisted of situations that would produce contrary decisions using the conventional handoff method and the proposed location-aided algorithm. In both scenarios the RF-based decision was to trigger the handoff to the discovered WLAN, whereas the proposed location-aided algorithm's decision was to refrain from triggering the handoff. The logic and functionality of the algorithm were demonstrated and discussed using these scenarios in Chapter 5. In both cases, the algorithm's ability to take into consideration the user's position, speed, and direction as well as the handoff delay values allowed it to predict that the user's visit to the discovered WLAN was not long enough to accommodate the handoff delays, the stabilization period, and the transfer of a sufficient amount of data.

6.2 Conclusion

The algorithm's ability to predict short visits to a WLAN's coverage area, prior to triggering handoff, allowed it to provide an observable benefit from the individual user's perspective as well as the network's perspective. Two metrics were used in Chapter 5 to address the benefit of the algorithm and its contribution in improving the handoff decision process in heterogeneous wireless overlay networks. The algorithm's contributions were mainly observed in terms of application-level throughput and application response time. Chapter 5 provided a demonstration of the impact of consecutive handoffs, which result from short visits to a discovered WLAN. These short visits caused interruptions in data transfer that could have lasted for a period as long as twenty seconds. These interruptions were shown to significantly throttle the application-level throughput and cause an increase in application-response time. Depending on the type of application, the length of the interruption period, and the frequency of interruptions, the increase in application response time could become noticeable by the user. Therefore, the algorithm's ability to predict short visits allowed it to prevent the occurrence of handoffs that would result in such performance degradation.

The algorithm was designed to perform its location-based evaluation in the background (i.e., at the UTRAN) while the MN continued to transfer data over its current connection and without interruptions. Only when the user's visit to a WLAN's coverage

area was foreseen to be long enough to result in benefit does the algorithm trigger the handoff to such WLAN. This allowed the task, for which the data transfer was initiated, to complete quicker than if the short visit to the discovered WLAN's coverage area had triggered the handoff (i.e., using the conventional method). The quicker completion of such task resulted in a quicker release of the channels that were used to transfer data, thus it allowed the network to offer more bandwidth to complete other users' task sooner. In addition, the algorithm's ability to prevent the triggering of handoffs associated with short visits avoided the unnecessary loading of the servers used to support mobility (i.e., DHCP servers, AAA servers, Mobile IP servers, etc.). Moreover, it prevented the unnecessary consumption of bandwidth and transmission-power associated with exchanging handoff control messages.

In addition to the algorithm's benefits, Chapter 5 also addressed situations where the algorithm's limitations could result in undesirable or inaccurate handoff decisions. The absolute worst-case scenarios were shown in Chapter 5 in order to demonstrate the worst possible impact of such limitations. However, unless the absolute worst-case scenario takes place, the algorithm's limitations were not expected to have a significant impact on the outcome. In fact, unless the absolute worst-case scenario occurs, the algorithm's mechanisms were designed to recover from errors and take self-correcting measures. Moreover, one algorithm limitation, which depended mainly on the latency of the algorithm itself, was expected to become obsolete as a result of advancements in user-tracking technologies and location-based services in the observable future.

Finally, the objective of this research effort was satisfied as demonstrated and addressed throughout this document. An extensive analysis of the problem, which highlighted the need for a location-aware handoff decision approach, was conducted during the course of this study. Based on such analysis, a location-aided handoff decision algorithm was developed to make better handoff decisions in heterogeneous environments characterized by higher user density, smaller cells, and higher user mobility, which result in shorter user visits to WLAN coverage areas (i.e., a higher rate of unnecessary handoffs). The proposed algorithm was shown to meet the underlying objective, which was to reduce or eliminate the triggering of handoffs that result in performance degradation and wasting of resources. By applying the proposed algorithm,

the only handoffs allowed to trigger are the ones that have the potential to improve performance, which supports the original objective of internetworking heterogeneous technologies such as 3G/UMTS and IEEE 802.11 networks.

6.3 Future Work

The concept of using a WLAN coverage database requires a tradeoff between granularity and access time. Future work is needed to further study the tradeoff and the balance that would produce the best results in terms of making accurate decisions within an acceptable amount of time.

Also, further research is needed to determine the best method for assessing the value of the "mean handoff latency" associated with completing the handoff procedures between an inter-networked 3G/UMTS-WLAN pair. This value was needed in the calculation of the minimum required visit duration (RVD), which was used in the computation of the TEP coordinates. Therefore, increasing the accuracy of the RVD value would result in an increase in the accuracy of the location-aided handoff decision.

In most location-based services, including the algorithm/service described in this research effort, there is a need to obtain and store (at least temporarily) the user's position for processing at the network. Furthermore, such information as well as other location-based information, which is typically stored in a database, must be shared among different network access providers/operators in order to support location-based services on a large scale. This requires strict security measures in order to ensure the user's privacy and minimize the risk associated with user tracking. Therefore, future work is needed to establish the requirements and measures needed to support location-based services.

Finally, this research effort focused on mitigating a handoff-related issue by using location-based information. This is only one aspect of the benefits that could results from utilizing location-based information in the wireless domain. As the accuracy, reliability, and availability of location-based information increases with the increase in user demand, such information can be used to improve other aspect of wireless communication. For instance, future work is needed to study the potential and benefit of using location-based information to improve coverage discovery in homogeneous/heterogeneous networks,

heterogeneous network load-balancing, and heterogeneous network/cell design and planning.

References

[1] B. Walke, P. Seidenberg, M.P Althoff, UMTS, The Fundamentals, John Wiley & Sons, Ltd, 2003.

[2] ANSI/IEEE Std 802.11, "IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", 1999 Edition.

[3] 3GPP TS 23.060: "General Packet Radio Services (GPRS); Service Description", Release 1999, Version 3.13.0, November 2002.

[4] Tero Ojanpera, Ramjee Prasad, "An overview of Third-Generation Wireless Personal Communication: A European Perspective", IEEE Personal Communications Magazine, Volume 5, no. 6, pp. 59-65, December 1998.

[5] 3GPP TR 22.934: "Feasibility Study on 3GPP system to WLAN interworkking", Version 6.2.0, September 2003.

[6] M. Cappiello, A. Floris, L. Veltri, "Mobility amongst Heterogeneous Networks with AAA Support", ICC. 2002 IEEE International Conference on Communications, Vol. 4, pp. 2064-2069, April-May 2002.

[7] 3GPP TS 25.321: "Medium Access Control (MAC) Protocol Specification", Release 5, Version 5.6.0, September 2003.

[8] D. H. Stojanovic, S. J. Djordjevic-Kajan, "Developing location-based services from a GIS perspective", Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2001. TELSIKS 2001. 5th International Conference, Volume 219, pp. 459-, September 2001.

[9] 3GPP TS 22.071: "Location Services (LCS); Service Description; Stage 1", Version 6.5.0, Release 6, September 2003.

[10] 3GPP TS 23.271: "Functional Stage 2 Description of LCS", Version 6.5.0, Release 6, September 2003.

[11] 3GPP TS 24.030: "Location Services (LCS); Supplementary Service Operations; Stage 3", Version 5.1.0, Release 5, June 2002.

[12] ITU-T. Recommendation Q.1701: Framework for IMT-2000 networks.

[13] 3GPP TS 23.002: "Network Architecture", Version 6.1.0, Release 6, June 2003.

[14] Jonathan P. Castro, The UMTS Network and Radio Access Technology-Air

Interface Techniques for Future Mobile Systems, John Wiley & Sons, Ltd., 2001.

[15] 3GPP TR 23.922: “Architecture for an All IP Network”, Version 1.0.0, October 1999.

[16] 3GPP TS 23.060: “General Packet Radio Services (GPRS); Service Description”, Version 6.2.0, Release 6, September 2003.

[17] 3GPP TS 25.331: “RRC Protocol Specification”, Release 99, October 1999.

[18] 3GPP TS 25.435: “UTRAN Iub Interface User Plane Protocols for Common Transport Channel data streams”, Version 6.1.0, March 2004.

[19] ETSI TR 101.031, “HIPERLAN Type 2; Requirements and Architectures for Wireless Broadband Access”, Version 2.2.1, January 1999.

[20] 3GPP TS 29.234: “3GPP System to WLAN Internetworking”, Release 6, Version 1.1.0, November 2003.

[21] Angela Doufexi, Simon Armour, Araceli Molina, “Hotspot Wireless LANs to Enhance the Performance of 3G and Beyond Cellular Networks”, IEEE Communications Magazine, pp. 58-65, July 2003.

[22] IETF, “Generic AAA Architecture”, IETF RFC 2903, August 2000.

[23] P. R. Calhoun, “Diameter Base Protocol”, IETF RFC 3588, September 2003.

[24] C. Rigney, “Remote Authentication Dial In User Service”, IETF RFC 2865, June 2000.

[25] C. Perkins, ed., “IP Mobility Support for IPv4”, IETF RFC 3344, August 2002.

[26] Shiao-Li Tsao, Chia-Ching Lin, “Design and Evaluation of UMTS-WLAN Interworking Strategies”, Vehicular Technology Conference, Vol. 2, pp. 777-781, 2002.

[27] Eva Gustafsson, Annika Jonsson, “Always Best Connected”, IEEE Wireless Communications, pp. 49-55, February 2003.

[28] ISO, “OSI Routing Framework” ISO/TR 9575, 1989.

[29] 3GPP TS 25.305: “Stage 2 Functional Specifications of User Equipment (UE) positioning in UTRAN”, Version 5.7.0, Release 5, September 2003.

[30] Yilin Zhao, “Standardization of Mobile Phone Positioning for 3G Systems”, IEEE Communications Magazine, Vol. 40, No. 7, pp. 108-116, July 2002.

[31] R. Chakravorty, "Performance issues with vertical handovers - experiences from GPRS cellular and WLAN hot-spots integration", Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications, pp.14-17, March 2004.

[32] J. Makela, M. Ylianttila, K. Pahlavan, "Handoff decision in multi-service networks", The 11th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Volume: 1, pp. 655-659, September 2000.

[33] M. Ylianttila, J. Makela, K. Pahlavan, "Geolocation information and inter-technology handoff", Communications, 2000 IEEE International Conference, Volume: 3, pp. 1573-1577, June 2000.

[34] Dimitrios Makris, Tim Ellis, "Spatial and Probabilistic Modeling of Pedestrian Behavior", Information Engineering Center, City University, London, UK, 2002.

[35] P. Engelstad, T. Haslestad, F. Paint, "Authenticated access for IPv6 supported mobility", (ISCC 2003) Proceedings, Eighth IEEE International Symposium on Computers and Communication, Volume 1, pp. 569 – 575, 30 June-3 July 2003.

[36] Eun Kyoung Paik, Yanghee Choi, "Prediction-based fast handoff for mobile WLANs", ICT 2003. 10th International Conference on Telecommunications, Volume 1, pp. 748 – 753, 23 Feb.-1 March 2003.

[37] R. Caceres, L. Iftode, "Improving the performance of reliable transport protocols in mobile computing environments", IEEE Journal on Selected Areas in Communications, Volume 13 , Issue: 5 , pp. 850-857, June 1995.

[38] A.C. –F. Chan, D. H. K. Tsnag, S. Gupta, "Impacts of handoff on TCP performance in mobile wireless computing", 1997 IEEE International Conference on Personal Wireless Communications, pp.184 – 188, 17-19 December 1997.

[39] H. Kim, H. Afifi, "Improving mobile authentication with new AAA protocols", IEEE International Conference on Communications, Volume 1, pp. 497-501, 11-15 May 2003.

[40] M. Georgiades, N. Akhtar, C. Politis, R. Tafazolli, "AAA Context Transfer for Seamless and Secure Multimedia Services over All-IP Infrastructures", 5th European Wireless Conference (EW'04), Barcelona, Spain, February 24-27, 2004.

[41] P. R. Calhoun, "Diameter Base Protocol", IETF RFC 3588, September 2003.

[42] IETF, "Remote Authentication Dial In User Service (RADIUS)", IETF RFC

2865, June 2000.

[43] TruePosition, Inc., **"**Uplink Time Difference of Arrival (U-TDOA) Added to GSM Standards Standardization ensures seamless interoperability and easy integration for wireless mobile operators", Press Release, KING OF PRUSSIA, Pa., May 5, 2003.

[44] 3GPP TR 45.811: "Feasibility Study on Uplink TDOA in GSM and GPRS", Release 6, Version 6.0.0; June 2002.

[45] A. M. Bin Ahmad, M. D. Bin Baba, "Handover strategy for mobile wireless LAN", NCTT 2003 Proceedings. 4th National Conference on Telecommunication Technology, pp.141 – 143, 14-15 January 2003.

[46] Ming-Hsing Chiu; M. A. Bassiouni, "Predictive schemes for handoff prioritization in cellular networks based on mobile positioning"**,** Selected Areas in Communications, IEEE Journal on , Volume 18 , Issue 3 , pp. 510 – 522, March 2000.

[47] D.-S. Lee, Y.-H. Hsueh, " Bandwidth-Reservation Scheme Based on Road Information for Next-Generation Cellular Networks", IEEE Transactions on Vehicular Technology, Volume 53, Issue 1, pp. 243 – 252, Janueary 2004.

[48] R. L. Knoblauch, M.T. Pietrucha, M. Nitzburg, "Field Studies of Pedestrian Walking Speed and Start-Up Time," Transportation Research Record 1538: Pedestrian and Bicycle Research, Transportation, Research Board, National Research Council, Washington, DC, p. 27, 1996 (http://trb.org/).

[49] J. Vatn, "Long random wait times for getting a care-of address are a danger to mobile multimedia", (MoMuC '99) IEEE International Workshop on Mobile Multimedia Communications, pp. 142 – 144, 15-17 November 1999.

[50] Jon-Olov Vatn, Gerald Q. Maguire Jr., "The effect of using co-located care-of addresses on macro handover latency", 14th Nordic Teletraffic Seminar, Lygby, Denmark, August 1998.

[51] L. Garber, L., "Will 3G really be the next big wireless technology", Computer, Volume 35, Issue 1, pp. 26 – 32, January 2002.

[52] J. De Vriendt, J., P. Laine, C. Lerouge, Xu Xiaofeng, "Mobile network evolution: a revolution on the move", Communications Magazine, IEEE , Volume 40, Issue 4, pp. 104 – 111, April 2002.

[53] G. M. Koien, T. Haslestad, "Security aspects of 3G-WLAN interworking", Communications Magazine, IEEE, Volume 41, Issue 11, pp. 82 – 88, November 2003.

[54] K. Pahlavan, Li Xinrong, J. P. Makela, “Indoor geolocation science and technology”, Communications Magazine, IEEE, Volume 40 , Issue 2 , pp. 112–118, November 2003.

[55] E.D. Kaplan, Understanding GPS: Principle and Applications, Artech House, 1996.

[56] Badache, N.; Tandjaoui, D., “A seamless handoff protocol for hierarchical Mobile IPv4”, 4th International Workshop on Mobile and Wireless Communications Network, pp. 651 – 655, 9-11 September 2002.

[57] M. Jaseemuddin, “An architecture for integrating UMTS and 802.11 WLAN networks”, Eighth IEEE International Symposium on Computers and Communication, (ISCC 2003). Proceedings, Volume 2, pp. 716 – 723, 30 June-3 July 2003.

[58] A. Mohammad, A. Chen, “Seamless mobility requirements and mobility architectures”, Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE, Volume 3, pp. 1950 – 1956, 25-29 November 2001.

[59] Eunsoo Shim, Hung-yu Wei, Yusun Chang, R. D. Gitlin, “Low latency handoff for wireless IP QoS with NeighborCasting”, ICC 2002. IEEE International Conference on Communications, Volume 5, pp. 3245 – 3249, 28 April-2 May 2002.

[60] A. Stephane, A. Mihailovic, A. H. Aghvami, “Mechanisms and hierarchical topology for fast handover in wireless IP networks”, Communications Magazine, IEEE, Volume 38, Issue 11, pp. 112 – 115, November 2000.

[61] IETF, “Dynamic Host Configuration Protocol”, IETF RFC 2131, March 1997.

[62] Guangbin Fan, I. Stojmenovic, Jingyuan Zhang, “A triple layer location management strategy for wireless cellular networks”, Eleventh International Conference on Computer Communications and Networks, pp. 489 – 492, 14-16 October 2002.

[63] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. P. Makela, R. Pichna, J. Vallstron, ”Handoff in hybrid mobile data networks”, Personal Communications, IEEE [see also IEEE Wireless Communications], Volume 7, Issue 2, pp. 34 – 47, April 2000.

[64] M. Ylianttila, M. Pande, J. Makela, P. Mahonen, “Optimization scheme for mobile users performing vertical handoffs between IEEE 802.11 and GPRS/EDGE networks”, Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE, Volume 6, pp. 3439 – 3443, 25-29 November 2001.

[65] D. Findlay, H. Flygare, R. Hancock, T. Haslestad, E. Hepworth, D. Higgins, S. McCann, "3G interworking with wireless LANs", Third International Conference on 3G Mobile Communication Technologies (Conf. Publ. No. 489), pp. 394 – 399, 8-10 May 2002.

[66] Keith C. Clarke, Getting Stared with Geographic Information Systems, Second Edition, Printice Hall, 1999.

[67] S. Rogers, "Creating and evaluating highly accurate maps with probe vehicles", Intelligent Transportation Systems, IEEE, pp. 125 – 130, 1-3 October 2000.

[68] Rong-Hong Jan; Yung Rong Lee, "An indoor geolocation system for wireless LANs", International Conference on Parallel Processing Workshops, pp. 29 – 34, 6-9 October 2003.

[69] 3GPP TS 25.301: "Radio Interface Protocol Architecture", Release 5, Version 5.2.0, September 2002.

[70] Ekahau's Site Survey™ 2.0 (www.ekahau.com)

Vita

Areej Saleh was born and raised in Kuwait. She received a scholarship after high school to pursue an Engineering degree in the United States. Areej relocated to Blacksburg, Virginia in 1996 where she chose to pursue her degree at Virginia Tech's Bradley Department of Electrical and Computer Engineering. She completed her B.S. degree in Computer Engineering and a minor in Computer Science in 2001. After graduating, she relocated to Orlando, Florida to work at a Sun Microsystems reseller and training center, and was training to become a Sun Instructor. She returned to Virginia Tech to pursue a Master's degree in Computer Engineering with a specific interest in Network Engineering. Her academic interests are in wireless technology and the integration of heterogeneous wireless networks. Her career interests also lie in network security, particularly on a strategic and design level. In July of 2004 Areej will start work at Ernst & Young's Security Architecture Center in Baltimore, Maryland. Her personal interests include distance running, horseback riding, and Jazz music.