Architecture and Cross-Layer Mobility Management Protocols for Next-Generation Wireless Systems

A Thesis
Presented to
The Academic Faculty

by

Shantidev Mohanty

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2005

Architecture and Cross-Layer Mobility Management Protocols for Next-Generation Wireless Systems

Approved by:

Professor Ian F. Akyildiz, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor George Riley
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Gordon L. Stüber
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Mostafa H. Ammar
College of Computing
Georgia Institute of Technology

Professor Raghupathy Sivakumar
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: November 21, 2005

To my friends, well wishers,

teachers, and parents-

Haripriya Mohanty and Narayan Chandra Mohanty.

ACKNOWLEDGEMENTS

Many people have given me invaluable inspiration, guidance, and support and have been influential in my career. First and foremost, I would like to express my sincere thanks to my dissertation advisor, Dr. Ian F. Akyildiz, for his continuous support and guidance throughout my doctoral study. This work would not have been possible without all the insightful discussions with him. Not only did Dr. Akyildiz lead me into the world of networking research, he has also taught me in many other ways that could lead me to success in my future career.

I would like to acknowledge Dr. Gordon L. Stüber, Dr. Raghupathy Sivakumar, and Dr. George Riley for being on my dissertation proposal committee and defense committee. I would also like to thank Dr. Mostafa H. Ammar for serving on my dissertation defense committee. Their invaluable comments and enlightening suggestions have helped improve the quality for this dissertation.

I would like to extend my sincere gratitude towards Dr. Jiang Xie for her patience and support during my dissertation process. I would also like to thank the present and past members of the Broadband and Wireless Networking (BWN) Laboratory. Special thanks go to Chao Chen, Ozgur Baris Akan, Jian Fang, and Mehmet C. Vuran. Their friendship and assistance made my last four years an enjoyable experience.

I would also like to thank my parents for their love, inspiration, guidance, and encouragement. My friends provide me with everything that is required to reach my dreams. They have been there always sharing my happiness and extending their hands when going was not easy. I am deeply thankful to my friends without whom I can not imagine myself. I am blessed to extend them my sincere and heartily gratitude for providing me with selfless motivation, courage, confidence, and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AAA	Authentication, Authorization, and Accounting
AIMD	Additive Increase Multiplicative Decrease
AMC	Architecture for ubiquitous Mobile Communications
AMMF	Adaptive Multi-layer Mobility management Framework
APME	Access Point Management Entity
AR	Access Router
AU	Authentication Unit
AuC	Authentication Center
BLR	Boundary Location Register
BS	Base Station
CARD	Candidate Access Router Discovery
CDR	Call Detail Records
CDMA	Code Division Multiple Access
CH	Correspondent Host
CMP	Cross-layer (Layer 2 + 3) handoff Management Protocol
CN	Correspondent Node
CoA	Care-of-Addresses
DECT	Digital Enhanced Cordless Telephone
DHCP	Dynamic Host Configuration Protocol
EAP	Extensible Authentication Protocol
EAPOL	EAP over LAN
FA	Foreign Agent
FN	Foreign Network
FEC	Forward Error Correction

FER	Frame Error Rate
GEO	Geostationary Earth Orbit
GFA	Gateway Foreign Agent
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Services
GRX	GPRS Roaming eXchange
GS	Gateway Station
GSM	Global System for Mobile communications
HA	Home Agent
HMIP	Hierarchical Mobile IP
HN	Home Network
IAPP	Inter-Access Point Protocol
IG	Interworking Gateway
IP	Internet Protocol
ISHO	Inter-System Handover
LCR	Level Crossing Rate
LEO	Low Earth Orbit
MAC	Message Authentication Code
MAHO	Mobile Assisted network controlled Handoff
MAP	Mobile Application Part
MH	Mobile Host
MIP	Mobile IP
MN	Mobile Node
MT	Mobile Terminal
NAHO	Network Assisted mobile controlled Handoff
NAI	Network Access Identifier
NBS	New Base Station
NGWS	Next-Generation Wireless Systems

NIA	Network Inter-operating Agent
OBS	Old Base Station
PCF	Packet Control Function
PDSN	Packet Data Serving Node
PLMN	Public Land Mobile Network
PSD	Power Spectral Density
QoS	Quality of Service
RAN	Radio Access Network
RLP	Radio Link Protocol
RSS	Received Signal Strength
RTO	Retransmission Timeout
RTP	Real-time Transport Protocol
RTT	Round-Trip Time
SCTP	Stream Control Transmission Protocol
SIM	Subscriber Identity Module
SIP	Session Initiation Protocol
SLA	Service Level Agreement
TDMA	Time Division Multiple Access
UMP	User Mobility Profile
UMTS	Universal Mobile Telecommunication System
VGSN	Virtual GPRS Support Node
WLAN	Wireless Local Area Network

SUMMARY

As a result of rapid progress in research and development, today's wireless world exhibits several heterogeneous communication networks, such as cellular networks, satellite networks, wireless local area networks (WLAN), mobile ad hoc networks (MANET), and sensor networks. These networks are complementary to each other. Hence, their integration can realize a unified wireless system that has the best features of the individual networks. This has spurred much research interest in designing integrated next-generation of wireless systems (NGWS).

While existing wireless networks have been extensively studied individually, the integrated wireless system brings new challenges in architecture design, system management, and protocol design. The different wireless networks use different communication technologies and are based on different networking paradigms. Therefore, it is challenging to integrate these networks such that their heterogeneities are hidden from each other and a harmonious inter-operation among them is achieved. The objective of this research is to design a scalable, secure, and robust architecture and to develop seamless mobility management protocols for NGWS.

More specifically, an architecture that integrates the heterogeneous wireless systems is first proposed for NGWS. Next, a cross-layer (Layer 2 + 3) handoff management protocol is developed for NGWS. Afterward, analytical modeling is developed to investigate the handoff performance of the existing mobility management protocols for different types of applications. Finally, a framework for multi-layer mobility management is developed to support the seamless handoff support to all types of applications in NGWS.

CHAPTER I

INTRODUCTION

1.1 Next-Generation Wireless Systems

During the past few years, advances in mobile communication theory have enabled the development and deployment of different wireless access technologies. Alongside the revolutionary progress in wireless access technologies, advances in wireless access devices (such as laptops, palmtops, and cell phones) and mobile middleware have paved the way for the delivery of beyond-voice-type services while on the move. This sets the platform for high-speed mobile communications that provide high-speed data and both real and non-real time multimedia to mobile users. Today's wireless world uses several communication infrastructures such as Bluetooth for personal area, IEEE 802.11 for local area, Universal Mobile Telecommunication System (UMTS) for wide area, and satellite networks for global networking. These networks are designed independently for some specific service needs of mobile users and vary widely in terms of their service parameters [74], as summarized below:

- **Data Rate**: The satellite and cellular networks can deliver a maximum data rate of 2 Mbps. On the other hand, the local area and personal area networks such as IEEE 802.11 can support data rates in excess of 100 Mbps.

- **Access Delay**: While the one-way access delay in the wireless link may not be significant in a Wireless Local Area Network (WLAN), a typical round-trip time (RTT) varies between a few hundred milliseconds and one second in 3G links because of the extensive physical-layer processing, e.g., forward error correction (FEC), interleaving, and transmission delays [45]. The access delay is much

higher in satellite links, which have high propagation delay up to 270 ms [10].

- **Coverage Area**: The satellite networks and cellular networks can provide global and wide area coverage, respectively. However, 802.11 and other local area networks have only limited coverage.

Therefore, none of the existing wireless networks can simultaneously satisfy the high data rate, low latency, and ubiquitous coverage needs of the mobile users' service demands [74]. On the other hand, since these wireless networks are complementary to each other [12], their integration and coordinated operation can provide ubiquitous "always best connection" [34] quality mobile communications to the users. Figure 1 shows an example architecture of an integrated wireless system that consists of a UMTS/3G network, a satellite network, and a WLAN. It may be noted that other networks such as Bluetooth, Home RF, and sensor networks can also be included in Figure 1. In this architecture, mobile users use multi-mode terminals that are equipped with multiple air interfaces and adaptive protocols so that the same terminal can be used for different networks. Using these terminals, mobile users are always connected to the best available network or networks. For example, when users reside inside WLAN coverage areas such as WLANs available in offices, airports, and shopping complexes, they communicate using the WLANs. On the other hand, when away from a WLAN network, for example on highways, they use the available UMTS/3G networks. If neither a WLAN nor UMTS/3G is available, then they use satellite networks. When users move out of the coverage of the serving network, their terminals automatically switch to another network such that the applications do not experience connection interruption. Therefore, users perceive different wireless networks as a single integrated system. We refer to this integrated system as the next-generation wireless systems (NGWS).

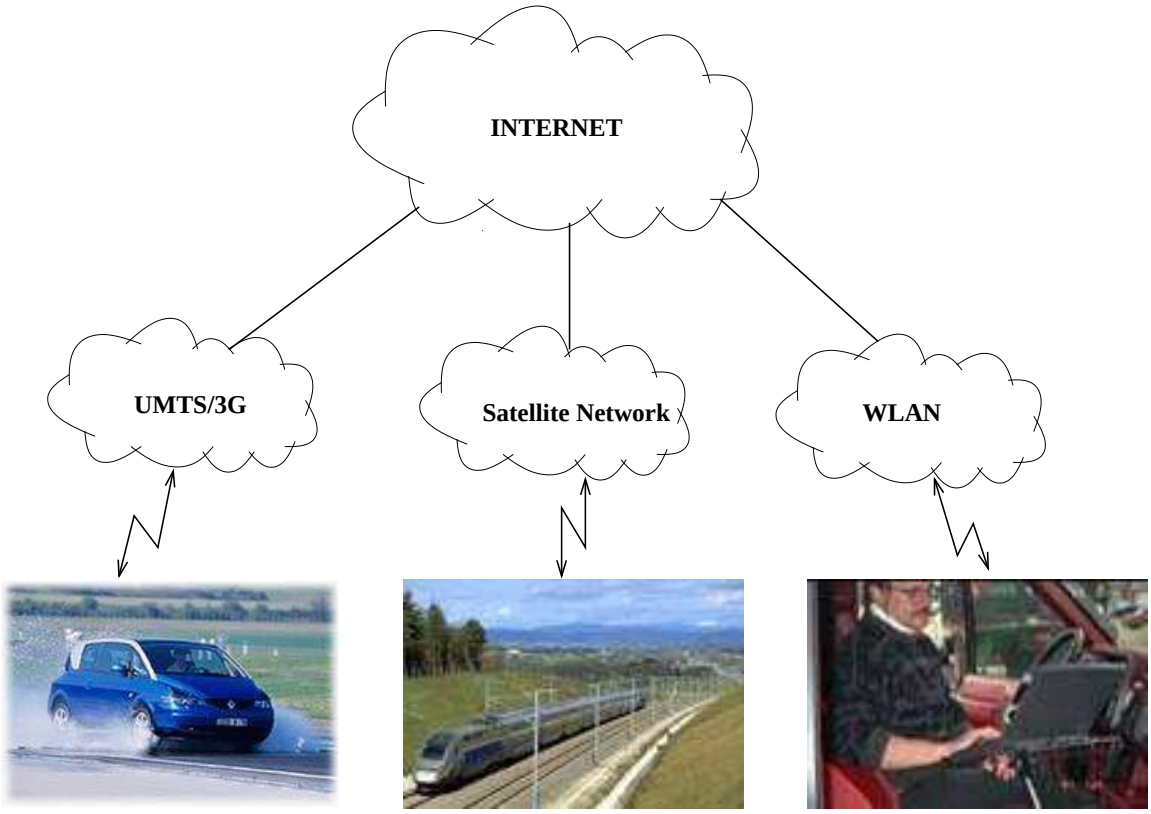


Figure 1: An example architecture for integrated wireless systems.

The design of NGWS is challenging because of the following inherent heterogeneities of different wireless networks:

- **Access Technologies**: Different networks use different technologies for radio access, e.g., General Packet Radio Services (GPRS) use time division multiple access (TDMA), cdma2000 and UMTS use code division multiple access (CDMA), and WLANs use random access schemes such as carrier sense multiple access with collision avoidance (CSMA/CA).

- **Network Protocols**: Different networks use different protocols for transport, routing, mobility management, authentication, billing, etc.

- **Service Providers**: These networks belong to different service providers who may not have direct service level agreements among them.

Therefore, innovative techniques are required to integrate these networks such that their heterogeneities are hidden from each other and a harmonious inter-operation among them is achieved. This necessitates a new direction in the design of NGWS architecture.

Once a suitable architecture is designed for NGWS, the next challenge is to support seamless mobility management. In NGWS, two types of mobility scenarios arise: horizontal handoff (i.e., intra-system handoff) and vertical handoff (i.e., inter-system handoff) [12], as shown in Figure 2. A mobile user's movement between two base stations (BSs) of the same system (e.g., the movement from BS10 to BS11 in Figure 2) is known as a horizontal handoff and that between the BSs of two different systems (e.g., the movement between BS12 to BS20 in Figure 2) is known as a vertical handoff. It is essential that applications running on the mobile terminals remain unaware of a user's movement, both horizontal and vertical, to ensure uninterrupted services with minimum quality of service (QoS) degradation. This can be achieved by reducing the handoff failure probability and by restricting the handoff latency and packet loss during handoffs to the values that are tolerable by the applications. This requires the design of efficient mobility management protocols for NGWS.

1.2 Research Objectives and Solutions

In this research, a new architecture is proposed to integrate the heterogeneous wireless systems to realize a scalable architecture for NGWS. To support efficient mobility management in NGWS, analytical models are developed to investigate the performance of the existing mobility management protocols for different types of application. Based on the results of this analysis, application adaptive mobility management protocols are developed. Moreover, cross-layer techniques are proposed to further enhance the handoff performance of application adaptive mobility management. Specifically, the following four areas are investigated under this research:

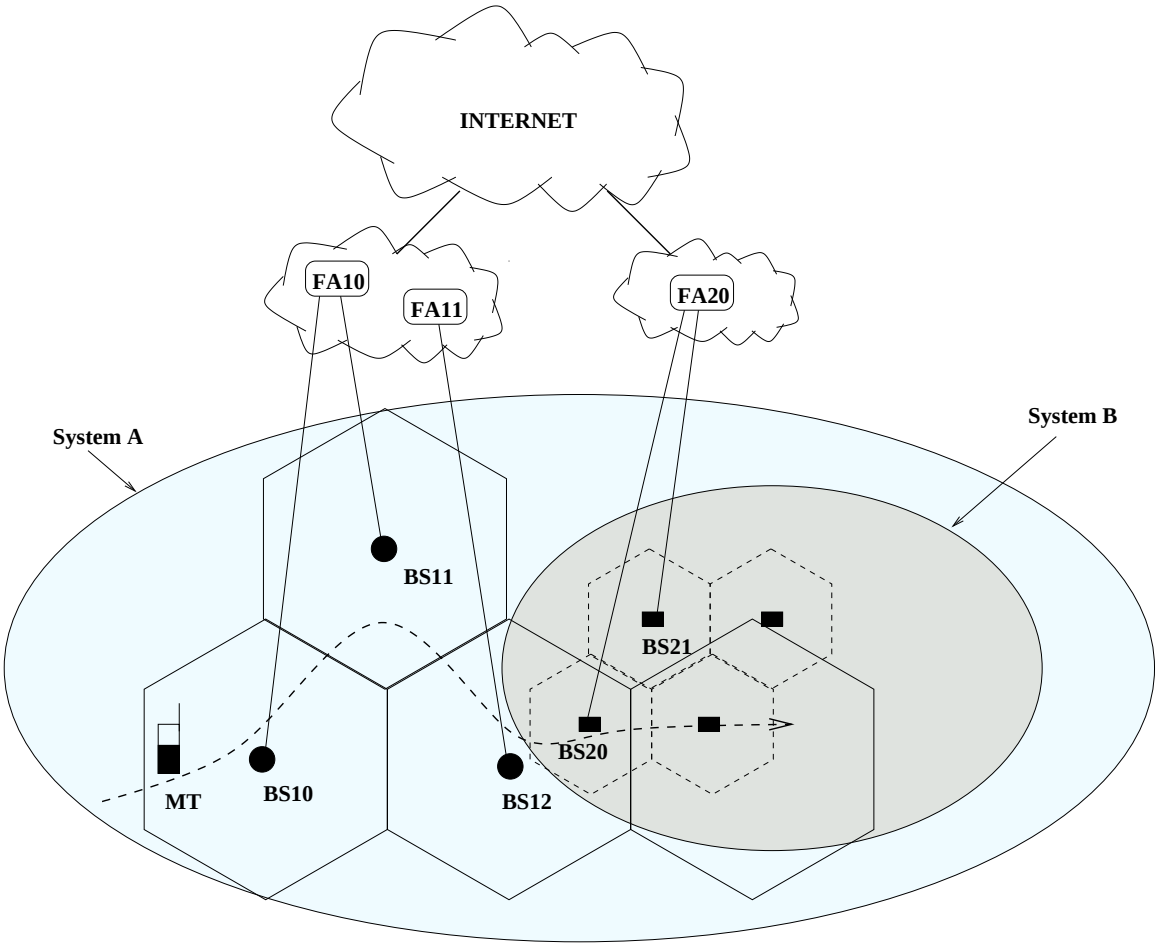


Figure 2: Mobility in the next-generation wireless systems.

1. **A ubiquitous mobile architecture for next-generation wireless systems:** Rapid progress in the research and development of wireless networking and communication technologies has created different types of wireless systems (e.g., Bluetooth, IEEE 802.11, UMTS, and satellite networks). These systems are envisioned to coordinate with each other to provide ubiquitous high-data-rate services to mobile users. A novel architecture, Architecture for ubiquitous Mobile Communications (AMC), is proposed that integrates these heterogeneous wireless systems. AMC eliminates the need for direct service level agreements among service providers by using a third party, a network interoperating agent (NIA). Instead of deploying a totally new infrastructure, AMC

extends existing infrastructure to integrate heterogeneous wireless systems. It uses Internet Protocol (IP) as the interconnection protocol. By using IP as the interconnecting protocol, transparency to the heterogeneities of the individual systems is achieved in AMC. Third-party-based authentication and billing algorithms are designed for AMC. New mobility management protocols are also developed to support seamless roaming between different wireless systems.

2. **A Cross-Layer (Layer 2 + 3) Handoff Management Protocol for Next Generation Wireless Systems:** In NGWS different wireless networks, each of which is optimized for some specific services and coverage area, will be integrated with each other to provide ubiquitous communications to the mobile users. It is an important and challenging issue to support seamless handoff management in this integrated architecture. The existing handoff management protocols are not sufficient to guarantee handoff support that is transparent to the applications in NGWS. A cross-layer (Layer 2 + 3) handoff management protocol, CMP, is proposed to support seamless intra- and inter-system handoff management in NGWS. CMP uses users' speed and handoff signaling delay information to enhance the performance of Mobile IP that is standardized to support handoff management in wireless IP networks. First, the performance of Mobile IP is analyzed with respect to its sensitivity to the link layer (Layer 2) and network layer (Layer 3) parameters. Afterwards, a cross-layer handoff management architecture is developed using the insights learnt from the analysis. Based on this architecture, the detailed design of CMP is carried out. Finally, extensive simulation experiments are carried out to evaluate the performance of CMP. The theoretical analysis and simulation results show that CMP significantly enhances the performance of both intra- and inter-system handoffs.

3. **Performance Analysis of Handoff Techniques based on Mobile IP, TCP-Migrate, and SIP:** Mobility management protocols operating from different layers of the classical protocol stack (e.g., link layer, network layer, transport layer, and application layer) are developed to support mobility in next-generation wireless systems. These protocols offer different handoff performance when used for different types of applications. To understand the effect of handoffs, first different types of mobile applications are grouped into five different classes, *Class A* through *Class E*, based on their mobility management requirements. Then, analytical modeling is developed to investigate the performance of the existing mobility management protocols for these classes of applications. The analysis shows that applications of a particular class experience different handoff performance when different mobility management protocols are used. Handoff performance comparison of different mobility management protocols are carried out to decide the suitable mobility management protocol for a particular class of application. The results of analysis advocate the use of transport layer, Mobile IP, and Session Initiation Protocol based mobility management for applications using TCP, non-real time applications using UDP, and real time applications using UDP, respectively. Moreover, through analytical modeling and performance investigation the parameters that influence the handoff performance of mobility management protocols are identified. The information about these parameters can be used to design new techniques to enhance the handoff performance of the existing mobility management protocols.

4. **Application Adaptive Multi-Layer Handoff Management in NGWS:** Different types of applications have different requirements in terms of handoff performance from a mobility management protocol. None of the existing

mobility management protocols is capable of supporting efficient handoff management for every type of application. An adaptive multi-layer mobility management framework (AMMF) is proposed that uses the mobility management protocol that best suits the handoff requirements of a particular application enabling application adaptive handoff support. To further enhance the handoff performance of mobility management protocols, AMMF uses information from different layers of the protocol stack realizing cross-layer interactions in the handoff process. Thus, it eliminates the adverse effects of other layers when mobility management protocols operate from one particular layer. First, the working principles of AMMF are developed. This is followed by the design of architectural components of AMMF. Then, analytical modeling is developed to investigate the handoff performance of AMMF. Finally, simulation experiments are carried out using the analytical modeling to evaluate the handoff performance of AMMF for different types of applications. The results show that AMMF significantly enhances the handoff performance for different classes of applications.

1.3 Thesis Outline

The objective of the proposed research is to design a scalable, secure, and robust architecture and to develop efficient mobility management protocols for next-generation wireless systems. More specifically, a new architecture is designed to integrate heterogeneous wireless systems. In addition, a cross-layer (Layer 2 + 3) handoff management protocol is developed for NGWS. Finally, a mobility management framework is developed to support mobility management that is adaptive to different types of applications. The rest of this thesis is organized as follows. Chapter 2 starts with the design of a new architecture for ubiquitous mobile communications in next-generation wireless systems. Next, a cross-layer (Layer 2 + 3) handoff management protocol,

CMP, is proposed in Chapter 3 to support seamless intra- and inter-system handoff management in NGWS. In Chapter 4, analytical modeling is developed to investigate the performance of the existing mobility management protocols for different types of applications. Next, a framework for multi-layer mobility management is developed in Chapter 5 to support the seamless handoff support to different types of applications in NGWS. Finally, Chapter 6 summarizes the research contributions and identifies several future research directions.

CHAPTER II

A UBIQUITOUS MOBILE COMMUNICATION ARCHITECTURE FOR NEXT GENERATION WIRELESS SYSTEMS

2.1 Introduction

Mobile users are demanding anywhere and anytime access to high speed data, real and non-real time multimedia services from next generation wireless systems (NGWS). These services have different requirements in terms of latency, bandwidth, and error rate.

Currently, there exist disparate wireless networks, such as Bluetooth for personal area, WLANs for local area, Universal Mobile Telecommunication System (UMTS) for wide area, and satellite networks for global networking. These networks are designed for specific service needs and vary widely in terms of bandwidth, latency, area of coverage, cost, and quality of service (QoS) provisioning. For example, satellite networks can provide global coverage, but are limited by high cost and long propagation delay (from 20-25 ms for Low Earth Orbit (LEO) satellite to 250-280 ms for Geostationary Earth Orbit (GEO) satellite). Third generation (3G) wireless systems, e.g., UMTS can deliver maximum data rate of 2 Mbps at a lower cost and has wide area of coverage. Whereas, WLANs support bandwidth up to 54 Mbps at extremely low cost. It may be noted that the future generation of WLANs are expected to provide data rate in excess of 100 Mbps. However, WLANs can support only low mobility users and have small coverage area. Therefore, none of the existing wireless systems can simultaneously satisfy the low latency, high bandwidth, and ubiquitous-coverage needs of mobile users at low cost. This necessitates a new direction in the design of NGWS.

There can be two possible approaches in designing NGWS:

- One way is to develop a new wireless system with radio interfaces and technologies which can satisfy the requirements of the services demanded by future mobile users.

- The other approach is to intelligently integrate the existing wireless systems so that the users may receive their services via the best available wireless system.

The first approach is expensive and needs more time for development and deployment, hence, it is not practical. Therefore, we advocate the use of the second approach which is a more feasible option [44]. Following the second approach, heterogeneous wireless systems, each of which is optimized for some specific service demands and coverage area, will co-operate with each other to provide ubiquitous "always best connection" [34] to the mobile users. In this integrated heterogeneous network architecture, each user is always connected to the best available network or networks.

The integrated NGWS keeps the best features of the individual networks, i.e., global coverage of satellite networks, wide mobility support of 3G systems, and high-speed and low-cost of WLAN. At the same time, it eliminates the weaknesses of the individual systems. For example, the low data rate limitation of 3G systems can be overcome when a WLAN coverage is available, through handover of the user to the WLAN. When the user moves out of a WLAN coverage area, it can be handed over to the overlaying 3G system. Similarly, a satellite network can be used when neither a 3G system nor WLAN is available. The basic idea is to use the best available network at anytime.

The above integrated NGWS must have the following characteristics: 1.) support for the best network selection based on users' service needs, 2.) mechanisms to ensure high quality security and privacy, and 3.) protocols to guarantee seamless inter-system mobility. Moreover, the architecture should be scalable, i.e., able to integrate

any number of wireless systems of different service providers who may not have direct service level agreements (SLAs) among them.

In this chapter, a novel architecture for NGWS, AMC (Architecture for ubiquitous Mobile Communications) is proposed that integrates heterogeneous wireless systems using a third party, Network Inter-operating Agent (NIA). AMC eliminates the need for direct SLAs among different network operators. It achieves transparency to the heterogeneities of individual systems by using Internet Protocol (IP) as the inter-connection protocol. AMC implements protocols for authentication, authorization, and billing when users move among different wireless systems. In addition, it also implements algorithms for the best network selection and protocols for inter-system mobility management.

2.2 Design Goals

In NGWS, users move between different networks as discussed in the previous section. They want to maintain their ongoing communications while moving from one network to another. These heterogeneous networks (WLANs, 3G cellular networks, and satellite networks) may or may not belong to the same service provider. Hence, the support for inter-system movement between networks of different service providers is required in NGWS.

One way of achieving roaming among networks of different service providers is to have bilateral SLAs among them. This approach is not feasible due to the following reasons.

- First, operators have reservations to open their network databases (which is required for authentication, billing, and service provisioning when SLA is established between operators) to all other operators.

- Second, each time a new operator deploys its wireless network, it has to create

a SLA with every other operator separately. The number of operators of wireless networks is very large, e.g., the number of GSM/GPRS operators alone is around 620. Similarly, there are a large number of operators for 3G networks, satellite networks, and WLANs. Given the large number of operators, it is almost impractical for network operators to create direct SLAs with every other operator. It may be noted that to overcome this problem in GPRS global roaming support, GSM association has proposed the use of GPRS roaming networks instead of direct SLAs among GPRS operators [4].

Therefore, there is a need for a new architecture to achieve roaming among heterogeneous networks of different service providers who may not necessarily have direct SLAs among them. We advocate that architecture of NGWS should have the following characteristics.

- **Economical:** The architecture should try to use as much of the existing infrastructure as possible and minimize the use of new infrastructures. This will ensure economical and speedy deployment.

- **Scalable:** The architecture should be able to integrate any number of wireless systems of both existing and future service providers.

- **Transparency to heterogeneous access technologies:** The architecture should be transparent to different access technologies of different networks.

- **Secure:** The architecture should be able to provide security and privacy equivalent to the existing wireless networks.

- **Seamless mobility support:** The architecture should support seamless mobility management to eliminate connection interruption and QoS degradation during inter-system roaming.

We survey the architectures for the integration of different communication systems proposed in the literature in the next section.

2.3 Related work

The concept of integrating two or more communication systems to get better performance is already in use and has been proven to be highly efficient. The existing integration architectures address the following issues: integration of two specific systems, integration of two general systems, integration of networks of multiple operators but of the same technology, and integration of networks of different operators employing different technologies. These architectures are described below.

In [31] [32], specific pairs of different systems are integrated through an additional gateway, such as interworking of DECT (Digital Enhanced Cordless Telephone) with GSM and interworking of IS-41 with GSM. The additional gateway proposed between a pair of systems takes care of interworking and inter-operating issues such as transformation of signaling formats, authentication, and retrieval of user profiles. Similarly, the integration of satellite and terrestrial networks have been studied in [66]. Appropriate interworking units, which are specific for the considered terrestrial networks, are placed at the interface between the satellite system and the terrestrial systems. In addition, different architectures are proposed to integrate WLAN and 3G systems [23]. All the above architectures are limited to the integration of a specific pair of systems and hence are not scalable to integrate multiple systems.

The Boundary Location Register (BLR) approach [11] is proposed to integrate any two adjacent networks with partially overlapping areas. However, this approach is not scalable in the sense that one BLR gateway is needed for each pair of adjacent networks when integrating multiple networks. Moreover, the above architecture assumes the existence of SLAs between the systems, which is not desirable as discussed in Section 2.2.

GSM association has proposed an inter-PLMN (public land mobile network) backbone using GPRS Roaming eXchange (GRX) [4] to globally integrate the GPRS networks deployed by various providers who may not necessarily have direct SLAs among them. This architecture uses multiple peer GRX nodes for connecting several GPRS networks. This architecture is limited to only one technology, i.e., GPRS networks.

In SMART project [38], a new architecture is proposed to integrate the heterogeneous wireless systems. This architecture uses two distinct networks: *basic access network*, and *common core network* for signaling and data traffic, respectively. This architecture is scalable, but requires the development and deployment of the new basic access and common core networks and hence is not cost-effective.

Heterogeneous network integration using Mobile IP and SIP are proposed in [33] and [69], respectively. In these architectures, Mobile IP and SIP use Authentication, Authorization, and Accounting (AAA) agents to carry out authentication and accounting during inter-network roaming. However, these architectures do not have any mechanism to decide the best available network. Moreover, although Mobile IP and SIP protocols are used to carry out inter-system handoff, seamless support of inter-system handoff is not always guaranteed [13].

None of the above architectures satisfy all the requirements of the NGWS outlined in Section 2.2. This is the motivation behind designing a new architecture for NGWS with all the design goals. The details of the proposed architecture are presented in the next section.

2.4 The Proposed Architecture

First, the motivation for selecting IP to integrate different wireless systems is discussed. This is followed by the detailed description of AMC.

2.4.1 IP-Based Inter-Connection

The integrated NGWS has the following heterogeneities:

- Access technologies: NGWS will include many heterogeneous access networks using different radio technologies, e.g., GPRS, cdma2000, UMTS, WLAN, etc.
- Network protocols: NGWS will have different protocols for transport, routing, mobility management, etc.

These heterogeneities ask for a common infrastructure to inter-connect the heterogeneous networks. Since IP provides a globally successful infrastructure for supporting applications in a scalable and cost effective way, it is recognized to become the core backbone network of NGWS.

By using IP as the common inter-connection protocol, mobile users may roam among multiple wireless networks in a manner that is transparent to different radio technologies. This is achieved by using Mobile IP [63] protocol to support roaming between different access technologies. This IP-based inter-connection solution hides the heterogeneities of the lower layer technologies from higher layers. Moreover, in NGWS, IP-based mobile devices with multiple radio interfaces may switch from one network interface to another by using multiple care-of-addresses (CoAs), one for each interface. In this scenario, the interface switching is carried out as defined in [84]. Therefore, this approach requires no modifications to the existing heterogeneous radio technologies and provides the greatest transparency to ubiquitous communications in a heterogeneous network environment.

2.4.2 Architecture for Next Generation Wireless Systems

Architectures requiring direct SLAs among different providers are not feasible because of the reasons mentioned in Section 2.2. We propose the use of a third party to integrate heterogeneous wireless systems of different service providers. In this case, an individual network operator needs to establish the direct SLA only with the third party instead of establishing separate SLAs with every other operators.

The proposed Architecture for ubiquitous Mobile Communications (AMC) for

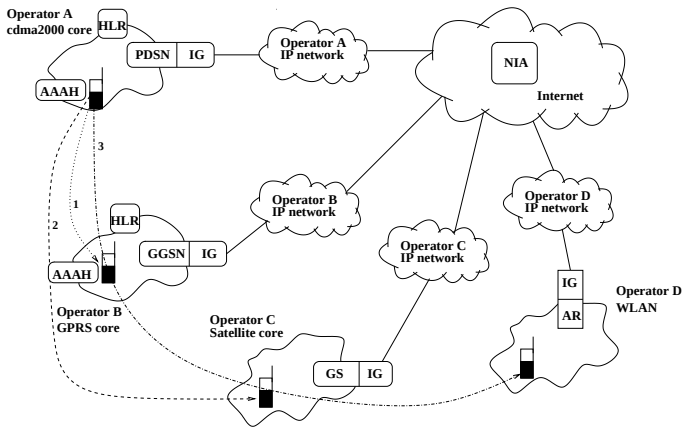


Figure 3: NIA based integrated architecture for NG wireless systems.

NGWS is shown in Figure 43, which consists of a cdma2000, GPRS, satellite network, and WLAN of service providers A, B, C, and D, respectively. These systems are connected to the Internet through gateways, e.g., cdma2000 is connected to the Internet via Packet Data Serving Node (PDSN), GPRS through Gateway GPRS Support Node (GGSN), satellite network through Gateway Station (GS) and WLAN through Access Router (AR). It may be noted that AMC can integrate any number of systems of different service providers.

Two new entities **Network Inter-operating Agent** (NIA) and **Interworking Gateway** (IG) are defined for AMC. The NIA functions as the third party and IG as the gateway between a particular system and the NIA. The NIA resides in the Internet, whereas IG resides in each system and acts as the gateway as shown in Figure 43. Instead of getting connected to every other system, IG is connected to only one entity, NIA. It can be implemented as a separate entity or can be integrated with the gateways through which individual systems are connected to the Internet, e.g., PDSN, GGSN, GS, AR, in case of cdma2000, GPRS, satellite, and WLAN, respectively as shown in Figure 43. We advocate the latter choice because in this case IG can be plugged into the existing infrastructure, hence it is easy to implement and manage.

In AMC, the network providers do not have to create separate SLAs with every other providers. Instead, they offer roaming services to subscribers of other providers with only one SLA with the NIA. This eliminates the need for separate SLAs between each pair of systems and makes AMC scalable. The NIA is supported by a third party provider. It is assumed that the operator of the NIA generates revenue from the network providers who have SLAs with the NIA. It may be noted that network providers charge more from their subscribers when the later communicate through a foreign network. We advocate that the providers share a part of this revenue generated during the inter-system roaming of their subscribers with the NIA. The operators will be interested to use NIA to support roaming to the networks of other operators as a value-added-service feature to their subscribers. For example, a similar business model is used by iPass company to provide global remote access services.

The NIA handles the authentication, authorization, billing, and mobility management issues of inter-system roaming. Currently, the Authentication, Authorization, and Accounting (AAA) broker networks support authentication, authorization, and billing for users belonging to different service providers. However, they can not handle the mobility management issues, and hence, can not be used as the third party.

The components of the NIA and the IG are described in the following subsection.

2.4.3 Components of the NIA and IG

The sub-systems of the NIA are as follows. These are shown in Figure 44(a).

- The *authentication unit* is used to authenticate the users moving between two systems belonging to two different service providers as discussed in Section A.2.1.1.
- The *accounting unit* handles the billing issues between different systems as discussed in Section A.2.1.2.
- The *operators database* stores information about the network operators who have SLAs with the NIA.

- The *handover management unit* decides if the inter-system handover (ISHO) request should be granted. The handover management unit derives the *Network Access Identifier* (NAI) from the Mobile IP *Registration Request* message and verifies with the *operators database* for the existence of the SLA with the home operator of the mobile terminal (MT). When applicable, it also acts as the mediator between different networks, e.g., for transferring user service profiles. In addition, the handover management unit decides the best available network as discussed in Section 2.6.1.

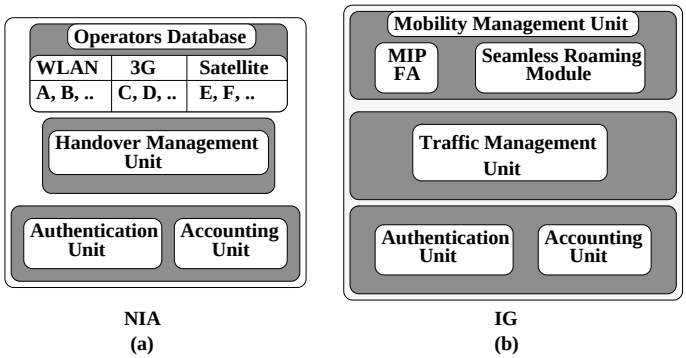


Figure 4: Logical diagram showing the subsystems of NIA and IG.

The components of the IG are described below. These are shown in Figure 44(b).

- The *mobility management unit* implements Mobile IP [63] (MIP) functionalities using the MIP foreign agent (FA). Note that when a particular wireless system already implements Mobile IP, e.g., cdma2000, there is no need to implement the FA in the IG. In this case, the FA in the IG refers to the FA already implemented in the system. The *mobility management unit* has a *seamless roaming module* which will implement mobility management protocols for seamless inter-system roaming as discussed in Section 2.6.2.

- The IG implements traffic monitoring function in its *traffic management unit*. The specific implementation of this unit may be different for different providers based on their policies.

- The *authentication unit and accounting unit* provide authentication and billing support, respectively, to the roaming users (refer to Section A.2.1).

2.5 Security and Billing Support in AMC

In AMC, authentication, authorization, and billing mechanisms are carried out as follows.

2.5.1 Authentication and Authorization

The proposed security architecture for AMC is shown in Figure 45, where the foreign network (FN) is the network the MT is currently visiting and the HN is the home network of the MT. This architecture glues the security architectures of the FN and HN through *Authentication Unit* (AU) of the NIA (AU_NIA). The use of AU_NIA eliminates the need for any direct security association/agreement between the FN and HN. Both the FN and HN have separate security associations/agreements with AU_NIA. Thus, AU_NIA functions, in essence, as a trusted third party for authentication and authorization dialogs between the FN and HN. The working principle of this third-party-based security architecture is as follows. When a mobile user requests services from a FN and the FN determines that it has no SLAs with the HN provider, it forwards the request to AU_NIA to authenticate and authorize the user. Then, AU_NIA talks to the HN provider and mediates between the FN and HN for authentication and authorization message exchanges. Once the user is authenticated, AU_NIA mediates for the creation of security associations/keys that are required between the FN and HN. At the end, the HN and FN will be mutually authenticated and will have session keys for secured data transfer.

In AMC the authentication, authorization, and Mobile IP registration processes are integrated as defined in [33]. The architecture in Figure 45 shows the existing security associations along with the required MIP security associations so that the

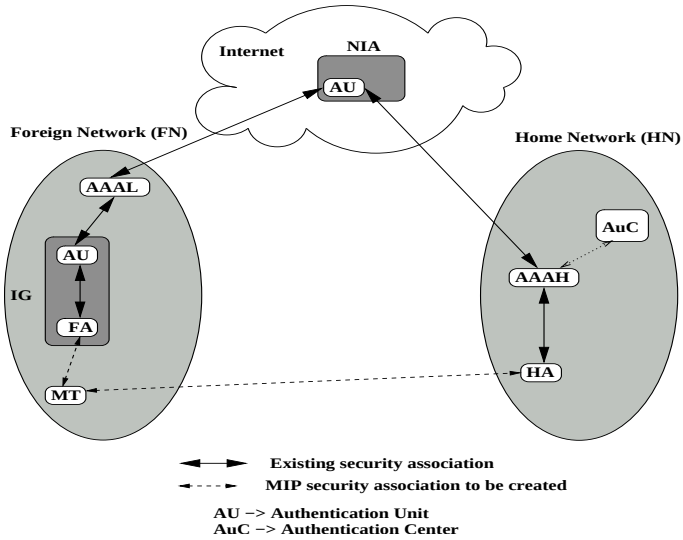


Figure 5: The proposed security architecture for AMC.

FN will be able to deliver services to a roaming MT. Extensible Authentication Protocol (EAP) over Diameter is used for end-to-end mutual authentication between an MT and its home AAA server (AAAH). When the MT roams into the FN domain, the authentication, authorization, and MIP registration are carried out as described below. The signaling messages for this procedure are shown in Figure 46. Here, EAP-SIM [37] is used to illustrate the authentication process. Note that any other authentication schemes, e.g., EAP-AKA, EAP-SKE, EAP-TLS etc., can also be used.

1. When the MT hears MIP *Agent Advertisement* (*step 1*), it sends MIP *Registration Request* including *Mobile-AAA Authentication and Authorization* extensions (as defined in [24]) to the FA located in the IG (*step 2*). The MT also includes a *SIM Key Request* extension [36] and a *Network Access Identifier* (NAI) [25], e.g., *MT@relam*, in its MIP *Registration Request*.

2. When the FA receives the MIP *Registration Request* and finds the *Mobile-AAA Authentication and Authorization* extensions, it learns that the MT is a roaming user and forwards the MIP *Registration Request* to the Authentication Unit in the IG (AU_IG) (*step 3*). Based on the NAI in the MIP *Registration Request*,

the AU_IG recognizes that the FN does not have a direct SLA with the HN of the MT and forwards the MIP *Registration Request* to the Authentication Unit in the NIA (AU_NIA), either directly or through other AAA proxies (*step 4*).

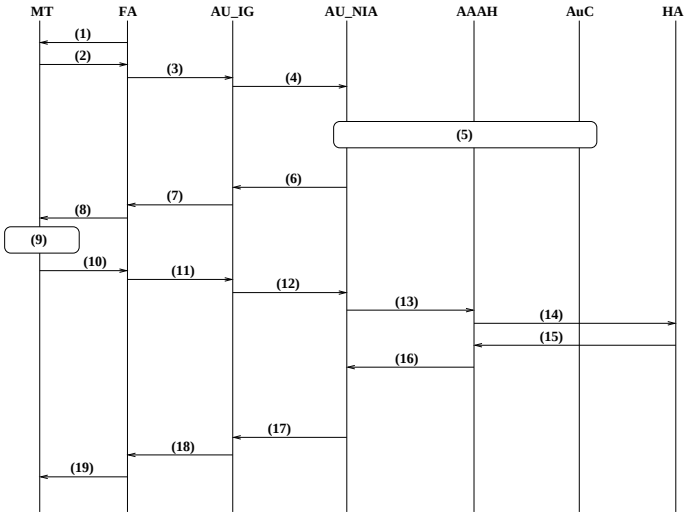


Figure 6: The authentication and authorization signaling messages for AMC.

3. The AU_NIA examines the NAI of the received MIP *Registration Request* message and forwards it to the Home AAA server (AAAH). Once the AAAH receives the MIP *Registration Request*, it first verifies the *Mobile-AAA Authentication and Authorization* extensions. If these extensions are valid, it contacts the home authentication center (AuC) of the MT. The AuC generates n triplets (RAND, SRES, K_c), where RAND denotes a random number, SRES denotes the response, and K_c is the key used for encryption. Then the AuC calculates message authentication code (MAC) for the RANDs (MAC_RAND) as defined in [36]. The AuC sends the RANDs and MAC_RAND to the AAAH, which forwards those to the AU_NIA (all these constitute *step 5*). Then, the AU_NIA forwards these to the AU_IG (*step 6*). Finally, the AU_IG forwards these to the FA (*step 7*). The FA sends an MIP *Registration Reply* message to the MT containing the RANDs and MAC_RAND (*step 8*). The MT derives the corresponding SRES and K_c values using its SIM card and the received RANDs.

It also calculates MAC_RAND and validates the authenticity of RANDs by comparing the calculated MAC_RAND with the received MAC_RAND, thus confirming that the RANDs are generated by the HN. If the MAC_RAND is valid, the MT calculates an MAC for its SRES values as defined in [36] (*step 9*). The MAC_SRES is used by the AuC to know if the SRES values are fresh and authentic. The MT also generates security association keys: K_{MT_FA} for the FA and K_{MT_HA} for the HA as defined in [36]. These keys are used to authenticate subsequent MIP registrations until the key lifetime expires.

4. Now, the MT sends another MIP *Registration Request* message to the FA containing *SRES* extension [36] and *Mobile-AAA Authentication and Authorization* extensions (*step 10*). When the FA detects the presence of *Mobile-AAA Authentication and Authorization* extensions, it forwards the MIP *Registration Request* message to the AU_IG (*step 11*), which forwards it to the AU_NIA (*step 12*). The AU_NIA forwards the MIP *Registration Request* message to the AAAH (*step 13*). After successful authentication and authorization (this may require the interaction of the AAAH and AuC), the AAAH forwards the MIP *Registration Request* to the HA (*step 14*) containing K_{MT_HA} security key. The HA carries out the registration for the MT as defined in [63], extracts the K_{MT_HA} key, and sends MIP *Registration Reply* to the AAAH (*step 15*). The AAAH forwards the MIP *Registration Reply* (containing K_{MT_FA} and the K_c keys) to the AU_NIA (*step 16*). Then, the AU_NIA forwards the MIP *Registration Reply* to the AU_IG (*step 17*). The AU_IG forwards it to the FA (*step 18*). The FA extracts K_{MT_FA} and K_c keys and sends an MIP *Registration Reply* to the MT (*step 19*). The K_c keys are used for secure data transfer between the MT and the FA providing confidentiality and integrity to the data traffic. If necessary, a FA-HA security association key can be generated by the AuC as defined in [36] and distributed to the FA and the HA as a part of authentication process.

2.5.2 Billing

Once the MT is authenticated and authorized by the FN, *Accounting Unit* of the IG (ACU_IG) maintains a per user accounting record based on the charging policy of the FN provider (e.g., connection duration, amount of data transferred, etc.). It transfers the accounting information either on per session basis or in real-time to the AAAL server of the FN domain. The AAAL server collects and consolidates the accounting information for the MT and forwards it as FN access call detail records (FN CDRs) to the *Accounting Unit* of the NIA (ACU_NIA). The NIA is capable of interpreting FN CDRs. However, it may happen that HN of the MT supports a different CDR format. Then, the NIA first converts the FN CDR format to the CDR format supported by the HN and forwards the final CDRs to AAAH for billing purposes. ACU_NIA is responsible for the inter-operation of different billing schemes supported by different network providers.

2.6 Inter-system Handover Protocols

Two types of inter-system roaming may arise in AMC. They are:

1. roaming between fully overlapping systems, which can be further classified as
    - roaming from a lower tier system to a higher tier system, e.g., (1) and (3) in Figure 7
    - roaming from a higher tier system to a lower tier system, e.g., (2) and (4) in Figure 7
2. roaming between partially overlapping systems, e.g., (5) in Figure 7.

Note that a lower tier system supports greater bandwidth than a higher tier system.

When any type of the above inter-system roaming occurs, inter-system handover (ISHO) is carried out. ISHO is also referred as vertical handoff. It is essential that

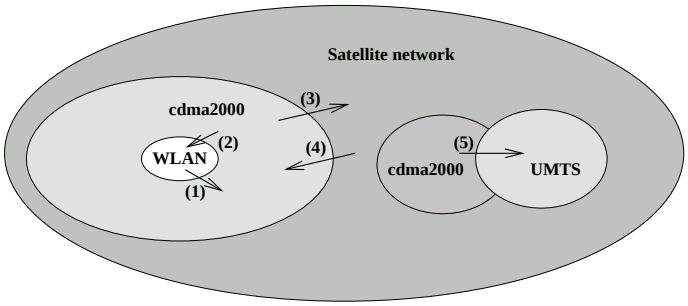


Figure 7: Next generation integrated systems scenario.

the applications running on an MT remain unaware of the roaming to ensure uninterrupted services with minimum QoS degradation. This can be achieved by reducing the handoff failure probability and latency to the values that are tolerable by the applications.

Several issues need to be addressed during the vertical handoff. When an MT is accessible through multiple fully overlapping systems, first, based on the service needs of a user, the best communication network should be determined. Then, the handoff initiation time is determined to guarantee a successful inter-system roaming. The authentication, authorization, and accounting procedures are then carried out before the MIP registration process.

2.6.1 Best Network Selection

The NIA helps each MT to be "always best connected" [34] by selecting the best available network for communications. Several factors influence the design of policies on the best network selection for vertical handoff. Monetary cost, network conditions, power consumption, user activity history, and the required QoS from applications are considered as the decision metrics. Moreover, the best network selection also affects the distribution of the overall system load.

AMC uses the hybrid network selection scheme [92] that combines terminal-based and network-based selection mechanisms to select the best available network.

Terminal-based mechanism allows MTs periodically collect dynamic network conditions and determine the best reachable network for handoff by themselves. Network-based mechanism makes globally optimized selection and achieve load balancing for the whole system. The objective of this best network selection scheme is to provide satisfactory overall performance of the whole system as well as take into account the user preferences. It is a two-level decision-making scheme. At the first level, each MT monitors and collects the dynamically varying network conditions for decision-making at the terminal side. At the second level, the *handover management unit* inside the NIA finds the optimal user distribution for each individual network based on global observations. The decision made by this central controller is fed back to the first-level decision as adjustments. The details of the hybrid network selection scheme is in [92].

2.6.2 Handoff Initiation Time Estimation

After the best available network is selected, the next challenge is to determine the right time to start handoff procedures. Currently, there are several proposals which use the physical and MAC layer sensing to determine the appropriate time for vertical handoff initiation. In these algorithms, the implicit assumption is that the signaling delay associated with vertical handoff is constant. Based on this assumption, these algorithms initiate the vertical handoff when the received signal strength (RSS) of the serving network goes below a certain fixed threshold value, S_{th}. However, in a real scenario, the vertical handoff signaling delay varies from few seconds to several tens of seconds depending on several factors, e.g., traffic level in the backbone network, the wireless link quality, and the distance between the user and its home network. Therefore, the protocols that are designed based on a fixed vertical handoff signaling delay have poor performance.

In Chapter 3, we propose the use of dynamic RSS threshold to eliminate the effect of signaling delay variation. Towards this, AMC predicts the handoff signaling delay

and estimates MT's speed. Then, it determines a dynamic threshold value for the RSS, S_{dth}, based on the handoff signaling delay and speed information such that if the vertical handoff procedures are initiated when the RSS of the serving network goes below S_{dth}, they are completed before the user moves out of the coverage area of the serving network. The *seamless roaming module* in the IG implements the algorithm for the estimation of S_{dth}. Details of this proposed scheme is presented in Chapter 3.

2.6.3 ISHO Protocols for Fully Overlapping Systems

2.6.3.1 ISHO protocols for lower to higher tier roaming

When an MT is moving out of the coverage area of a lower tier system, the goal is to switch it to the overlaying higher tier system before the lower tier link breaks. The associated mobility management protocols for this scenario are described using Figure 49.

The MT first enables its interfaces for the higher tier systems and determines the best network to be handed off to (*step 1*). When the handoff initiation time is determined, it registers with the higher tier system using Mobile IP (MIP) [63] registration procedures. Authentication and authorization procedures are combined with MIP registrations as discussed in Section A.2.1.1 (*step 2*). The MT also maintains its registration with the lower tier system using simultaneous mobility binding [84] with both the systems.

After successful registration with the higher tier system, the MT uses both the lower and higher tier systems for downlink traffic, but uses only the lower tier system for uplink traffic as long as it is within the coverage area of the lower tier system to take advantage of the higher data rate of the lower tier system (*step 3*). With the established connection with the higher tier system, the ongoing communications of the MT can be immediately switched to the higher tier system, when it moves out of the lower tier system. This ensures a seamless ISHO. Once it moves out of the coverage area of the lower tier system, it uses only the higher tier system for its

communications (*step 4*).

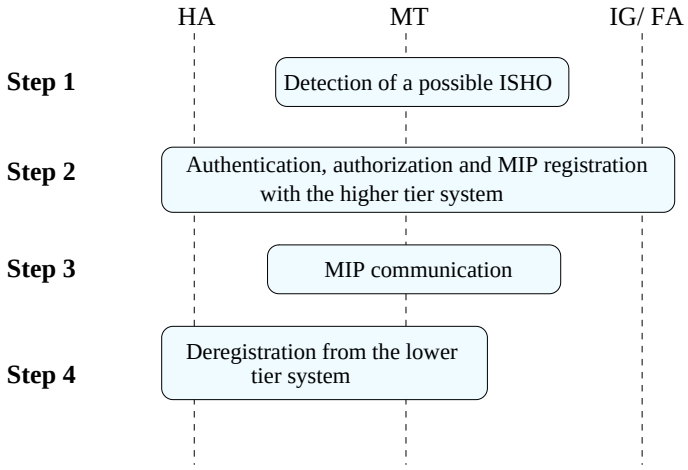


Figure 8: Steps for lower to higher tier roaming.

2.6.3.2 ISHO protocols for higher to lower tier roaming

A higher tier system completely overlaps a lower tier system. Therefore, when an MT roams from a higher tier system to a lower tier system, the MT can always keep its connection with the higher tier system to ensure no connection loss.

The MT initiates an ISHO by sending an MIP *Registration Request* message to the FA located in the IG of the corresponding lower tier system. The FA determines that the MT is a roaming user and starts the process of ISHO by forwarding the MIP *Registration Request* message to the NIA through AU_IG (refer to Section A.2.1.1). The NIA determines if the MT has the permission to access the lower tier system using its *operators database* as discussed in Section 2.4.3. If the outcome is yes, the NIA proceeds with the MIP registration process along with authentication and authorization as discussed in Section A.2.1.1 (*step 2*). After successful registration with the lower tier system, the MT starts communicating through the lower tier system and de-registers from the higher tier system (*step 3 and 4*).

2.6.4 ISHO Protocols for Partially Overlapping Systems

In case of adjacent systems, when the MT detects that it is moving out of the coverage of the serving system, it enables the interfaces and searches for an available system (*step 1*). When it finds the new system, it registers with that system using Mobile IP [63] registrations procedures. Authentication and authorization procedures are combined with MIP registration as discussed in Section A.2.1.1 (*step 2*). The MT also maintains its registration with the old system using simultaneous mobility bindings to both the systems for a predefined period of time to avoid ping-pong effect during the ISHO. After successful registration with the new system, it uses both the old and new systems for downlink traffic. It uses only the new system for uplink traffic (*step 3*). After the specified time period, if it does not move back to the old system, it de-registers from the old system and uses only the new system for its communications (*step 4*).

2.7 Performance Evaluation of AMC

In this section, qualitatively evaluation of the proposed architecture, AMC, is carried out in the context of the design goals stated in Section 2.2.

Economical: AMC uses the access and core network infrastructure of the existing wireless systems. It does not require any change to the infrastructure of the existing networks. AMC achieves the integration of heterogeneous networks by adding only one new entity, integration gateway (IG), to the individual networks. Hence, it is economical.

Scalability: AMC can integrate any number of wireless systems of different providers who may not have SLAs among them by using the NIA as the third party. Therefore, it is scalable. To further enhance the scalability, we propose the hierarchical NIA structure to integrate the wireless networks globally. In this hierarchical structure, wireless networks of various providers are integrated at the regional (e.g., city) level

through the first-tier NIAs. These regional NIAs of a particular country or several countries are then integrated through the second-tier NIAs, followed by the integration of the second-tier NIAs through the third-tier NIAs to realize global integration. The exact number of tiers and the number of NIAs at each tier depend on several factors, such as the number of network providers in that tier and the number of roaming users. Determination of the number of NIAs required for a particular deployment scenario can be carried out. This is beyond the scope of this paper.

These NIAs can be owned by a single operator or multiple operators with SLAs among them. Note that the number of NIA operators is small. Hence the required SLAs among NIA operators is only a few. Therefore, the scalability of AMC is not compromised when multiple operators own NIAs. An NIA operator is responsible for aspects of heterogeneous wireless system integration of a particular region and supports their inter-working with other wireless systems globally, through the establishment of SLAs with other NIA operators.

In this hierarchical NIA structure, a network operator only needs to have SLAs with a set of nearby first tier (aka regional) NIA operators to be able to provide its subscribers with global access.

The NIA is involved only during the ISHO process and transfers the control signals between two systems. Once the ISHO is over, the data traffic of the roaming users does not go through the NIA as discussed in Section A.4. Therefore, the load on the NIA is limited.

Transparency to heterogeneous access technologies: By using IP as the common inter-connection protocol in AMC, mobile users may roam among multiple wireless networks in a manner that is completely transparent to different radio technologies.

Security: AMC adopts the state-of-the-art security mechanisms such as SIM to provide security and privacy equivalent to the existing wireless networks.

Seamless mobility support: AMC supports seamless inter-system mobility using

Mobile IP (MIP) as the mobility management protocol. AMC further improved the performance of ISHO by using a dynamic RSS threshold for ISHO initiation. This reduces latency and packet loss during ISHO.

2.8 Summary

In this research, a third-party-based architecture, AMC, is proposed to integrate heterogeneous wireless systems. The design goals of AMC are cost, scalability, transparency, security, and seamless mobility support. AMC reduces the cost of architecture deployment by using the access and core network infrastructure of the existing wireless systems. AMC integrates heterogeneous wireless systems of different operators who may not necessarily have direct service level agreements among them. Furthermore, security equivalent to the existing wireless systems is achieved in AMC. Finally, inter-system handover is implemented in AMC to achieve seamless roaming. The integration of third generation wireless networks (3G) and WLANs is gaining increasing importance to provide broadband access to the mobile users. AMC is used to specifically integrate these two networks. The details of the 3G and WLAN integration using AMC is described in Appendix A.

CHAPTER III

A CROSS-LAYER (LAYER 2 + 3) HANDOFF MANAGEMENT PROTOCOL FOR NEXT GENERATION WIRELESS SYSTEMS

3.1 Introduction

As a result of rapid progress in research and development, today's wireless world witnesses several heterogeneous communication networks, such as Bluetooth for personal area, IEEE 802.11 for local area, UMTS (Universal Mobile Telecommunications System) for wide area and satellite networks for global networking. These networks are complementary to each other and hence their integration can realize a unified next generation wireless system (NGWS) that has the best features of the individual networks to provide ubiquitous 'always best connection' [34] to the mobile users [9]. A novel architecture for NGWS, AMC, is proposed in Chapter 2.

In AMC [9], users are always connected to the best available networks and switch between different networks based on their service needs [9]. It is an important and challenging issue to support seamless mobility management in AMC. Mobility management contains two components: location management and handoff management [8]. Location management enables the system to track the locations of mobile users between consecutive communications. On the other hand, handoff management is the process by which a user keeps its connection active when it moves from one base station (BS) to another. There exist efficient location management techniques in the literature for NGWS [91] [90]. These can be used in AMC. However, seamless support of handoff management in NGWS is still an open issue [13].

Figure 9 shows a typical handoff scenario in the NGWS. The integrated architecture in Fig. 9 consists of two different wireless systems. System A is a macro-cellular

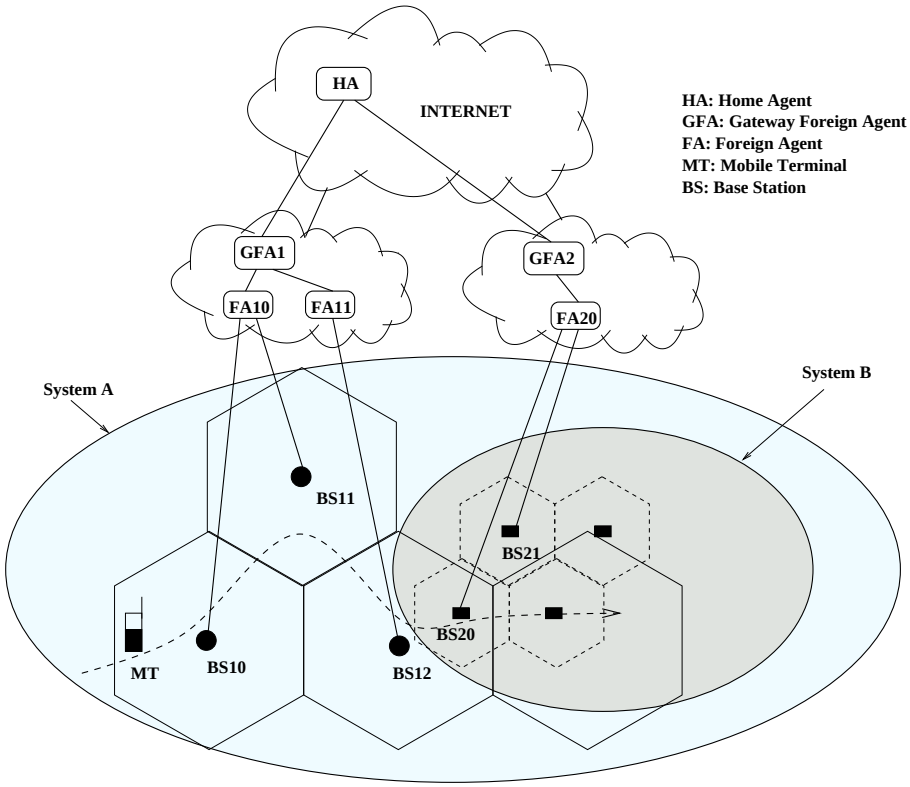


Figure 9: Mobility in the integrated NGWS architecture.

system such as cdma2000 based 3G network, whereas System B is a micro-cellular system such as 802.11 based WLAN. It may be noted that System B can also be another macro-cellular network. These two systems are integrated through the Internet backbone [9]. It may be noted that in a real scenario the integrated architecture may consist of many different wireless systems. Figure 9 shows the architectural components of hierarchical Mobile IP [35] protocol. In NGWS, two types of handoff scenarios arise: horizontal handoff and vertical handoff [13] [75].

- *Horizontal Handoff*: Handoff between two BSs of the same system. Horizontal handoff can be further classified into

  - *Link-Layer Handoff*: Horizontal handoff between two BSs that are under the same foreign agent (FA), e.g., the handoff of the MT from BS10 to BS11 in Fig. 9.

  - *Intra-System Handoff*: Horizontal handoff between two BSs that belong to two different FAs and both the FAs belong to the same system and hence to same gateway foreign agent (GFA), e.g., the handoff of the MT from BS11 to BS12 in Fig. 9.

- *Vertical Handoff (Inter-System Handoff)*: Handoff between two BSs that belong to two different systems and hence to two different GFAs, e.g., the handoff of the MT from BS12 to BS20 in Fig. 9.

Efficient algorithms are present in the literature that support link-layer handoff transparent to the applications [95]. Therefore, in this work we do not address the link-layer handoff. On the other hand, seamless support for intra- and inter-system handoff is still an open issue [13]. The large value of signaling delay associated with the intra- and inter-system handoff [27] can be above the threshold required for the support of delay-sensitive or real-time services [50]. In addition, the packets in transit can not be delivered to the MT during this high handoff latency period causing significant packet loss during handoff. We advocate that efficient intra- and inter-system handoff protocols should have the following characteristics to support seamless roaming in NGWS.

- **Minimum handoff latency:** The handoff management protocols should introduce only minimum handoff latency to the ongoing communications.

- **Low packet loss:** Packet loss during handoff should be minimized.

- **Limited handoff failure:** Handover failure should be limited to a predefined value.

Handoff management protocols operating from different layers of the classical protocol stack (e.g., link layer, network layer, transport layer, and application layer) are proposed in the literature [13]. Mobile IP [63] that operates from the network layer

is proposed to support mobility in IP-based networks. It forwards packets to mobile users that are away from their home networks using IP-in-IP tunnels [63]. Transport layer mobility management protocols are proposed to support mobility between networks that eliminates the need for tunneling of the data streams. An architecture called MSOCKS is proposed in [51] for transport layer mobility. MSOCKS implements transport layer mobility using a split-connection proxy architecture and a new technique called TCP Splice that gives split-connection proxy systems the same end-to-end semantics as usual TCP connections [51]. Moreover, work is going on in the IETF to modify the Stream Control Transmission Protocol [76] to allow it to dynamically change endpoint addresses in the midst of a connection [29] [39]. Recently, application layer mobility using Session Initiation Protocol (SIP) is proposed in [87]. SIP based mobility does not require any changes to the IP stack of the mobile users. In addition, device independent personal mobility and location services are supported by SIP mobility.

The standard network layer mobility management protocol, Mobile IP [63], is simple to implement, but has several shortcomings, such as *triangular routing*, *high global signaling load*, and *high handoff latency* [13]. Mobile IP route optimization [64] eliminates the triangular routing problem. Hierarchical Mobile IP [35] and other micro-mobility protocols such as Cellular IP [82], IDMP [55], and HAWAII [67] address the problem of high global signaling load and high handoff latency by introducing another layer of hierarchy to the base Mobile IP architecture to localize the signaling messages to one domain. Mobile IP handoff latency is composed of latencies for handoff requirement detection and Mobile IP registration [94]. The proposed hierarchical Mobile IP and micro-mobility solutions [55, 67, 82] particularly achieve reduction in registration signaling delay, but fail to address the problem of handoff requirement detection delay [94].

Therefore, recently the use of link layer information to reduce the handoff requirement detection delay has gained attention [12] [13] [50]. The basic idea behind this approach is to use the link layer information to anticipate the possibility of an intra- or inter-system handoff in advance so that the handoff procedures can be carried out successfully before the MT moves out of the coverage area of the serving base station (BS). The use of link layer information significantly reduces the handoff latency and the handoff failure probability of handoff management protocols [13].

The user mobility profile (UMP) is used in [12] to support enhanced mobility management. The concept of inter-system boundary cells are used in [52] to prepare the users for a possible inter-system handoff in advance. Thus, significant reduction of inter-system handoff failure probability is achieved. A generic link layer technique is used in [50] to aid the handoff protocols operating from the upper layers. However, it does not specify any particular mechanism for obtaining the link layer triggers. Different link layer assisted handoff algorithms that use the received signal strength (RSS) value to reduce the handoff latency and handoff failure are proposed in [23] [94] [96]. However, these studies are limited to the mobility between 3G and WLAN systems. There are some other studies that use the RSS measurements to track the mobile nodes (MNs) and then use the tracking information to support low latency Mobile IP handoff such as S-MIP [42].

The above link-layer assisted handoff protocols implicitly assume that the handoff latency of the intra- and inter-system handoffs are constant. Based on this assumption, the link-layer assisted handoff protocols initiate the handoff when the RSS of the serving BS goes below a pre-defined fixed threshold value. However, in a real scenario the signaling delay of the intra- and inter-system handoffs depends on the traffic level in the backbone network, the wireless link quality [18], and the distance between the user and its home network at the handoff instance. Therefore, the protocols that are designed assuming a fixed delay for intra- and inter-system handoff have

poor performance when the handoff signaling delay varies. Moreover, the existing link-layer assisted handoff protocols do not consider the influence of users' speed on the performance of the handoff protocols. Our analysis in Section 3.2 shows that users' speed has significant effect on the performance of the handoff protocols. In addition, to the best of our knowledge there is no existing work that determines how the link layer information can be used to guarantee desired performance in terms of handoff latency and handoff failure probability.

In this chapter, first the performance of the existing network layer handoff management protocol, hierarchical Mobile IP (HMIP), is analyzed with respect to its sensitivity to the link layer (Layer 2), e.g., users' speed and network layer (Layer 3), e.g., handoff signaling delay parameters. Next, a cross-layer handoff management architecture is developed using the results of the analysis. Then, using the cross-layer architecture a cross-layer protocol, CMP, is designed to support enhanced handoff management in NGWS. CMP uses users' speed and handoff signaling delay information and enhances the performance of HMIP handoff significantly. Finally, extensive simulation experiments are carried out to evaluate the performance of CMP. The theoretical analysis and simulation results show that CMP significantly enhances the performance of both intra- and inter-system handoffs. CMP jointly addresses all the desired characteristics of an efficient handoff management protocol mentioned earlier.

3.2 Effect of Layer 2 and Layer 3 Parameters on the Performance of Handoff Management Protocols

In this section, an analytical framework is developed to answer the question: **how should the Layer 2 and Layer 3 information be used to make sure that the handoff performance remains the same irrespective of users speed and network dynamics?**

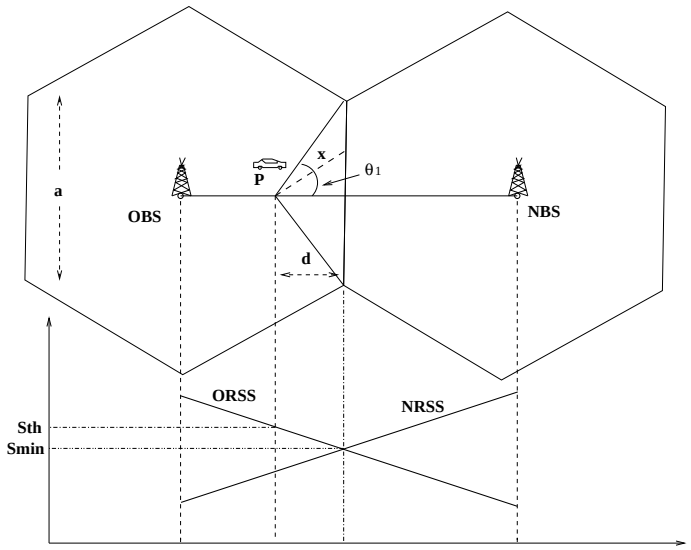


Figure 10: Analysis of the handoff process.

The following notations are defined with reference to Figure 10 that shows the handoff from the current BS referred as old BS (OBS) to the future BS referred as new BS (NBS).

- S_{th}: the threshold value of the RSS to initiate the HMIP [35] handover process. This implies, when the RSS of OBS goes below S_{th}, the HMIP registration procedures are initiated for MT's handover to the NBS.
- S_{min}: the minimum value of RSS required for successful communication between an MT and OBS.
- a: the length of each side of hexagonal cells.

A scenario where an MT is currently served by OBS is considered for the analysis. We consider that the MT is moving with a speed v. v is assumed to have uniformly distributed in $[v_{min}, v_{max}]$. Therefore, the probability density function (pdf) of v is given by

$$f_v(v) = \frac{1}{v_{max} - v_{min}} \quad v_{min} < v < v_{max} \tag{1}$$

During its course of movement the MT discovers that it is going to move into the subnet served by NBS and hence needs to perform the HMIP registration with the FA serving the NBS. This FA is referred as new FA (NFA). The MT may learn about the possibility of moving into another subnet when the RSS of OBS decreases continuously.

Once the MT discovers that it may enter into the coverage area of NBS, the next challenge is to decide the right time to initiate the HMIP registration procedures with the NFA. The existing link-layer assisted HMIP protocols propose to initiate the HMIP registration when the RSS from the serving BS, i.e., OBS in the above scenario, goes below a fixed threshold value (S_{th}). Below, the performance of these solutions is analyzed.

It is assumed that during the course of its movement when the MT reaches the point P (the distance of P from the boundary is d) as shown in Fig. 10, the RSS from OBS goes below S_{th}. Therefore, when the MT reaches P, the HMIP registration is initiated with the NFA. At this point, the RSS received by the MT from NBS may not be sufficient for the MT to send the HMIP registration messages to NFA through NBS. Hence, the MT may send the HMIP registration messages to NFA through OBS. This is called pre-registration [50]. For a smooth and successful handoff from OBS to NBS, MT's HMIP registration with NFA and link and MAC layer associations with NBS must be completed before the RSS of OBS goes below S_{min}, i.e., before the MT moves beyond the coverage area of OBS.

When the MT is located at point P (as shown in Figure 10), it is assumed that it can move in any direction with equal probability, i.e.,

$$f_\theta(\theta) \quad = \quad \tfrac{1}{2\pi} \qquad -\pi < \theta < \pi \tag{2}$$

with a speed of v that is uniformly distributed in $[v_{min}, v_{max}]$. It is also assumed that MT's direction of motion and speed remain the same from point P until it moves away from the coverage area of OBS. As the distance of P from the boundary of OBS

is not very large, this assumption is realistic. For example for an MT moving at 100 km/h, considering the handoff signaling delay of 2 sec, $d = 50$ m. A vehicle moving at this speed is not quite expected to change speed of direction within a distance of 50 meters. For smaller value of v and handoff delay, d will be much smaller (typically in the order 10-30 meters).

From Fig. 10, it is clear that the need for handoff to NBS arises only if MT's direction of motion from P is in the range [$\theta \in (-\theta_1$,$\theta_1)$], where $\theta_1 = \arctan(\frac{a}{2d})$. Otherwise, the handoff initiation is a false one. Therefore, using (2) the probability of false handoff initiation is

$$\begin{aligned} p_a &= 1 - \int_{-\theta_1}^{\theta_1} f_\theta(\theta) d\theta \\ &= 1 - \frac{2\theta_1}{2\pi} = 1 - \frac{1}{\pi} \arctan\left(\frac{a}{2d}\right) \end{aligned} \tag{3}$$

When the direction of motion of the MT from P, $\beta \in [(-\theta_1, \theta_1)]$, the time it will take to go beyond the coverage area of OBS is given by

$$t = \frac{d \sec \beta}{v}. \tag{4}$$

It is known that the pdf of β is

$$f_\beta(\beta) = \begin{cases} \frac{1}{2\theta_1} & -\theta_1 < \beta < \theta_1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

From (4), t is a function of β, i.e., $t = g(\beta)$ where $g(\beta) = \frac{d \sec \beta}{v}$. Therefore from [62],

$$f_t(t) = \sum_i \frac{f_\beta(\beta_i)}{|\acute{g}(\beta_i)|} \tag{6}$$

where β_i are the roots of the equation $t = g(\beta)$ in $[-\theta_1, \theta_1]$. The equation $t = g(\beta)$ has two roots in the interval $[-\theta_1, \theta_1]$ and for each of these roots $f_\beta(\beta_i) = \frac{1}{2\theta_1}$. Therefore, (6) becomes

$$f_t(t) = \frac{1}{\theta_1 |g'(\beta_i)|} \tag{7}$$

where $g'(\beta)$ is the derivative of $g(\beta)$ given by

$$g'(\beta) = \frac{d \sec\beta \tan\beta}{v} = t\sqrt{\frac{v^2t^2}{d^2} - 1} \tag{8}$$

Using (7) and (8), the pdf of t is given by

$$f_t(t) = \begin{cases} \frac{d}{\theta_1 t\sqrt{v^2t^2-d^2}}, & \frac{d}{v} < t < \frac{\sqrt{\frac{a^2}{4}+d^2}}{v} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The probability of handoff failure is given by

$$p_f = \begin{cases} 1 & \tau > \frac{\sqrt{\frac{a^2}{4}+d^2}}{v} \\ p(t<\tau) & \frac{d}{v} < \tau < \frac{\sqrt{\frac{a^2}{4}+d^2}}{v} \\ 0 & \tau \leq \frac{d}{v} \end{cases} \tag{10}$$

where τ is the handoff signaling delay and $p(t < \tau)$ is the probability that $t < \tau$. When, $\frac{d}{v} < \tau < \frac{\sqrt{\frac{a^2}{4}+d^2}}{v}$, using (82)

$$\begin{aligned} p(t<\tau) &= \int_0^\tau f_t(t)dt \\ &= \int_{\frac{d}{v}}^\tau \frac{d}{\pi t\sqrt{v^2t^2-d^2}}dt \\ &\approx \frac{1}{\theta_1}\arccos\left(\frac{d}{v\tau}\right) \end{aligned} \tag{11}$$

Now using (10) and (11),

$$p_f = \begin{cases} 1 & \tau > \frac{\sqrt{\frac{a^2}{4}+d^2}}{v} \\ \frac{1}{\theta_1}\arccos\left(\frac{d}{v\tau}\right) & \frac{d}{v} < \tau < \frac{\sqrt{\frac{a^2}{4}+d^2}}{v} \\ 0 & \frac{d}{v} \geq \tau \end{cases} \tag{12}$$

In the following subsections, detailed discussion about the performance of the handoff algorithms is presented using the above mathematical formulations.

3.2.0.1 False Handoff Initiation Probability

It is clear from (3) that if an unnecessarily large value for d (hence, corresponding value of S_{th}) is used for handoff initiation, the probability of false handoff initiation increases. This leads to the wastage of limited wireless system resources. Moreover, this increases the load on the network that arises because of the handoff initiation. The relationship between probability of false handoff initiation and d is shown in Figure 11 for different cell size, a. Figure 11 shows that for a particular value of a, the probability of false handoff initiation increases as d increases. It also shows that the problem of false handoff initiation becomes more and more severe when the cell size decreases. The cell size of wireless systems is decreasing so that the capacity and data rate may increase. Hence, in NGWS it is important to select the proper value of d to reduce the false handoff initiation probability.

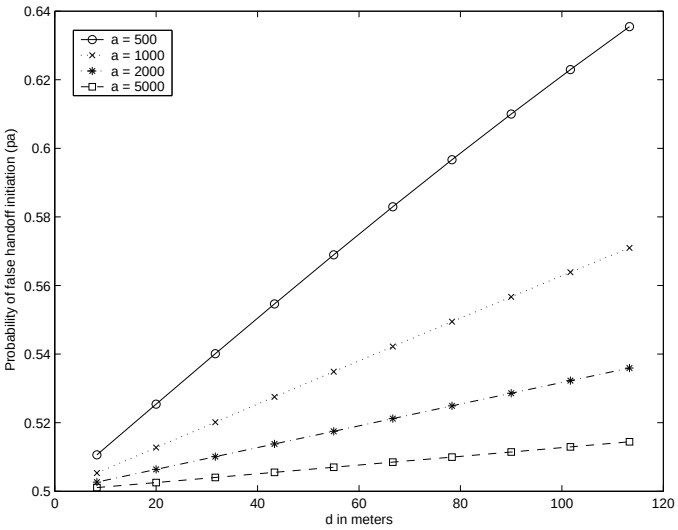


Figure 11: Relationship between false handoff initiation probability and d.

3.2.0.2 Relationship Between Handoff Failure Probability and Speed

When $\frac{d}{v} < \tau < \frac{\sqrt{\frac{a^2}{4}+d^2}}{v}$, (12) shows that if a fixed value of S_{th} (hence a fixed value of corresponding d) is used, the handoff failure probability depends on the speed of the MT. As the speed increases the probability of handoff failure also increases. The relationship between the handoff failure probability and MT's speed is shown in Fig. 12 (a) and Fig. 12 (b) for intra- and inter-system handoff, respectively. These figures show the numerical value of p_f for different values of d (corresponding to different values of S_{th}). Cell size of 1 km is considered for this simulation. As pointed out earlier the main difference between intra- and inter-system handoff is the latency associated with the handoff process. The latency of inter-system handoff is significantly larger than that of intra-system handoff because during an inter-system handoff before HMIP registration authentication and billing procedures are carried out [26] adding extra delay to the handoff process. Moreover, the inter-system HMIP signaling messages are handled by the HA instead of GFA adding extra delay to the signal propagation as the distance of MT from HA is typically large than that of MT from the GFA. Handoff latency of 1 sec and 3 sec are considered for intra- and inter-system handoff procedures, respectively. Figure 12 (a) and Figure 12 (b) show that for a particular value of d, as speed increases, the handoff failure probability increases for both intra- and inter-system handoff. This is because on average the MT requires less time to cross the coverage region of OBS. These figures also show that when a particular value of S_{th} is used p_f becomes higher for inter-system handoff compared to intra-system handoff for any speed value. Therefore, it is not efficient to use the same value of S_{th} for intra- and inter-system handoff. To summarize, this analysis shows that the value of d and therefore the value of S_{th} should be adaptive to the speed of the MT and to the type of handoff to guarantee a desired handoff failure probability.

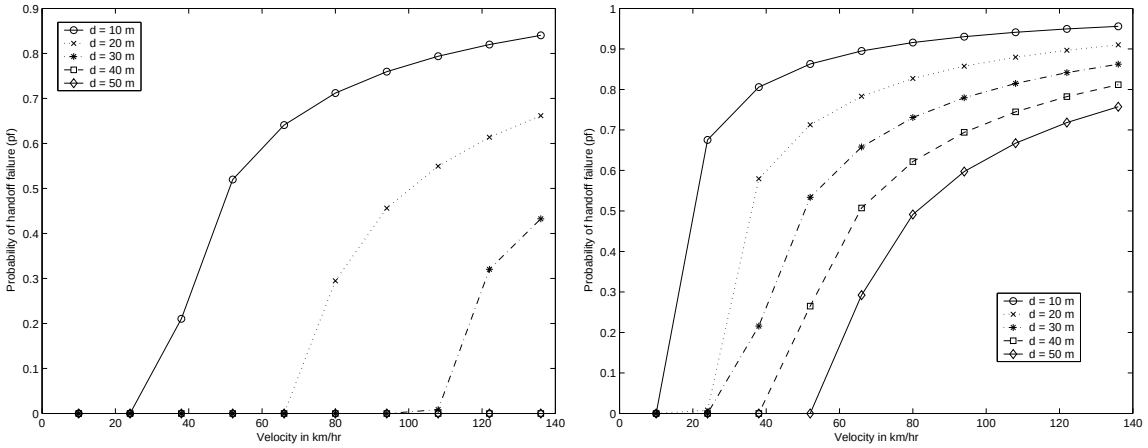


Figure 12: Relationship between handoff failure probability and v: (a) for intra-system handoff with $\tau = 1$ sec, (b) for inter-system handoff with $\tau = 3$ sec.

3.2.0.3 Relationship Between Handoff Failure Probability and Handoff Signaling Delay

As discussed earlier, the handoff signaling latency in case of intra- and inter-system handoff varies based on the network dynamics, e.g., congestion level, wireless link condition, and the location of the user from its home network. Figure 13 shows the relationship between the handoff failure probability and the handoff signaling delay when a fixed value of S_{th}, therefore a fixed value of d is used. The higher value of τ corresponds to the inter-system scenarios and the lower values of τ corresponds to the intra-system handoff scenarios. Figure 13 shows that when a fixed value for S_{th} is used, the handoff failure probability increases as the handoff signaling delay increases. Therefore, to keep the handoff failure probability limited it is essential to predict the handoff signaling delay in advance and accordingly use an adaptive value for S_{th}.

To summarize, the analysis shows that when a fixed value for S_{th} is used, the handoff failure probability increases as the speed of the MT increases (as shown in Figure 12 (a) and Figure 12 (b)). Also, for a fixed value of S_{th} the handoff failure probability increases as the handoff signaling delay increases (as shown in Figure 13). Moreover, the analysis shows that an unnecessarily large value of S_{th} should not be

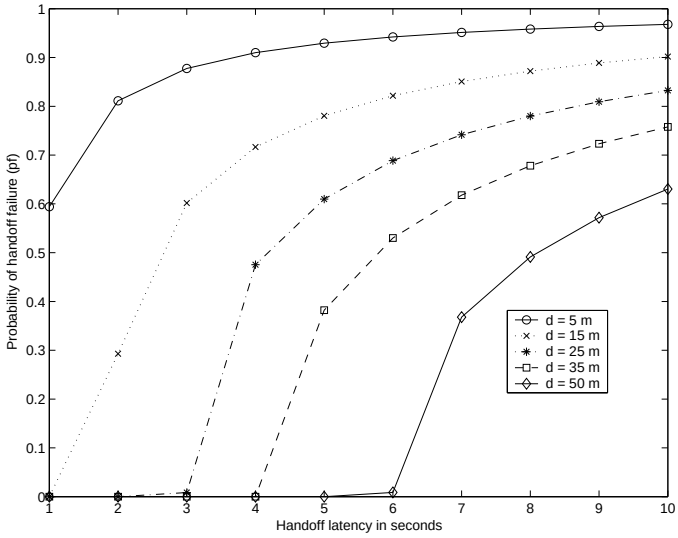


Figure 13: Relationship between handoff failure probability and τ.

used as it increases the probability of false handoff initiation (as shown in Figure 11) and hence, affects the performance of the system negatively. Therefore, we propose the use of adaptive S_{th} for handoff initiation. The exact value of S_{th} will depend on the speed of the user and handoff signaling delay at a particular time. Our objective is to use adaptive S_{th} to limit the handoff failure probability and at the same time to reduce the unnecessary load on the system that arises because of false handoff initiation.

3.3 Cross Layer (Layer 2+ 3) Handoff Management

The analysis in the previous section shows that the performance of intra- and inter-system handoff algorithms depends on the users' speed and handoff signaling delay. Therefore, using speed and handoff signaling delay information, the performance of the existing handoff management protocols (that do not consider the user's speed and network dynamics) can be enhanced to achieve the design goals pointed out in Section 5.1.

In this section, an architecture is proposed to implement handoff management

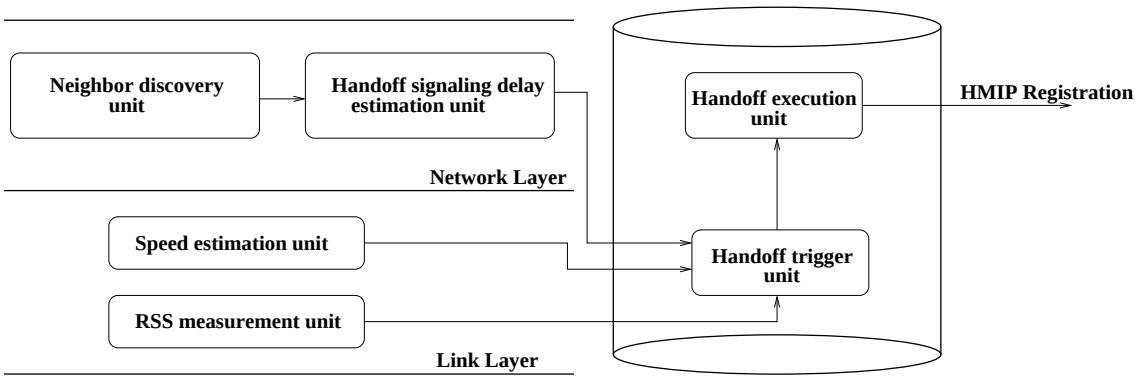


Figure 14: The modules of cross-layer handoff management architecture.

adaptive to the link layer (Layer 2) and network layer (Layer 3) parameters to support enhanced handoff management in NGWS. Afterwards, a handoff management protocol is developed using this architecture. As the proposed handoff management protocol uses information derived from different layers of network protocol stack (e.g., speed information from link layer and handoff signaling delay information from network layer), we call it cross-layer handoff management protocol (CMP). The architecture of our proposed CMP is shown in Fig. A.1 that shows the different modules of CMP. Some of these modules collect the link and network layer information useful for handoff management and the other modules use the information to decide about the appropriate time to initiate and execute the handoff procedures. The modules that collect information include *neighbor discovery unit*, *handoff signaling delay estimation unit* implemented in the network layer; and *speed estimation unit* and *RSS measurement unit* implemented in the link layer. The modules that use the Layer 2 and Layer 3 information to carry out the handoff procedures are *handoff trigger unit* and *handoff execution unit*. The functionalities of these units are as follows.

- The *neighbor discovery unit* assists the MT to learn about the neighboring BSs. It implements network discovery protocols or has interface with the network discovery protocols such as candidate access router discovery (CARD) [49].

- The *handoff signaling delay estimation unit* estimates the delay associated with

intra- and inter-system handoffs. More discussion about the handoff signaling prediction is provided later in this section.

- The *speed estimation unit* estimates the speed of the user using our own algorithm, VEPSD (velocity estimation using the power spectral density of the received signal envelope), proposed in [56]. The maximum Doppler frequency (f_m) is related to the speed (v) of a mobile user, speed of light in free space (c), and the carrier frequency of the received signal (f_c) through

$$v = \left(\frac{c}{f_c}\right) f_m. \tag{13}$$

  VEPSD uses f_m in the received signal envelope to estimate the speed of a mobile user. It estimates f_m using the slope of the power spectral density (PSD) of the received signal envelope. The slope of PSD of receive signal envelope has maximum values at frequencies $f_c \pm f_m$ in mobile environments [56]. VEPSD detects the maximum value of received signal envelope's PSD that corresponds to the highest frequency component ($f_c + f_m$) to estimate f_m. We select this algorithm over other speed estimation algorithms such as [16] [40] because the latter suffer from larger estimation errors [56]. The details of VEPSD is described in Appendix B.

- The *handoff trigger unit* collects the information from the handoff signaling delay estimation unit, speed estimation unit, and RSS measurement unit; and estimates the appropriate time to start the handoff procedures. The details about the estimation of handoff initiation time is discussed in Sec. 3.3.2.

- Finally, the *handoff execution unit* starts the HMIP registration process at the handoff initiation time estimated by the handoff trigger unit.

3.3.1 Operation of CMP

To give further insight into the guidelines behind the operation of CMP, the entire handoff process is sub-divided into the following steps.

3.3.1.1 Neighborhood Discovery

When an MT is served by a BS, it learns about its neighboring BSs using the neighbor discovery unit. The neighboring BSs refer to the BSs that are the immediate neighbor of the serving BS. Some of these BSs may belong to the serving FA, where as others may belong to different FAs. When the MT moves into the coverage of a neighboring BS that belongs to its serving FA the resulting handoff is a link-layer handoff. In this case, the MT uses the existing link-layer handoff algorithms [95] and CMP procedures are not invoked. When the neighboring BS belongs to a different FA under the serving system, the corresponding handoff is an intra-system handoff. Similarly, when the neighboring BS belongs to a different system the resulting handoff is an inter-system handoff. CMP is used for both intra- and inter-system handoffs. Using the neighbor discovery protocol the MT also learns the details of its neighboring BSs such as the IP addresses of the FAs that serve the BSs.

3.3.1.2 Handoff Signaling Delay Estimation

It is difficult to predict which particular BS the MT will move unless the handoff instance is very close. Our objective is to estimate the handoff signaling delay in advance without knowing which particular BS the MT will move. This can be done in many ways. For example techniques such as [43] [48] can be used to estimate the delay between different network entities that are involved in the handoff process and using this information the handoff signaling delay for intra- and inter-system handoff can be estimated. A simple technique is proposed that uses the existing HMIP protocol to estimate the handoff delay. From the neighborhood discovery step the MT learns the BSs and the corresponding FAs involved in a possible intra-

or inter-system handoff. Now the objective is to estimate the signaling delay for these handoffs. To estimate the signaling delay of a possible handoff to a particular neighboring BS, the MT sends the HMIP registration messages to the GFA with an invalid *Mobile-GFA Authentication Extension* if the corresponding handoff is intra-system. Otherwise, it sends the HMIP registration messages to the HA with an invalid *Mobile-Home Authentication Extension* if the corresponding handoff is inter-system. The objective of using invalid *Authentication Extension* is to just learn the handoff signaling delay without changing the mobility binding at GFA or HA. When GFA or HA receives the HMIP registration messages and learns the presence of the invalid *Authentication Extension*, they return the HMIP *Registration Reply* with appropriate code [63] that signifies mobile node (MN) failed authentication. The handoff signaling delay is estimated by comparing the time difference between the transmission time of HMIP registration request and the reception time of HMIP registration reply. This way the MT predicts the handoff signaling delay in the event of its movement to the BS. Similarly, it also learns the signaling delay of the associated handoffs to the other neighboring BSs. The proposed handoff signaling prediction technique introduces extra signaling overhead to the system. However, we advocate its use because of its simplicity. Moreover, this technique can be implemented using the existing HMIP protocol, hence no extra implementation is required. Considering the significant performance improvement (as discussed in Section 3.4), this signaling overhead is tolerable. If this extra signaling overhead is undesirable for a particular deployment scenario, then the existing delay estimation algorithms [43] [48] can be used to estimate the handoff signaling delay.

It may be noted that the prior estimation of handoff signaling delay captures different factors such as the type of handoff to be performed, the location of the MT from its home network, and the load on the network. For example, if the handoff is intra-system then there are fewer signaling messages exchanged [26] [35], hence,

the handoff delay is less compared to inter-system handoff. Similarly, if either the user is far from the home network or the network is experiencing higher load the handoff signaling delay increases. This shows that by estimating the handoff signaling delay in advance, CMP eliminates the adverse effect of the above parameters on the performance of the handoff management protocols.

3.3.1.3 Handoff Anticipation

When the RSS of the serving BS measured by the RSS measurement unit decreases continuously, a handoff is anticipated. Moreover, using the existing movement prediction techniques [12] [42] the MT learns the next BS it is going to move. Then the handoff trigger unit learns about the signaling delay for that particular BS from the handoff signaling delay estimation unit. Note that the objective of estimating the handoff delays for each neighboring BSs in advance is to avoid estimating the delay after learning which particular BS the MT will move. This eliminates the latency associated with handoff signaling delay estimation if it were to be done after the handoff anticipation. The extra delay associated with the signaling delay estimation may lead to delay in handoff initiation resulting in an unsuccessful handoff [50].

3.3.1.4 Handoff Initiation

Once the MT learns the BS that it is going to move, the next challenge is to estimate the right time to start the HMIP registration. The handoff trigger unit uses the speed and handoff signaling delay information to estimate the value of S_{ath} as discussed in Sec. 3.3.2. When the value of RSS goes below S_{ath}, the handoff trigger unit sends a trigger to the handoff execution unit to start the HMIP handoff procedures.

3.3.1.5 Handoff Execution

When the handoff execution unit receives the handoff trigger from the handoff trigger unit it starts the HMIP registration. Once the HMIP registration is completed

the mobile is switched to the new BS. The MT keeps its HMIP registration for a specified time period with the old BS to avoid ping-pong effect during handoff. This is implemented by using the simultaneous binding option of HMIP protocol. The MT binds the CoA of the old FA (OFA) and new FA (NFA) at the GFA in case of intra-domain handoff and at the HA in case of inter-domain handoff. Therefore, the GFA and HA forwards packets destined at both the CoAs during this time interval. It may be noted that in case of inter-system handoff these two CoAs may belong to two different network interfaces when the MT moves between networks employing different wireless access technologies. Therefore, the multiple interfaces of the MT can be used to reduce the ping-pong effect during inter-system handoff. If the MT returns to the old BS during this time period, there is no need to carry out the HMIP handoff procedures again. If the MT does not return to the old BS with in this time duration, it deregisters from the old BS.

The operation of CMP is summarized in Fig. 15. First the MT learns about its neighborhood using the neighbor discovery protocol. Then it determines the type of handoff (e.g., link-layer handoff, intra-system, or inter-system handoff) in the event of its movement to these BSs. When the MT learns about the neighboring BSs, the handoff signaling delay unit estimates the signaling delay associated with the handoff to the neighboring BSs that would result in either intra- or inter-system handoff. The RSS monitoring unit starts to monitor the RSS of the serving BS and anticipates a handoff when this RSS decreases continuously. The MT learns about the next BS using the existing movement detection techniques [12] [42]. Then one of the following three steps is carried out.

- If the associated handoff to the next BS is an link-layer handoff, the existing link-layer handoff algorithms [95] are used and CMP does not take any action.

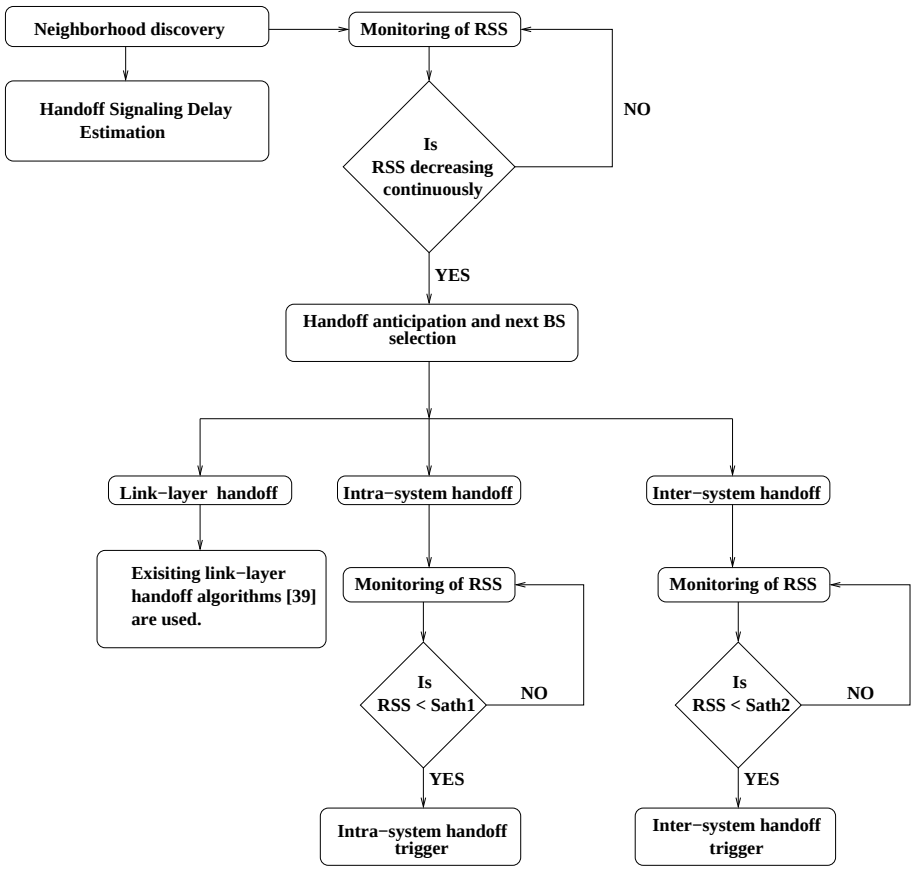


Figure 15: Flow diagram of CMP operation.

- If the associated handoff to the next BS is intra-cell handoff, the handoff trigger unit estimates the value of dynamic RSS threshold, S_{ath1} as discussed in Sec. 3.3.2. When the RSS of the current BS drops below S_{ath1}, if the RSS of the next BS is sufficient, then MT starts HMIP handoff procedures with the next BS directly. Otherwise, it carries out HMIP registration with the next BS through the serving BS [50].

- If the associated handoff to the next BS is inter-system handoff, the steps are similar to that of intra-system handoff. The dynamic RSS threshold corresponding to inter-system handoff is referred as S_{ath2} in Fig. 15. The HMIP inter-system handoff procedures are carried out when the RSS of the serving BS drops below S_{ath2}.

The different functionalities of CMP can be implemented either at the MT or at the network side. Accordingly, the handoff management using CMP can be classified into mobile assisted network controlled handoff (MAHO) or network assisted mobile controlled handoff (NAHO). In case of MAHO the MT implements the speed estimation, RSS measurement and handoff signaling delay units of CMP. The network implements the handoff trigger unit that collects the information about users speed and handoff signaling delay measurement from MT and estimates the numerical value of dynamic RSS threshold (S_{ath}). When the RSS of the MT goes below S_{ath}, the network generates the handoff trigger for intra- or inter-system handoff referred to as *HT_intra* or *HT_inter*, respectively. Then the network initiates the handoff procedures by sending *Proxy Router Advertisement* message [50] to the MT. On the other hand, in NAHO, the network assists the MT with the neighborhood discovery and in the selection of next BS. The MT calculates the dynamic value of RSS threshold (S_{ath}) and generates the handoff triggers *HT_intra* or *HT_inter* and initiates the handoff procedures when the RSS of serving drops below (S_{ath}) by sending *Proxy Router Solicitation* message [50] to the new FA. The timing diagrams of intra- and inter-system handoff using CMP for both MAHO and NAHO scenarios are shown in Figure 16 and Figure 17, respectively. In NGWS, there exist two types of intra-system handoff scenarios and four types of inter-system handoff scenarios depending on the cell-size of the wireless systems. The intra-system handoff can be between two cells of a macro-cellular system, referred as macro-intra handoff (Intra_MA_HO) or between two cells of a micro-cellular system, referred as micro-intra handoff (Intra_MI_HO).

Similarly, the inter-system handoff can be one of the following four types.

- Inter-system handoff between one macro-cellular system to another macro-cellular system, referred as macro-inter handoff (Inter_MA_HO).

- Inter-system handoff between one macro-cellular system to another micro-cellular

system, referred as macro-micro-inter handoff (Inter_MAMI_HO).

- Inter-system handoff between one micro-cellular system to another micro-cellular system, referred as micro-inter handoff (Inter_MI_HO).

- Inter-system handoff between one micro-cellular system to another macro-cellular system, referred as micro-macro-inter handoff (Inter_MIMA_HO).

It may be noted that micro-cellular systems are usually overlapped with the macro-cellular systems. Therefore, during a macro-cell to micro-cell inter-system handoff (Inter_MAMI_HO), there is no handoff failure as the macro-cell coverage is always available.

3.3.2 Handoff Initiation Time Estimation

The handoff trigger unit determines the value of adaptive RSS threshold (S_{ath}) to initiate the HMIP handoff procedures using the speed and handoff signaling delay information. S_{ath} is estimated as follows. First, we calculate the value of d for a desired value of p_f using

$$p_f = \frac{1}{\theta_1} \arccos\left(\frac{d}{v\tau}\right) \; ; \; \frac{d}{v} < \tau < \frac{\sqrt{\frac{a^2}{4}+d^2}}{v} \tag{14}$$

where v is the speed of the MT and τ is the handoff signaling delay. The derivation of (14) is carried out in Section 3.2. (14) is a non-linear equation of d. A closed form expression may not be always possible. However, an approximate value of d can be calculated using

$$p_f = \frac{\cos^{-1}\left(\frac{d}{v\tau_1}\right)}{\tan^{-1}\left(\frac{a}{2d}\right)} = \frac{\frac{\pi}{2} - \frac{d}{v\tau_1}}{\frac{\pi}{2} - \frac{2d}{\sqrt{4d^2+a^2}}} \tag{15}$$

Moreover, numerical methods can be used to calculate d. The Bisection numerical method [61] is used to solve for d (it takes only few iterations to calculate d when the

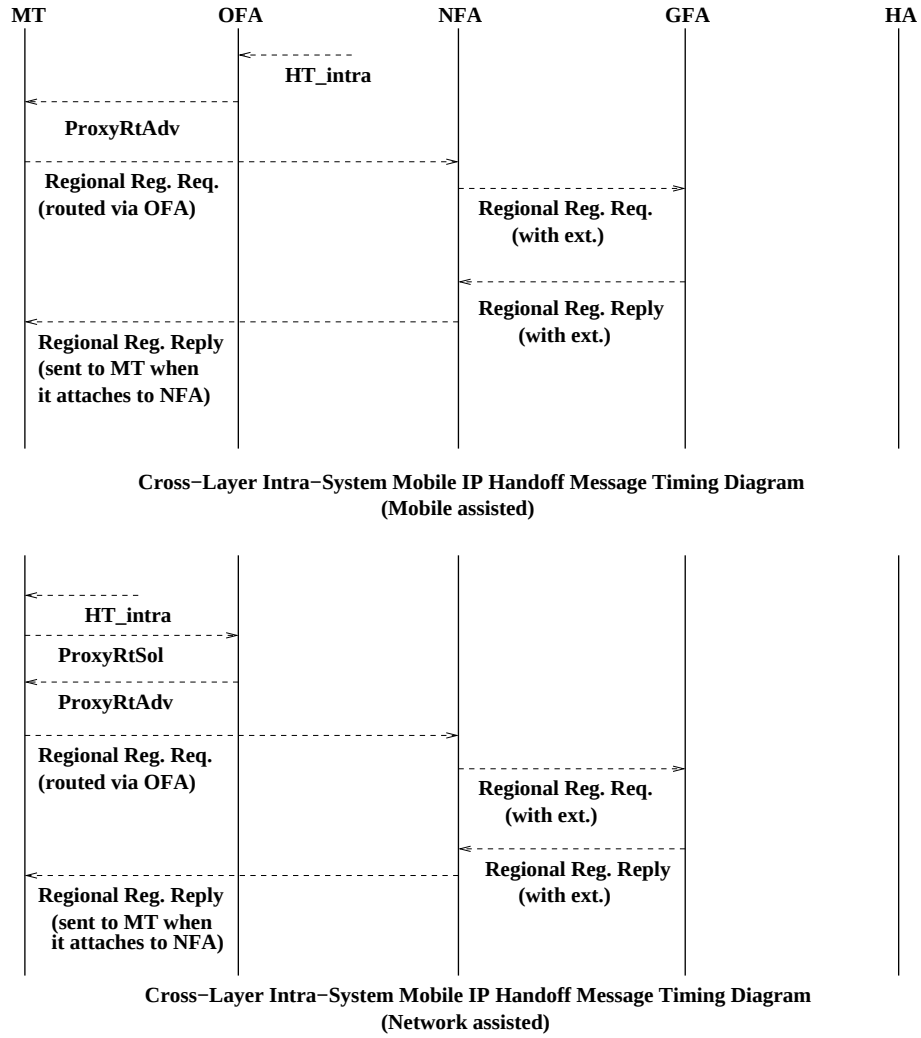


Figure 16: Timing diagram for cross-layer intra-system HMIP handoff.

Bisection method [61] is used. Hence, calculation of d does not have much computational complexity and can be easily implemented at the MT or at the network side, e.g., the BS or FA). Once d is calculated, the corresponding value of S_{ath} is calculated using the path loss model and the cell size of the serving BS. We use the path loss model given by [77]

$$P_r(x) = P_r(d_0)\left(\frac{d_0}{x}\right)^{\alpha} + \epsilon \tag{16}$$

where x is the distance between the base station and MT, $P_r(d_0)$ is the received power at a known reference distance, which is in the far field of the transmitting antenna. Typical value of d_0 is 1 km for macrocells, 100 m for outdoor microcells, and 1 m for

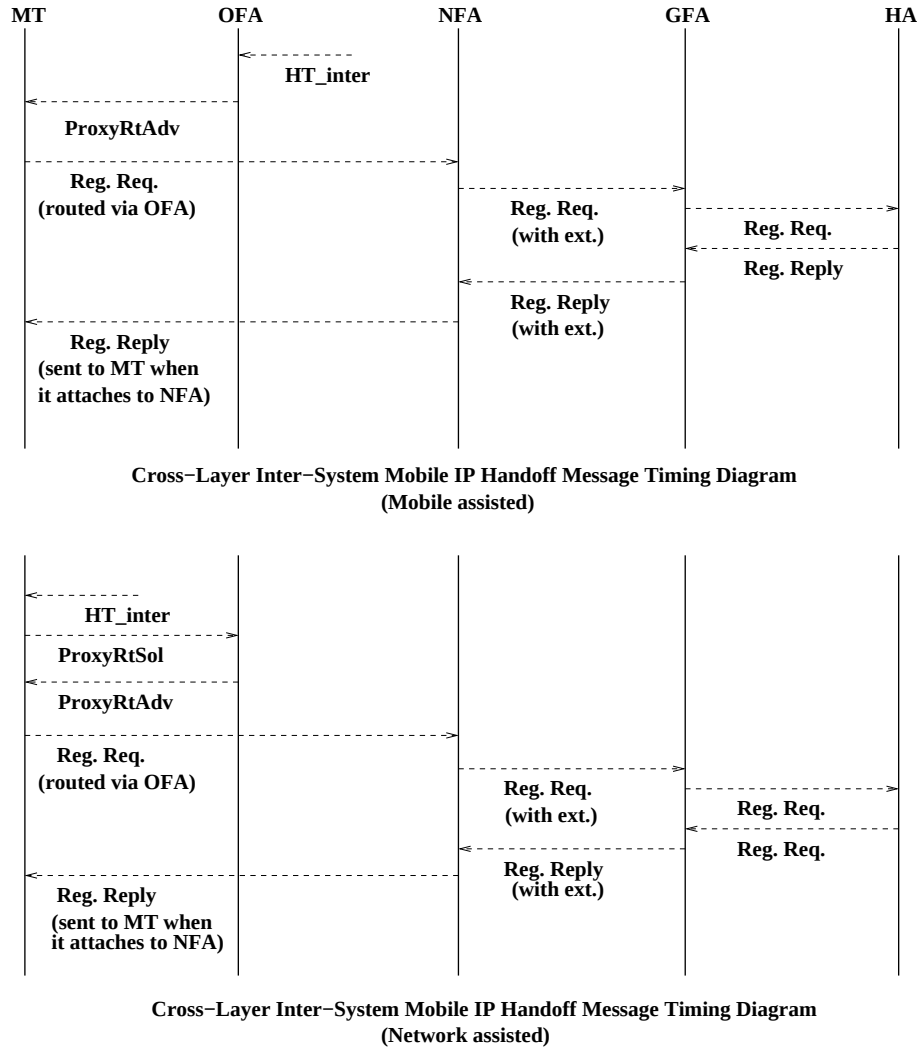


Figure 17: Timing diagram for cross-layer inter-system HMIP handoff.

indoor pico cells [77]. The numerical value of $P_r(d_0)$ depends on different factors such as frequency, antenna heights, and antenna gains. α is the path loss exponent. The numerical value of α is dependent on the cell size and local terrain characteristics. The typical value of α ranges from 3 to 4 and 2 to 8 for a typical macro-cellular and micro-cellular environment, respectively. ϵ is a zero-mean Gaussian random variable that represents the statistical variation in $P_r(x)$ caused by shadowing. Typical standard deviation of ϵ is 8 dB [77]. Its actual value depends on the cell size. Using (62), the RSS value when the MT is at a d distance from the cell boundary is given by

$$S_{ath} = 10 \log_{10}[P_r(a-d)] \tag{17}$$

The value of S_{ath} is defined as S_{ath1} and S_{ath2} for intra-system and inter-system handoff, respectively, in Figure 15. Once S_{ath} is calculated, the handoff trigger unit monitors the RSS from the serving BS and when the RSS goes below S_{ath}, it sends a trigger to handoff execution unit to start the HMIP registration procedures.

3.4 Performance Evaluation of CMP

In this section, the performance evaluation of CMP is carried out. For simulation the following scenarios and parameters are considered: a macro cellular system with cell size of $a = 1$ km, a micro cellular system with cell size of $a = 30$ meters, macro-cell reference distance $d_0 = 100$ m, micro-cell reference distance $d_0 = 1$ m, standard deviation of shadow fading parameter $\epsilon = 8$ dB and path-loss co-efficient $\alpha = 4$ for macro and micro-cells. We assume that the target handoff failure probability is $p_f = 0.02$. We consider that the maximum value of users' speed in micro-cellular and macro-cellular system are 14 km/h and 140 km/h, respectively. Moreover, we assume that the value of S_{min} is -64 dBm.

3.4.1 Relationship between S_{ath} and Speed

The relationship between S_{ath} and MT's speed (v) for different values of handoff signaling delay (τ) is analyzed. For different values of v, first the required value of d is determined using (14). Then using (61), the required value of S_{ath} is calculated. Figure 18 (a) shows the relationship between S_{ath} and v for different value of τ when the serving BS (OBS) belongs to a micro-cellular system. Figure 18 (b) shows the similar results when the OBS belongs to a macro-cellular system. It may be noted that the results shown in Figure 18 (a) are applicable for Intra_MI_HO, Inter_MI_HO, and Inter_MIMA_HO, whereas the results shown in Figure 18 (b) are applicable for

Intra_MA_HO, Inter_MA_HO, and Inter_MAMI_HO. Figure 18 (a) and (b) show that for particular value of τ, the value of S_{ath} increases as MT's speed increases. This implies that for a MT with high speed, the handoff initiation should start earlier compared to a slow moving MT to guarantee the desired handoff failure probability to users independent of their speed. Slight variation in the S_{ath} estimation is introduced because of the error in handoff signaling delay estimation and the effect of shadow fading. Figure 18 also shows that S_{ath} increases as τ increases. This is because when τ is high the handoff must start earlier compared to when τ is small. The lower and higher values of τ correspond to intra- and inter-system handoff, respectively. Therefore, CMP calculates S_{ath} that is adaptive to v and τ.

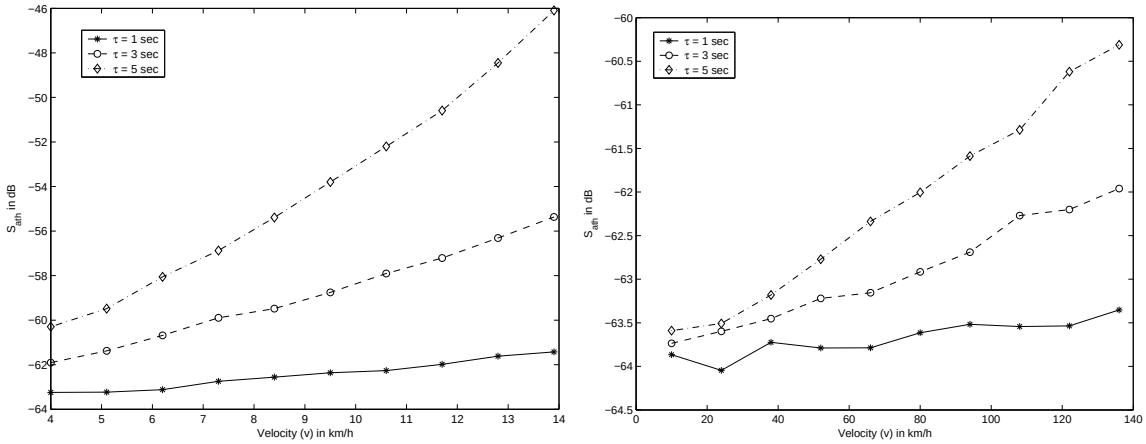


Figure 18: RSS threshold (S_{ath}) for different speed values when the serving BS (OBS) belong to a : (a) micro-cellular system, (b) micro-cellular system.

3.4.2 Handoff Failure Probability of CMP

To determine the handoff failure probability of CMP, we investigate the handoff failure probability of different types of intra- and inter-system handoff and compare that with the handoff failure probability of the fixed RSS threshold based handoff protocols [94]. To analyze the handoff failure probability, the required value of S_{ath} is calculated using the speed and handoff signaling delay information. Then this S_{ath} value is

used to initiate the HMIP handoff and determine the handoff failure probability. Figure 19 shows the handoff failure probability of CMP and the existing fixed RSS threshold based handoff algorithms, for different values of speed when the serving BS belongs to a micro-cellular system. Figure 19 (a) shows the numerical value of p_f for Intra_MI_HO, whereas Figure 19 (b) shows the numerical value of p_f for Inter_MI_HO and Inter_MIMA_HO. These figures show that when the speed of the MT is known, 70% to 80% reduction in handoff failure probability is achieved in CMP compared to fixed RSS based handoff algorithms [94]. It also shows that when CMP is used the probability of handoff failure (p_f) is independent of the speed. On the other hand, for fixed RSS threshold based algorithms p_f depends on the numerical value of S_{th}. Comparison of Figure 19 (a) and Figure 19 (b) shows that for a particular value of fixed RSS threshold the numerical value of p_f is different for intra- and inter-system handoff. This shows that the handoff protocols need to be adaptive to the type of handoff. CMP implements this by learning the neighboring BSs and then determining the type of handoff in case of MT's movements to those BSs. Figure 20 (a) and Figure 20 (b) show the similar results when the serving BS belongs to a macro-cellular system.

3.4.3 CMP Performance for Different Signaling Delay

Figure 21 shows the handoff failure probability of CMP for different values of handoff signaling delay (τ). The results show that unlike the fixed RSS based handoff protocols, p_f remains independent of τ in case of CMP. This is because CMP estimates τ and uses it for the calculation of dynamic RSS threshold. Figure 21 shows that 70-80 % reduction in p_f is achieved in case of CMP compared to the fixed RSS based handoff protocols. The lower and higher values of τ correspond to intra- and inter-system handoffs, respectively. Therefore, by incorporating the estimated value of τ into dynamic RSS the p_f is limited to the desired value irrespective of users speed

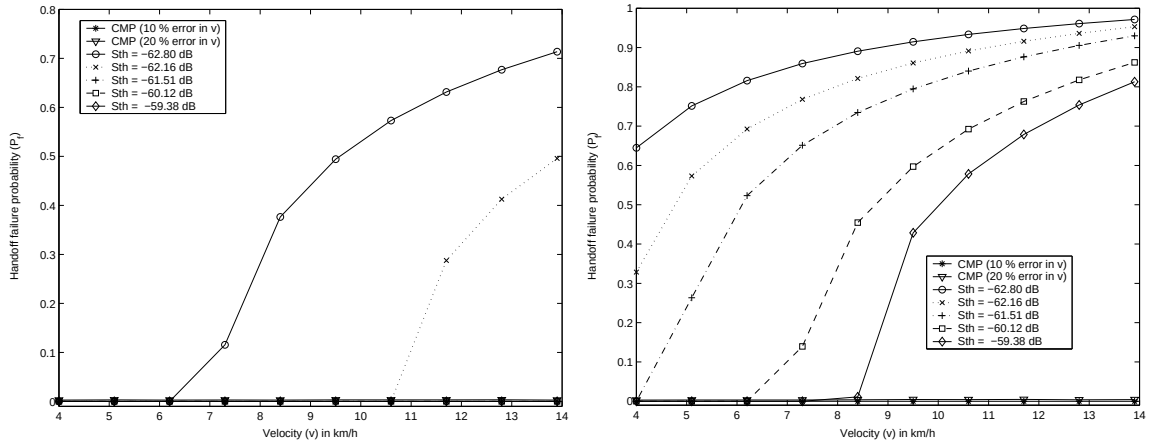


Figure 19: RSS threshold (S_{ath}) for different speed values when the serving BS (OBS) belong to a micro-cellular system: (a) intra-system handoff scenario (b) inter-system handoff scenario.

and variation of handoff signaling delay.

3.4.4 Fixed vs. Adaptive Value of RSS Threshold

The use of adaptive RSS threshold initiates the handoff procedures in such a way that just enough time is there for the successful execution of the handoff. Therefore, an adaptive value of RSS threshold (S_{ath}) avoids too early or too late initiation of the handoff process. The former limits the value of handoff failure probability. The later ensures that handoff is carried out smoothly. Thus, CMP optimizes the false handoff initiation probability and handoff failure probability. We consider that when the fixed value of RSS threshold S_{th} is used, it is calculated such that the user with highest speed is guaranteed the desired value of handoff failure probability (p_f). Figure 22 (a) and Figure 22 (b) show the comparison of the false handoff initiation probability of CMP with the fixed RSS threshold based algorithms [94] when the serving BS belong to a micro-cellular system and macro-cellular system, respectively. These figures show that the false handoff initiation probability of CMP is 5 % to 15 % less compared to the fixed RSS threshold based algorithms [94]. Thus, CMP achieves up to 15 % reduction in the cost associated with false handoff initiation.

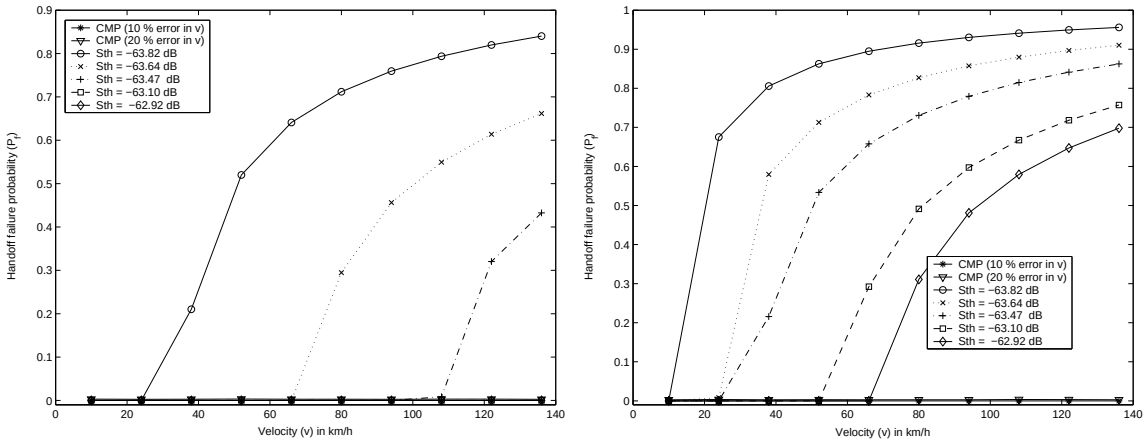


Figure 20: RSS threshold (S_{ath}) for different speed values when the serving BS (OBS) belong to a macro-cellular system: (a) intra-system handoff scenario (b) inter-system handoff scenario.

3.5 Summary

In this research, discussions about the different types of handoff in next generation wireless systems and the recent trend of link layer assisted mobility management protocol design is presented first. Then, the performance of mobility management protocols that use a fixed value of RSS threshold (S_{th}) to initiate the handoff process is analyzed. Through this analysis, it is observed that when a fixed value of S_{th} is used, the handoff failure probability increases when either speed or handoff signaling delay increases. Using the insights from this analysis, a cross-layer mobility management protocol called CMP is proposed. CMP estimates users' speed and predicts the handoff signaling delay of possible handoffs. CMP uses this information to estimate the appropriate instance for handoff initiation. Performance analysis and simulation results show that CMP significantly enhances the performance of both intra- and inter-system handoffs. CMP also significantly reduces the cost associated with the false handoff initiation because it achieves lower false handoff initiation probability.

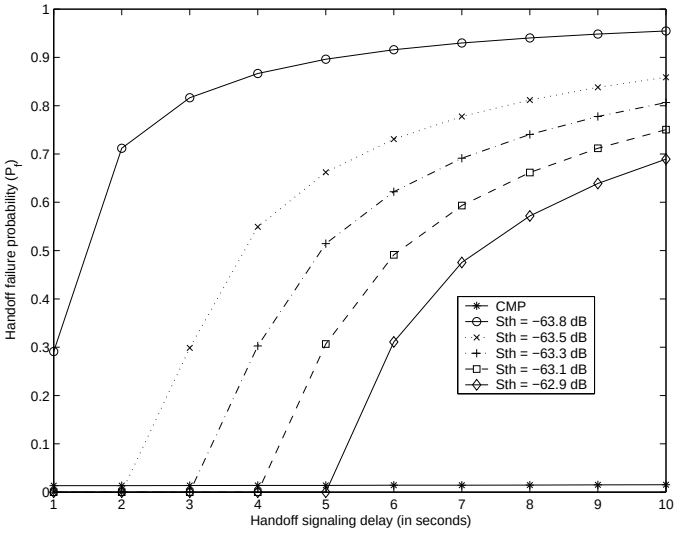
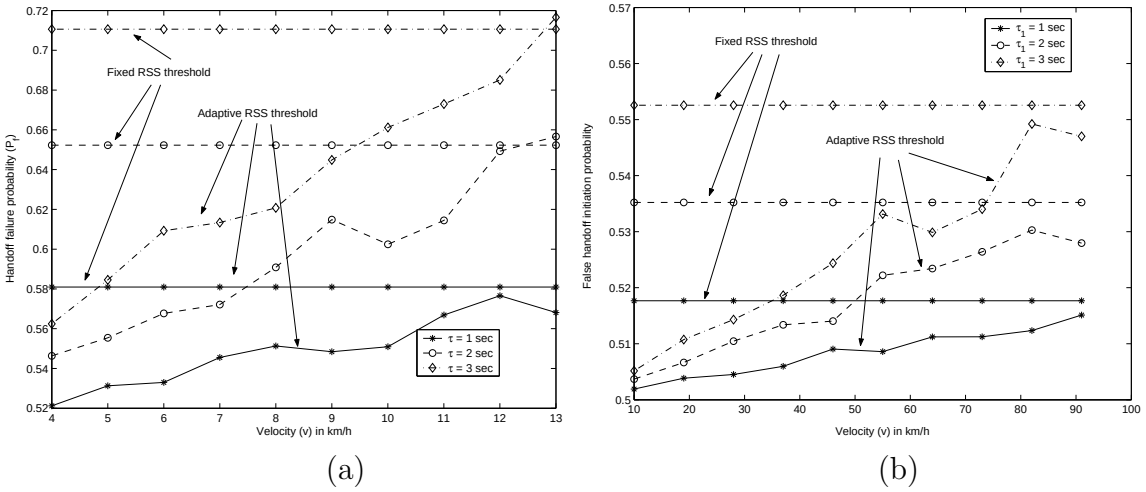


Figure 21: RSS threshold (S_{ath}) for different speed values.



(a) (b)

Figure 22: Probability of false handoff initiation when the serving BS (OBS) belong to a: (a) micro-cellular system (b) macro-cellular system.

CHAPTER IV

PERFORMANCE ANALYSIS OF HANDOFF TECHNIQUES BASED ON MOBILE IP, TCP-MIGRATE, AND SIP

4.1 Introduction

Next-Generation Wireless Systems (NGWS) integrate the existing wireless networks such as wireless local area networks (WLANs), third generation (3G) cellular networks, and satellite networks to realize a unified wireless communication system that has the best features of the individual networks to provide ubiquitous "always best connection" [34] to the mobile users [13]. In NGWN, mobile users are connected to the best available networks that suit their service requirements and switch between different networks based on their service needs. Therefore, it is required that a mobility management protocol supports mobility across heterogeneous access networks. The link layer mobility management protocols alone cannot be used in NGWS because of their inherent scope limitation to a single wireless access technology [13]. Because of intrinsic technology heterogeneity of different wireless networks, mobility management protocols supporting mobility outside the scope of a particular access technology are suitable for NGWS. These include mobility management protocols operating from network, transport, and application layers.

In NGWS, there will be different types of applications, e.g., voice, real and non-real time data, and multimedia services, which have different requirements in terms of *handoff latency*, *packet loss during handoff*, *end-to-end delay*, and *transport-layer transparency*. Based on their mobility management requirements we classify these applications into the following categories.

- **Class A Applications:** TCP or UDP applications that are short lived and

originated by a mobile node (MN) such as Domain Name Service (DNS) resolution [46] [73]. Here the Correspondent Node (CN) (usually a server) typically resides in the fixed backbone network and has a permanent IP address. We can assume that the MN knows about CN's IP address in advance. Since every Internet packet includes the IP address of the sender, the CN learns about the IP address of the MN from the first IP packet that it receives from the MN. As these applications are short lived (most are over in seconds from the initial service request by the client [46], i.e., MN in this case) there is no need for handoff support. If the transaction time happens to coincide with the handoff time, it is always possible to restart the transaction after the handoff [46]. As the transactions are initiated by the MN, there is no need for the CN to learn about the current location of the MN. Therefore, these applications do not require location or handoff support.

- **Class B Applications:** TCP applications that are long lived and originated by an MN such as web browsing and telnet sessions. These applications do not require location support as the MN initiates the connection. However, as they are long lived, they require handoff support as they may stay active over several cell transition instances. Therefore, these applications do not require location support but require handoff support.

- **Class C Applications:** TCP applications that are long lived and terminated at an MN such as telnet sessions. In this case, the originator of the application needs to learn the IP address of the MN before it can start the connection. Therefore, location support is required. Moreover, as these applications are long lived, handoff support is required. Thus, such applications require both location and handoff support.

- **Class D Applications:** UDP applications that are long lived and originated

by an MN such as mobile telephony where MN is the calling party. These applications require only handoff support.

- **Class E Applications:** UDP applications that are long lived and terminated at an MN such as mobile telephony where MN is the called party. In this case, the originator of the application needs to learn the IP address of the MN before it can start the connection. Therefore, location support is required. Moreover, as these applications are long lived, handoff support is required. Thus, these applications require both location and handoff support.

As Class A applications do not require location or handoff support, we do not consider these applications in this work. Class B, Class C, Class D, and Class E applications require handoff support. Therefore, it is essential that these applications remain transparent to the handoffs. The level of transparency to handoffs that these applications can achieve, depends on the mobility management protocol used to carry out the handoff. The effect of handoffs on these application classes can be specified in terms of the following parameters.

- *Handoff latency:* This is the time duration between handoff initiation and handoff completion. Real time applications using real-time transport protocol (RTP) over UDP such as Internet telephony and multimedia applications that belong to Class D and Class E require minimum handoff latency.

- *Packet loss during handoff:* Class D and Class E applications run over UDP. As UDP is not a reliable protocol, the packets that are lost during the handoff can not be recovered. Thus, Class D and Class E applications experience packet loss during handoffs. Class B and Class C applications run over TCP. As TCP is a reliable protocol, the packets that are lost during a handoff are recovered through TCP's retransmission mechanism. Therefore, there is no packet loss during a handoff for Class B and Class C applications.

- *Throughput degradation time:* For Class B and Class C applications that use TCP as the transport layer protocol the packets that are lost during a handoff trigger slow start mechanism of TCP leading to throughput degradation. The throughput degradation time should be kept minimum.

- *End-to-end delay:* Some applications require that the communicating hosts exchange packets directly without the intervention of any other network entities. This ensures that the end-to-end delay is minimum. It may be noted that if the mobility management protocol implements redirection of packets such as Mobile IP, then the end-to-end delay increases significantly. Class D and Class E applications that are real-time in nature require low end-to-end delay.

- *Transport-layer transparency:* Applications running over TCP require that if the transport layer connections are broken during a handoff, there should be a mechanism to resume them in such a way that applications remain transparent to the handoff. These include Class B and Class C applications. Therefore, mobility management protocols that hide the modifications of the IP-address of the mobile host upon handoff such as Mobile IP and TCP-Migrate are appropriate for these applications.

- *Security:* A particular application may have different levels of security requirements in different network environments. For example, while communicating inside a home network domain, an application does not require strict security mechanisms. On the other hand, while in a foreign domain or while communicating with CNs that are in foreign domains the same application may require strict security mechanisms. Thus, security is important for all classes of applications.

The above analysis shows that different classes of applications have different expectations from a mobility management protocol. In the next section, we discuss the

existing mobility management protocols and carry out their qualitative performance evaluation with respect to the above handoff performance metrics.

4.2 Qualitative Handoff Performance Analysis of Existing Mobility Management Protocols

Mobility management protocols operating from different layers such as link layer [12] [52], network layer [63], transport layer [73], and application layer [87] are proposed in the literature [8] [13] the last several years. Below we discuss these protocols and contrast them with respect to the handoff performance parameters discussed in the previous section.

4.2.1 Link layer (Layer 2) mobility management protocols

Link layer mobility management protocols focus on the issues related to inter-system roaming between heterogeneous access networks with different radio technologies and different network management techniques [13]. When an MN roams from one wireless access network to another which supports the same air interface and the same mobile application part (MAP), services are provided seamlessly. However, when the MAPs in the two systems are different, additional entities and signaling traffic are required during MN's handoff between these two systems [13]. The user mobility profile (UMP) is used in [12] to support enhanced mobility management. The concept of inter-system boundary cells are used in [52] to prepare the users for a possible inter-system handoff in advance. Thus, significant reduction of inter-system handoff failure probability is achieved. The performance of the link layer mobility protocols is summarized as follows.

- The inter-system handoff latency is high because several functions such as *format transformation and address translation*, *user profile retrieval*, *signaling message transmission and connection setup*, *mobility information related to inter-system movement recording*, *QoS negotiation*, and *authentication between systems* are carried during an inter-system handoff [13].

- The large value of handoff latency results in higher packet loss during inter-system handoff.

- After the inter-system handoff, an MN communicates with the new system without the need for any redirection agent. Thus, the end-to-end delay requirement of the applications is respected.

- Since an MN communicates with a new address in the new system, a transport layer connection has to be re-established after inter-system handoff. Therefore, link-layer mobility management protocols are not transparent to TCP and UDP applications.

- As authentication is carried out during an inter-system handoff, these handoffs are secure.

4.2.2 Network layer (Layer 3) mobility management protocols

Mobile IP [63], which is a network layer mobility management protocol, is proposed to support mobility in IP-based networks. Mobile IP forwards packets to mobile users that are away from their home networks using IP-in-IP tunnels [63]. Therefore, an additional tunnel state is introduced into the network. IP-in-IP tunneling introduces significant overhead into the data packets. Moreover, Mobile IP suffers from problem of triangular routing, high global signaling load, and high handoff latency [13]. The performance of the Mobile IP protocol is summarized below.

- Mobile IP registration introduces significant amount of latency during handoff. Hierarchical Mobile IP [35] and other micro-mobility protocols such as Cellular IP [82], IDMP [55], and HAWAII [67] reduce the handoff latency by introducing another layer of hierarchy to the base Mobile IP architecture to localize the signaling messages to one domain.

- The large value of Mobile IP latency results in significant packet losses during a handoff.

- Mobile IP triangular routing results in path asymmetry between a CN and an MN. Additional delay is introduced from the CN to MN path because of packet redirection through the home agent (HA). Measurements in [98] show that Mobile IP increases the end-to-end delay by 45% within a campus (from a CN to an MN), which can be expected to increase further in wide area networks. This is not acceptable for delay sensitive applications [87].

- Through packet redirection during handoff, Mobile IP hides the change of IP address from the applications. Therefore, Mobile IP handoff is transparent to the applications and the transport layer connections are kept intact during a handoff.

- Authentication of Mobile IP registration messages is carried out as a part of the Mobile IP registration [24]. Thus, Mobile IP handoff is secure.

4.2.3 Transport layer (Layer 4) mobility management protocols

Using transport layer mobility, a TCP peer can suspend an open connection and reactivate it from another IP address. This reactivation of the TCP connection is carried out in such a way that the applications can continue to use an established TCP connection across a handoff [73]. Transport layer mobility management protocols are proposed to support mobility between networks without introducing any

state into the individual networks. Thus, these leave the network untouched and still allow roaming between networks, thereby preserve the stateless nature of the Internet and other IP-based networks. As a result the transport layer mobility makes the IP networks robust and resilient. However, the transport layer mobility protocols require modifications to the transport layer of the network protocol stack. Therefore, to implement transport layer mobility the existing applications have to incorporate the required changes. This may be expensive considering the large number of applications that are existing in today's Internet world. An architecture called MSOCKS is proposed in [51] for transport layer mobility. An end-to-end approach for transparent layer mobility across is proposed in [41]. TCP-Migrate is proposed in [73] to support end-to-end transport layer mobility management. Moreover, work is going on in the Internet Engineering Task Force (IETF) to modify the Stream Control Transmission Protocol (SCTP) [76] to allow it to dynamically change endpoint addresses in the midst of a connection [29] [39]. Performance of transport layer mobility protocols is summarized as follows.

- Since only the communicating end points are involved in the handoff process, the latency is often lower than that of Mobile IP [73]. It may be noted that the use of the third party such as a HA in case of Mobile IP increases the handoff latency.

- During a transport layer mobility, a TCP connection maintains the same control block and state including the sequence number space [73]. Therefore, any necessary retransmissions can be requested in the standard fashion. Thus, the packets that are lost during the handoff can be recovered. Therefore, transport layer mobility management protocols can be designed to realize zero packet losses during a handoff.

- Since there is no packet redirection, the path between the communicating hosts

(i.e., the MN and the CN) is symmetric. Therefore, the end-to-end delay does not increase after handoff. This is in contrast to network layer Mobile IP handoff where due to triangular routing the end-to-end delay increases in the CN to MN path when the MN is away from its home network.

- As a transport layer connection is reactivated upon handoff, the applications remain transparent to mobility.

- Authentication is implicitly included during a transport layer mobility making it highly secure. The end-to-end approach to mobility simplifies the trust relationships required to securely support end-host mobility compared to the network layer approaches such as Mobile IP [73]. Since no third parties are required or even authorized to speak on the mobile host's behalf in an end-to-end mobility approach, the only trust relationship required for secure relocation is between the MN and the CN [73].

4.2.4 Application layer (Layer 5) mobility management protocols

Application layer mobility management using Session Initiation Protocol (SIP) is proposed in [87]. SIP based mobility does not require any changes to the IP stack of the mobile users. In addition, device independent personal mobility and location services are supported by SIP mobility. The performance of SIP mobility protocol is summarized below.

- Because redirecting agents such as SIP proxies and SIP redirect servers are used during handoff, the handoff latency of SIP is comparable to that of Mobile IP but is higher than the transport layer mobility protocols.

- The packets that are in transit during the handoff signaling procedures are lost making handoff packet loss comparable to that of Mobile IP handoff.

- Once the handoff signaling phase is over the communicating hosts i.e., the CN and the MN communicate directly without any redirection agent. Therefore, end-to-end delay does not increase when a MN is away from its home network.

- SIP can not support TCP connections [87]. Therefore, SIP mobility is not transparent to TCP protocol.

- Signaling messages that are used used during SIP mobility management are secured using different security mechanisms. Thus, SIP based mobility management is secure.

Table 1: Qualitative performance of mobility management protocols

Performance parameter	Layer 2	Layer 3	Layer 4	Layer 5
Handoff latency	Worst	Worse	Weak	Worse
Handoff packet loss	Worst	Worse	Weak	Worse
End-to-end delay	Good	Weak	Good	Good
Transport-layer transparency	Weak	Good	Good	Weak
Security	Good	Good	Good	Good

We summarize the performance of the mobility management protocols operating from different layers of the TCP/IP network protocol stack in Table 1, which shows that none of the existing mobility management protocols can support mobility management transparent to different types of applications. Since it is not possible to support transparent mobility management for every type of applications using one particular mobility management protocol in next-generation wireless systems, we advocate the use of a mobility management framework that adaptively selects a mobility management protocol based on applications' requirements. To determine the mobility management protocol that is suitable for a particular class of application, it is essential to understand the handoff performance of the mobility management protocols when they are used for different application classes.

In this Chapter, we develop analytical models to investigate the handoff performance of the existing mobility management protocols in the context of *Class B*, *Class C*, *Class D*, and *Class E* applications. As mentioned before, *Class A* applications do not require any mobility support. Based on the results of our mathematical analysis, we will be able to decide on the suitable mobility management protocol for a particular application class. Moreover, our analysis provides insights about the parameters that influence the handoff performance of the mobility management protocols.

4.3 Parameters and Basic Derivations for Analytical Modeling

To develop analytical modeling for the performance analysis of the existing mobility management protocols, we consider a mobile host (MH[1]) that is away from its home network (HN) moves from an Old Network (ON) to a New Network (NN) in the middle of its communication with a Correspondent Host (CH). The network entities that assist the MH for its mobility management such as a SIP [69] server, a Domain Name Server (DNS), and a home agent (HA) are located in the HN.

Below we carry out some basic derivations that we use for our analytical modeling in the remaining part of this Chapter.

4.3.1 End-to-end packet loss probability

To derive the end-to-end packet loss probability between the MH and the HA (or the CH) located in the Internet, we divide the path between the MH and the HA (or the CH) the into two parts: the wireless link connecting the MH and the BS and the wired link between the BS and the HA (or the CH). Then, the end-to-end packet loss probability p between the MH and the HA (or the CH) is given by

[1]MH and CH (Correspondent Host) are synonymous with MN (Mobile Node) and CN (Correspondent Node), respectively

$$p = 1 - (1 - p_w)(1 - p_c) \tag{18}$$

where p_w and p_c are the packet loss probabilities in the wireless link and the wired link, respectively.

Next, we derive the expressions for p for both no Radio Link Protocol (RLP) and RLP scenarios. We denote by L_p and L_f the length of a packet (typically an IP packet) and the length of a link-layer frame, respectively. Therefore, the number of frames per packet is $K = \lceil \frac{L_p}{L_f} \rceil$. When no RLP is used, the packet loss probability in the wireless link becomes $p_{wnr} = 1 - (1 - p_f)^K$, where p_f is the link layer frame error rate (FER). Therefore, the end-to-end packet loss probability p_{nr} between the MH and the HA (or the CH) without RLP can be derived by using $p = p_{nr}$ and $p_w = p_{wnr}$ in (18). Thus, p_{nr} is

$$p_{nr} = 1 - (1 - p_f)^K (1 - p_c) \tag{19}$$

When RLP is used, the packet loss probability in the wireless link p_{wr} is given by [20]

$$p_{wr} = 1 - \left[1 - p_f ((2 - p_f) p_f)^{\frac{(n^2+n)}{2}} \right]^K \tag{20}$$

where n is the maximum number of trials that the RLP carries out before aborting the attempt to transmit a frame over the link layer. Typically, $n = 3$ for RLP [20].

The end-to-end packet loss probability p_r between the MH and the HA (or the CH) with RLP is obtained from (18) by using $p_w = p_{wr}$ and $p = p_r$. Then p_r is

$$p_r = 1 - \left[1 - p_f ((2 - p_f) p_f)^{\frac{(n^2+n)}{2}} \right]^K (1 - p_c) \tag{21}$$

where p_f is the link layer FER and K is the number of link layer frames per packet.

4.3.2 End-to-end packet transportation delay

The end-to-end packet transportation delay between the MH and the HA (or the CH) is the sum of packet transportation delay over the wireless link from the MH to the

BS and the packet transportation delay in the wired link between the BS and the HA (or the CH). When no RLP is used, there is no frame retransmission in the link layer. Therefore, the end-to-end packet transportation delay, T_{nr}, between the MH and the HA (or the CH) is given by

$$T_{nr} = D + t_w \tag{22}$$

where D is the link-layer access delay and t_w is the delay in the wired link between the BS and the HA (or the CH).

The one way frame transportation delay T_f between the MH and the BS with RLP is given by [20]

$$T_f = D(1 - p_f) + \sum_{i=1}^{n} \sum_{j=1}^{i} P(C_{i,j})(2iD + 2(j-1)\tau) \tag{23}$$

where p_f is the link layer FER and τ is the link layer inter-frame interval, which is typically around 20 ms [20]. $P(C_{i,j})$ is the probability that the first frame transmitted by the MH is received correctly by the BS, being the ith retransmitted frame at the jth retransmission trial. The expression for $P(C_{i,j})$ is given by [20]

$$P(C_{i,j}) = p_f(1-p_f)^2((2-p_f)p_f)^{(\frac{i^2-i}{2}+j-1)} \quad \text{for } i = 1, 2, ..., n \text{ and } j = 1, 2, .., i \tag{24}$$

Therefore, when RLP is used, the end-to-end packet transportation delay, T_r, between the MH and the HA (or the CH) is then

$$T_r = T_f + (K-1)\tau + t_w \tag{25}$$

where K is the number of link layer frames per packet as defined in Section 4.3.1.

4.3.3 Average signaling packet transportation delay using UDP

A signaling packet is retransmitted by the sender until it is received correctly by the destination. In this case, if the sender does not receive a reply for its transmitted packet, it retransmits the packet when the retransmission timer of the packet expires.

The average one way signaling packet transportation delay, D_p, between the MH and the HA (or the CH) is computed then by

$$D_p = \sum_{i}^{\infty} p_i T_i \tag{26}$$

where T_i is the packet transportation delay when the packet is successfully transferred between the MH and the HA (or the CH) in the ith retransmission trial and p_i is the probability that a packet is successfully transferred between the MH and the HA (or the CH) in the ith retransmission trial. p_i is computed by

$$p_i = q^{i-1}(1-q) \tag{27}$$

where q is the end-to-end packet loss probability between the MH and the HA (or the CH). $q = p_{nr}$ when no RLP is used and $q = p_r$ when RLP is used. The expressions for p_{nr} and p_r are derived in (74) and (75), respectively.

The formulation for T_i is as follows.

$$T_i = \begin{cases} \Delta + \gamma\Delta + \gamma^2\Delta + \ldots + \gamma^{i-2}\Delta + B & i \leq m \\ \Delta + \gamma\Delta + \gamma^2\Delta + \ldots + \gamma^{m-2}\Delta + (i-m)\gamma^{m-2}\Delta + B & i > m. \end{cases} \tag{28}$$

where the special cases are $T_1 = B$ and $T_2 = \Delta + B$. Next, we define each term in (28). m is an integer such that after mth retransmission timeout the retransmission timer is frozen. $B = T_{nr}$ when no RLP is used and $B = T_r$ when RLP is used. The expressions for T_{nr} and T_r are derived in (72) and (73), respectively. Δ is the initial value of the retransmission timer, which is large enough to account for the size of the messages, twice the round trip time between the MH and the HA (or the CH), and at least an additional 100 ms to allow for processing the messages at the MH and the HA (or the CH). γ is the factor by which the retransmission timeout (RTO) duration is incremented after each failed retransmission. Typically, $\gamma = 2$.

Now, using the formulations for p_is and T_is from (27) and (28), respectively, we simplify (26) to obtain

$$D_p = \sum_{i}^{\infty} p_i T_i = p_1 T_1 + \sum_{i=2}^{m} p_i T_i + \sum_{i=m+1}^{\infty} p_i T_i$$

$$
\begin{aligned}
&= (1-q)B + \sum_{i=2}^{m} q^{i-1}(1-q)[\Delta + \gamma\Delta + \gamma^2\Delta + \ldots + \gamma^{i-2}\Delta + B] + \\
&\quad \sum_{i=m+1}^{\infty} q^{i-1}(1-q)[\Delta + \gamma\Delta + \gamma^2\Delta + \ldots + \gamma^{m-2}\Delta + (i-m)\gamma^{m-2}\Delta + B] \\
&= (1-q)\Big\{B + A\sum_{i=2}^{m} q^{i-1}(\gamma^{i-1} - 1) + \\
&\quad \sum_{i=m+1}^{\infty} q^{i-1}[A(\gamma^{m-1} - 1) + (i-m)\gamma^{m-2}\Delta\Big\} \qquad (29)
\end{aligned}
$$

where $A = \frac{\Delta}{\gamma - 1}$.

4.4 *Handoff of performance of Class B and Class C Applications (Mobile IP and TCP-Migrate)*

As *Class B* and *Class C* applications use TCP, we consider a TCP connection between a CH and MH to investigate their handoff performance. The handoff performance of *Class B* and *Class C* applications is synonymous with the handoff performance of a TCP connection. We consider a scenario where the MH while in the Old Network (ON) starts to download a file using FTP from the CH and moves into the New Network (NN) in the middle of this file transfer. We assume that the size of the file is long enough for the TCP connection to continue from the ON to the NN. We further assume that CH's FTP application creates packets continuously such that CH's TCP sends full-sized segments (packets) as fast as its congestion window allows. Moreover, we assume that the window size advertised by the receiver (the MH in this case) is always larger than the congestion window size. Therefore, the sending window size is always limited by the congestion window. We assume that while the MH is in the ON, the TCP connection between the CH and the MH operates in a steady state. During this steady state, TCP state parameters, e.g., congestion window size and round trip time (RTT) are decided by the path between the CH and the MH. To maintain the highest throughput performance in different types of wireless networks characterized by different p_f and D and achieve fairness to the wired TCP sources sharing the

same bottleneck, we consider the adaptive congestion control proposed in [7] that dynamically adjusts additive-increase multiplicative-decrease (AIMD) parameters (α, β) according to the current wireless link conditions. The expression for α is given by [7]

$$\alpha = \frac{bp(1-\beta)}{2(1+\beta)}\left[\hat{T}(2R + 3T_0 p(1+32p^2)(1+\beta))\right]^2 \tag{30}$$

where p is the end-to-end packet loss probability, $\hat{T}$ is the throughout achieved by a wired TCP source experiencing p_c, which is the packet loss probability due to congestion in the wired network and R_c, which is the end-to-end RTT in the wired network. R is the end-to-end RTT between the CH and the MH, T_0 is the initial retransmission timeout (RTO) for the TCP connection, and b is the number of data packets acknowledged with a single ACK. The numerical value of β can be set to be 0.75, 0.80, and 0.85, for a WLAN, a 3G cellular network, and a satellite network, respectively [7]. p is the end-to-end packet loss probability between the MH and the CH.

The throughput of a TCP connection with AIMD parameters (α, β) is given by [93]

$$T_{\alpha,\beta}(p, R, T_0, b) = \frac{1}{R\sqrt{\frac{2b(1-\beta)p}{\alpha(1+\beta)}} + T_0 \min\left(1, 3\sqrt{\frac{(1-\beta^2)bp}{2\alpha}}\right)p(1+32p^2)} \tag{31}$$

The numerical values of α and β for wired TCP are 1 and $\frac{1}{2}$, respectively. Therefore, using (31), the expression for $\hat{T}$ in (30) is given by

$$\hat{T} = \frac{1}{R_c\sqrt{\frac{2bp_c}{3}} + T_{0c}\min\left(1, 3\sqrt{\frac{3bp_c}{8}}\right)p_c(1+32p_c^2)} \tag{32}$$

where T_{0c} is the initial RTO.

The steady state congestion window size of TCP depends on the end-to-end packet loss probability and is given by [93]

$$E[W] = \frac{\alpha + b(1-\beta)}{2b(1-\beta^2)} + \sqrt{\left(\frac{\alpha + b(1-\beta)}{2b(1-\beta^2)}\right)^2 + \frac{2\alpha(1-p)}{bp(1-\beta^2)}} \tag{33}$$

We use (30) to determine the additive-increase parameter of a TCP connection and (33) to calculate the steady state congestion window size when the MH is in the ON and the NN. While the MH is in the ON, we assume that the TCP connection is operating in the steady state corresponding to the link-layer frame error rate (p_f) and the end-to-end frame transportation delay (D) of the ON. After MH's handoff to the NN, the TCP connection should reach the steady state corresponding to the NN as soon as possible. Ideally, the TCP connection should switch from the steady state of the ON to that of the NN immediately after the handoff.

- Definition: *Throughput degradation time* is the time required for a TCP connection to switch from the steady state of the ON to the steady state of the NN.

Moreover, an ideal handoff management protocol should ensure that the application running over TCP at the MH does not observe any *handoff latency* during MH's movement from the ON to the NN.

- Definition: *Handoff latency* is the time elapsed after the MH receives the last packet in the ON until the MH receives the first packet (with the sequence number one higher than the one last received in the ON) in the NN.

As TCP is a reliable protocol, there is no packet loss during a handoff as lost packets are recovered through retransmissions after the handoff is completed. Therefore, the handoff performance of a TCP connection can be represented by two parameters: (1) throughput degradation time and (2) handoff latency. Next, we investigate the performance of a TCP connection when Mobile IP [63] is used as the mobility management protocol followed by when TCP-Migrate is used.

4.4.1 Handoff performance analysis of a TCP connection when Mobile IP is used

Mobile IP [63] deploys a HA that intercepts packets destined for an MH currently away from its HN. The HA tunnels the intercepted packets to the MH via a foreign agent (FA) (when foreign agent care-of-address is used) or directly (when co-located care-of-address is used) in the foreign network (FN) [63]. Figure 23 shows the Mobile IP [63] handoff process for a TCP connection when the MH moves from the ON to the NN. In Figure 23 t_{ch} is the one way delay between the CH and the HA, t_{ho} is the one way delay between the MH and its HA when the MH is in the ON, t_{hn} is the one way delay between the MH and its HA when the MH is in the NN, t_o is the one way delay between the MH and the CH while the MH is in the ON, and t_n is the one way delay between the MH and the CH while the MH is in the NN. As shown in Figure 23 the HA intercepts the packets for the MH. Then the HA tunnels the packets to the MH.

In Figure 23, the MH enters the NN at time A. Therefore, at this time the MH starts the layer 2 handoff (L2 handoff) to the NN. As pointed out earlier, before time A, the TCP connection operates in the steady state corresponding to the ON. We denote the congestion window size of this steady state as CW_1. We assume that all packets received by the MH before time A are properly ACKed and all these ACKs are received by the CH. We denote the sequence number of the packet received at time A as n. Therefore, the MH is expecting the packet with sequence number $n+1$ next. As shown in Figure 23, the MH starts layer 2 handoff to the NN and IP address acquisition from the NN at time A. These procedures are completed at time B. Then, at time B the MH starts Mobile IP [63] registration with its HA. The new care-of-address (CoA) of the MH gets successfully registered at the HA at time instant C. Thus, packets received by the HA after time C are correctly forwarded to the MH in the NN. The packets received by MH's HA from the CH between time G and C are

lost as they are forwarded to MH's old CoA. The last ACK sent by the MH from the ON is received by the CH at time E. Therefore, the CH transmits all packets in its congestion window, i.e., CW_1 number of packets, after E and waits for ACKs. One of the following scenarios may occur:

- *Case A:* The new CoA of the MH is registered at the HA after the HA receives the packet transmitted by the CH at time F. In this case, all packets in the congestion window (from E to F) are lost as the HA tunnels these packets to MH's old CoA. Therefore, the CH does not receive the ACKs for these packets and waits until the RTO of the packet transmitted at time E to occur. Then, it reduces the congestion window to one and retransmits the packet for which RTO occurs at time RTO_1. If the new CoA of the MH is not registered at the HA by the time this retransmitted packet reaches the HA, the HA sends the packet to MH's old CoA and the packet is lost again. Then, CH's TCP doubles the value of RTO, waits until the second RTO, and retransmits the packet when the second RTO expires. If the retransmitted packet after Nth RTO reaches the HA after MH's new CoA is registered at the HA, then the HA tunnels the packet to MH's new CoA. In this case, the retransmitted packet is successfully received by the MH in the NN.

- *Case B:* The new CoA of the MH is registered at the HA before the HA receives the packet transmitted by the CH at time F. In this case, the packets that belong to the congestion window (from E to F) and arrive after the registration of MH's new IP address at the HA are tunneled to MH's new CoA. TCP takes one RTT to transmit all the segments in one congestion window. Typically the Mobile IP handoff latency is larger than the RTT. Therefore, this case occurs very rarely.

We determine the handoff latency and throughput degradation time of a TCP connection for Case A as described below.

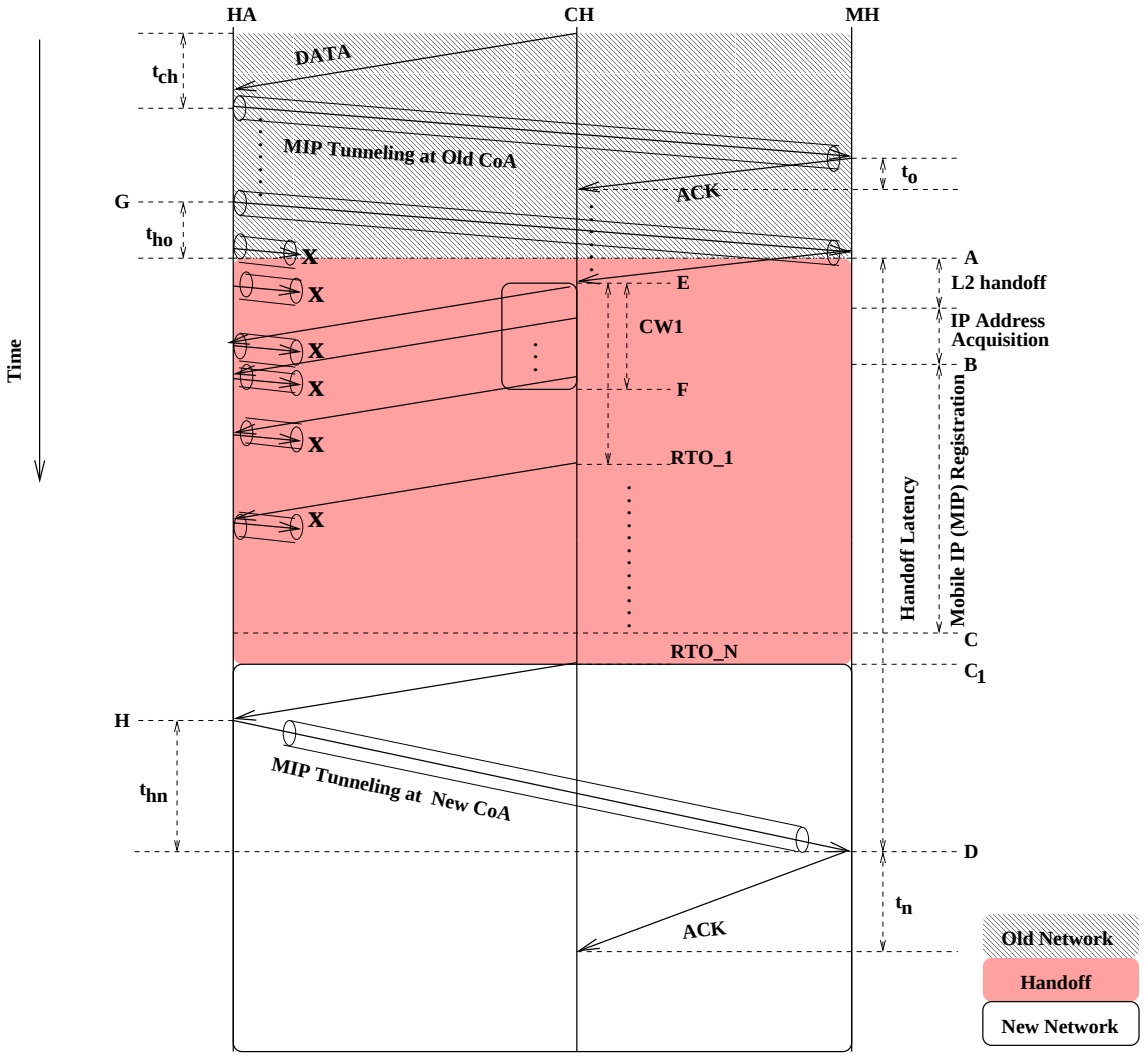


Figure 23: Handoff of a TCP connection using Mobile IP.

4.4.1.1 Handoff latency

To calculate handoff latency (time interval between the receipt of the packet with sequence number n by the MH in the ON and the receipt of the packet with sequence number $n+1$ by the MH in the NN), we first determine the time during which the packets transmitted by the CH are lost. This time is given by

$$T = C - A = \tau_{L2} + \tau_a + \tau_m \tag{34}$$

where τ_{L2} is the time required for MH's Layer 2 handoff to the NN, τ_a is the time required for new IP address acquisition by the MH in the NN, and τ_m is the time

required for Mobile IP [63] registration. If T is such that the sending window of CH's TCP is exhausted (which is usually the case), then CH's TCP goes through timeouts, reduces the congestion window to one and then starts to retransmit the packet whose RTO expires. If this retransmitted packet reaches the HA before the HA has MH's new CoA, then the retransmitted packet is lost again. CH's TCP doubles the value of RTO and waits until the timeout to transmit this packet again. In this way, if N number of timeouts occur before the HA receives the new CoA of the MH, then the handoff latency T_{h1} is given by

$$T_{h1} = D - A = D - C_1 + C_1 - A \tag{35}$$

$C_1 - A$ depends on the number of TCP timeouts N that occur before the HA receives MH's new CoA. From Figure 23, $C_1 - A$ is

$$\begin{aligned} C_1 - A &= C_1 - E + E - A \\ &= \begin{cases} TO_1 + \gamma TO_1 + ... + \gamma^N TO_1 + t_o & \text{if } N \leq m \\ TO_1 + \gamma TO_1 + ... + \gamma^m TO_1 + (N-m)\gamma^m TO_1 + t_o & \text{if } N > m \end{cases} \\ &= \begin{cases} TO_1 \frac{\gamma^{N+1}-1}{\gamma-1} + t_o & \text{if } N \leq m \\ TO_1 \frac{\gamma^{m+1}-1}{\gamma-1} + (N-m)\gamma^m TO_1 + t_o & \text{if } N > m \end{cases} \end{aligned} \tag{36}$$

Now using (36), (35) can be expressed as

$$T_{h1} = \begin{cases} TO_1 \frac{\gamma^{N+1}-1}{\gamma-1} + t_o + t_{ch} + t_{hn} & \text{if } N \leq m \\ TO_1 \frac{\gamma^{m+1}-1}{\gamma-1} + (N-m)\gamma^m TO_1 + t_o + t_{ch} + t_{hn} & \text{if } N > m \end{cases} \tag{37}$$

where TO_1 is the initial retransmission time out (RTO) period for the TCP connection when the MH is in the ON and is given by $TO_1 = \xi RTT_o$ where ξ is a constant weighting factor and RTT_o is the RTT of the TCP connection when the MH is in the ON. From Figure 23, $RTT_o = t_{ch} + t_{ho} + t_o$, where t_{ch} is the one way delay between the CH and the HA, t_{ho} is the one way delay between the MH and the HA when the MH is in the ON, and t_o is the one way delay between the MH and the CH while the MH is in the ON.

T in (34) required to determine the number of retransmission timeouts, N, that CH's TCP undergoes before the HA receives the new CoA of the MH. Once N is determined, the handoff latency can be calculated using (37). τ_{L2} and τ_a in (34) are usually constant for a particular wireless system such as a WLAN, 3G, Satellite network etc. On the other hand, τ_m depends on the distance between the MH and its HA and on the wireless link conditions.

We derive the expression for τ_m as follows. τ_m is equal to the time required for MH's *Mobile IP Registration Request* [63] message to reach the HA and HA's *Mobile IP Registration Reply* [63] to reach the MH, i.e., $\tau_m = 2D_{mh}$ where D_{mh} is the average one way delay to transport Mobile IP signaling packets between the MH and the HA. Note that Mobile IP signaling messages are transported using UDP [63]. Using steps similar to the derivation of (29), D_{mh} is given by

$$\begin{aligned} D_{mh} &= (1-q_1)\Bigg\{B_1 + A_1 \sum_{i=2}^{m} q_1^{i-1}(\gamma^{i-1}-1) + \\ &\qquad \sum_{i=m+1}^{\infty} q_1^{i-1}[A_1(\gamma^{m-1}-1) + (i-m)\gamma^{m-2}\Delta_1]\Bigg\} \end{aligned} \tag{38}$$

Next, we define each term in (38). B_1 is the end-to-end packet transportation delay between the MH and the HA. $B_1 = B_{1nr}$ when no RLP is used and $B_1 = B_{1r}$ when RLP is used. B_{1nr} is computed from (72) by using $T_{nr} = B_{1nr}$ and $t_w = t_{whn}$. B_{1r} is computed from (73) by using $T_r = B_{1r}$, $K = K_m$, and $t_w = t_{whn}$. $K_m = \lceil \frac{L_m}{L_f} \rceil$ is the number of link layer frames per one *Mobile IP Registration Request/Reply* message, where L_m is the length of a *Mobile IP Registration Request/Reply* message and L_f is the length of a link-layer frame. t_{whn} is the one way delay in the wired network between the NBS and the HA.

q_1 is the end-to-end packet loss probability between the MH and the HA. $q_1 = q_{1nr}$ when no RLP is used and $q_1 = q_{1r}$ when RLP is used. q_{1nr} is computed from (74) by using $p_{nr} = q_{1nr}$ and $K = K_m$. q_{1r} is computed from (75) by using $p_r = q_{1r}$ and $K = K_m$.

Δ_1 is the initial value of the retransmission timer for Mobile IP signaling messages, which is large enough to account for the size of the Mobile IP signaling messages, twice the round trip time between the MH and the HA, and at least an additional 100 ms to allow for processing the messages at the MH and the HA. γ is the factor by which the retransmission timeout (RTO) duration is incremented after each failed retransmission. Typically, $\gamma = 2$. $A_1 = \frac{\Delta_1}{\gamma - 1}$. m is an integer such that after mth retransmission timeout the retransmission timer is frozen.

4.4.1.2 Throughput degradation time

As discussed earlier, the HA receives the Nth retransmission packet after the successful registration of MH's new CoA. Therefore, the HA tunnels the Nth retransmitted packet and subsequent packets transmitted by CH's TCP to MH's new CoA. CH's TCP resumes TCP slow start operation at time C_1 as shown in Figure 23. Then, it increases the congestion window to the steady state value of the NN denoted by CW_2. The time required by TCP to increase its congestion window size from 1 to CW_2, τ_s, is given by

$$\tau_s = [1 + \log_2 CW_2] RTT_n \tag{39}$$

where RTT_n is the RTT when the MH is in the NN and is given by, $RTT_n = t_{ch} + t_{hn} + t_n$, where t_{ch} is the one way delay between the CH and the HA, t_{hn} is the one way delay between the MH and the HA when the MH is in the NN, and t_n is the one way delay between the MH and the CH while the MH is in the NN.

The time for which the TCP connection experiences throughput degradation T_{t1} is equal to $T_t = (C_1 - A) + \tau_s$. Using (36) and (39), the expression for T_{t1} is

$$T_{t1} = \begin{cases} TO_1 \frac{\gamma^{N+1} - 1}{\gamma - 1} + t_o + [1 + \log_2 CW_2] RTT_n & \text{if } N \leq m \\ TO_1 \frac{\gamma^{m+1} - 1}{\gamma - 1} + (N - m)\gamma^m TO_1 + t_o + [1 + \log_2 CW_2] RTT_n & \text{if } N > m \end{cases} \tag{40}$$

4.4.2 Handoff performance analysis of a TCP connection when TCP-Migrate is used

We select TCP-Migrate [73] as the representative transport layer mobility management protocol as it requires minimum change in the network infrastructure, whereas other solutions such as MSOCKS [51] require the introduction of an additional network entity such as a proxy to split the TCP connection [19]. Next, we briefly explain the operation of TCP-Migrate [73] during a handoff.

The MH and the CH negotiate a *token* through the Migrate option as described in [73] during the initial TCP connection establishment. Thus, a TCP connection can be uniquely identified at the MH and the CH by either <*MH's address, MH's port, CH's address, CH's port*> 4-tuple or a new <*CH's address, CH's port, token*> triple [73]. When the MH moves to the NN and receives a new IP address, it sends a SYN segment containing its new IP address and a Migrate Option to the CH. This SYN segment includes the *token* computed during the initial connection establishment in the Token field. The CH identifies the connection corresponding to this *token* and changes the address and port to match MH's new IP address. Then the CH resets the congestion-related states of the connection to the initial values and resumes the connection from the slow start operation of TCP. Further details about the operation of TCP-Migrate can be found in [73].

Figure 24 shows the TCP-Migrate handoff process of a TCP connection when the MH moves from the ON to the NN. At time A, the MH starts the handoff process to the NN. We assume that all packets received by the MH before time A are properly ACKed and all of them are received by the CH. We denote the sequence number of the packet received at time A as n. Therefore, the MH is expecting the packet with sequence number $n + 1$ next. As shown in Fig. 24, the MH starts layer 2 handoff to the NN and IP address acquisition from the NN at time A. These procedures are completed at time B. Then, the MH starts the TCP-Migrate handoff process that is

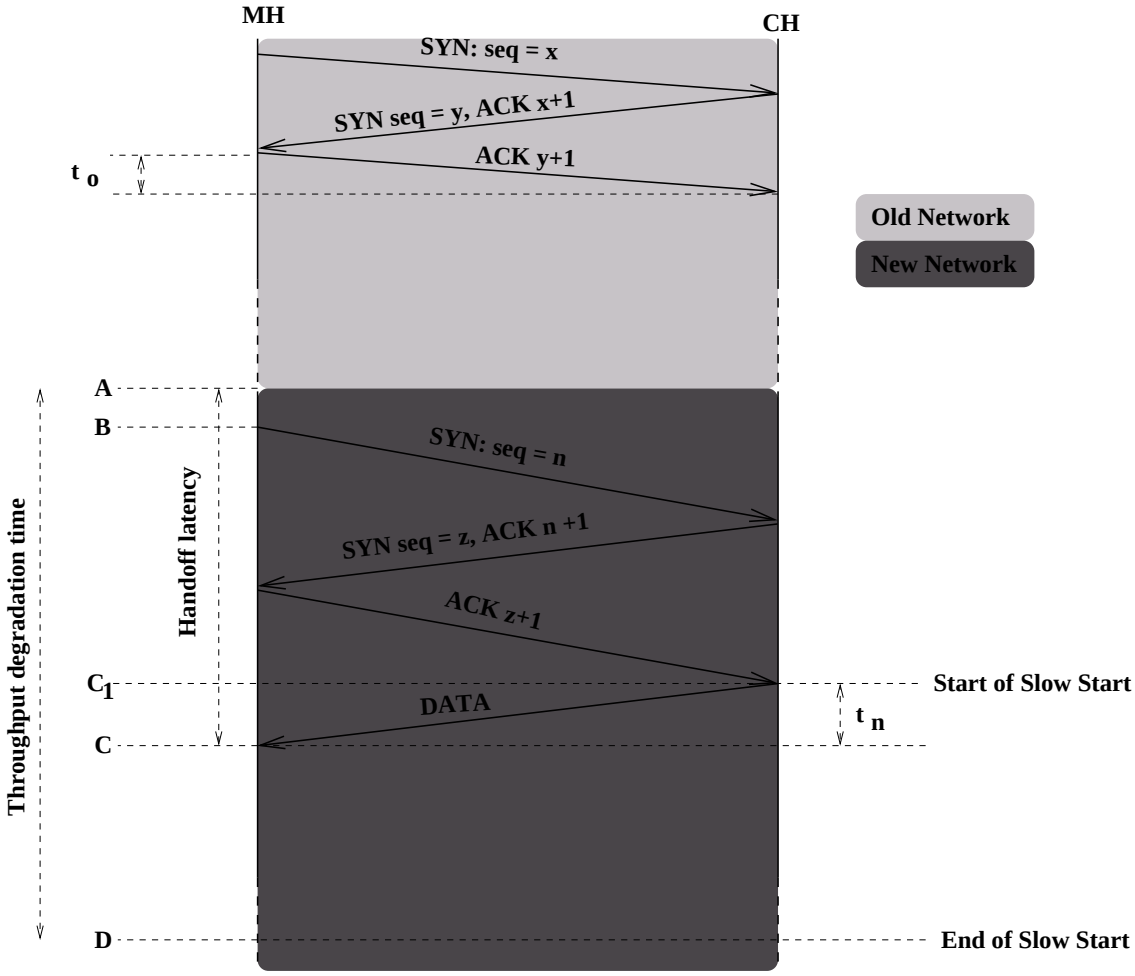


Figure 24: Diagram showing the operation of TCP-Migrate

completed at time C_1. Then the CH resumes the TCP connection from slow start at time C_1 as shown in Fig. 24. The slow start ends at time D, i.e., the TCP connection reaches the steady state corresponding to the NN. We determine the handoff latency and throughput degradation time of the TCP connection as described below.

4.4.2.1 Handoff latency

From Figure 24, the TCP-Migrate handoff latency, T_{h2}, is given by

$$T_{h2} = C - A = \tau_{L2} + \tau_a + E[L] + t_n$$

where τ_{L2} and τ_a are the time required for MH's L2 handoff to the NN and IP address acquisition in the NN, respectively. t_n is the one way delay between the MH and the

CH while the MH is in the NN. $E[L]$ is the average delay for the transportation of TCP-Migrate signaling messages. Next we derive the expression for $E[L]$.

First, the MH sends a SYN packet with TCP-Migrate options containing the MH's new IP address to the CH $i \geq 0$ times unsuccessfully until the $(i+1)$th SYN arrives successfully at the CH. Then, the CH repeatedly retransmits its SYN/ACK until it receives an ACK from the MH. Let CH sends SYN/ACK $j \geq 0$ times unsuccessfully and the $(j+1)$th SYN/ACK successfully arriving at the MH. Then, the MH retransmits the ACK to the CH that gets successfully transmitted in the $(k+1)$th trial ($k \geq 0$). Therefore, the probability $P_h(i,j,k)$ that the TCP-Migrate handoff is completed after the exchange of i unsuccessful SYNs, followed by one successful SYN, followed by exactly j SYN/ACK failures followed by one successful SYN/ACK, followed by k unsuccessful ACKs, followed by one successful ACK is given by

$$P_h(i,j,k) = p_1^i(1-p_1)p_2^j(1-p_2)p_2^k(1-p_2) \quad \text{for } i,\ j,\ k = 0,\ 1,\ 2,\ ...,N_m\text{-}1 \qquad (41)$$

where N_m is such that TCP abort connection establishment attempts after N_m number of retransmissions. p_1 is the end-to-end packet loss probability between the MH and the CH for a SYN packet and p_2 is the end-to-end packet loss probability between the MH and the CH for a SYN/ACK or ACK packet. $p_1 = p_{1nr}$ when no RLP is used and $p_1 = p_{1r}$ when RLP is used. p_{1nr} is computed from (74) by using $p_{nr} = p_{1nr}$ and $K = K_1$. p_{1r} is computed from (75) by using $p_r = p_{1r}$ and $K = K_1$. $K_1 = \lceil \frac{L_1}{L_f} \rceil$ is the number of link layer frames per one SYC packet. L_1 is the length of the SYC packet and L_f is the length of a link-layer frame. Similarly, $p_2 = p_{2nr}$ when no RLP is used and $p_2 = p_{2r}$ when RLP is used. p_{2nr} is computed from (74) by using $p_{nr} = p_{2nr}$ and $K = K_2$. p_{2r} is computed from (75) by using $p_r = p_{2r}$ and $K = K_2$. $K_2 = \lceil \frac{L_2}{L_f} \rceil$ is the number of link layer frames per one SYN/ACK or ACK packet. L_2 is the length of the SYN/ACK or ACK packet. The handoff latency for the above scenario is given by

$$
\begin{aligned}
L_h(i,j,k) &= 1.5RTT_n + \sum_{m=0}^{i-1} 2^m RTO + \sum_{m=0}^{j-1} 2^m RTO + \sum_{m=0}^{k-1} 2^m RTO \\
&= 1.5RTT_n + (2^i + 2^j + 2^k - 3)RTO \\
&\qquad \text{for } i,\ j,\ k = 0,\ 1,\ 2,\ ...,N_m\text{-}1
\end{aligned} \tag{42}
$$

where RTO is the initial retransmission time out for the TCP connection, $RTO = \xi RTT_o$ and RTT_o is the RTT in the ON. Therefore, the average TCP-Migrate handoff latency is

$$
E[L] = \sum_{i=0}^{N_m-1} \sum_{j=0}^{N_m-1} \sum_{k=0}^{N_m-1} P_h(i,j,k) L_h(i,j,k). \tag{43}
$$

4.4.2.2 Throughput degradation time

As described earlier, CH's TCP resumes TCP slow start operation at time C_1 as shown in Figure 24. Then, it increases the congestion window to the steady state value of the NN denoted by CW_2. We assume until the congestion window size reaches the steady state value of the NN at time D, TCP does not experience any packet loss. Therefore, the TCP connection experiences throughput degradation from time A to D. The expression for handoff degradation time, T_{t2}, is given by

$$
T_{t2} = D - A = \tau_{L2} + \tau_a + E[L] + [1 + \log_2 CW_2]RTT_n \tag{44}
$$

where τ_{L2} and τ_a are the time required for MH's L2 handoff to the NN and IP address acquisition in the NN, respectively. t_n is the one way delay between the MH and the CH while the MH is in the NN. RTT_n is the RTT when the MH is in the NN. $E[L]$ is given by (78).

4.4.3 Handoff performance comparison of Mobile IP and TCP-Migrate for a TCP connection

To compare the performance of Mobile IP (MIP) and TCP-Migrate based handoff for a TCP connection, we assume the following values for different parameters: the

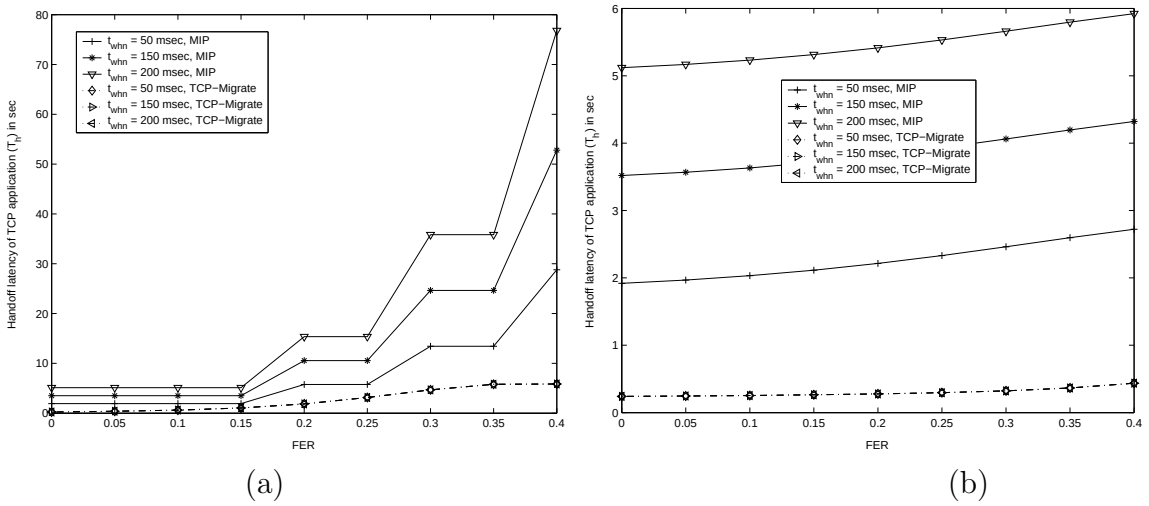


Figure 25: Handoff latency comparison of Mobile IP and TCP-Migrate (a) no RLP and (b) RLP.

time required for MH's L2 handoff to the NN $\tau_{L2} = 10$ ms, the time required for IP address acquisition in the NN $\tau_a = 20$ ms, one way delay between the CH and the HA $t_{ch} = 50$ ms, link layer access delay $D = 10, 50, 150$ ms for WLAN, 3G cellular, and satellite networks, respectively [7], length of link-layer frame $L_f = 19$ bytes, link-layer inter-frame interval $\tau = 20$ ms, one way delay in the wired network between the old BS (OBS) and the CH $t_{wco} = 100$ ms, one way delay in the wired network between the new BS (NBS) and the CH $t_{wcn} = 100$ ms, packet loss probability in the wired network $p_c = 1e - 5$. We denote the one way delay in the wired network between the OBS and the HA when the MH is in the ON by t_{who}. Similarly, t_{whn} is the one way delay in the wired network between the NBS and the HA when the MH is in the NN. We consider $t_{who} = t_{whn}$ and use different values for them in our simulations.

Figure 25 (a) shows the handoff latency comparison of Mobile IP and TCP-Migrate for a TCP connection when no RLP is used in the link layer. Similarly, Figure 25 (b) shows the handoff latency comparison for Mobile IP and TCP-Migrate when RLP is used. The results show that for both no RLP and RLP scenarios, the handoff latency of Mobile IP is always greater than that of TCP-Migrate. The reason is twofold. First, the Mobile IP signaling messages are transferred between the MH and

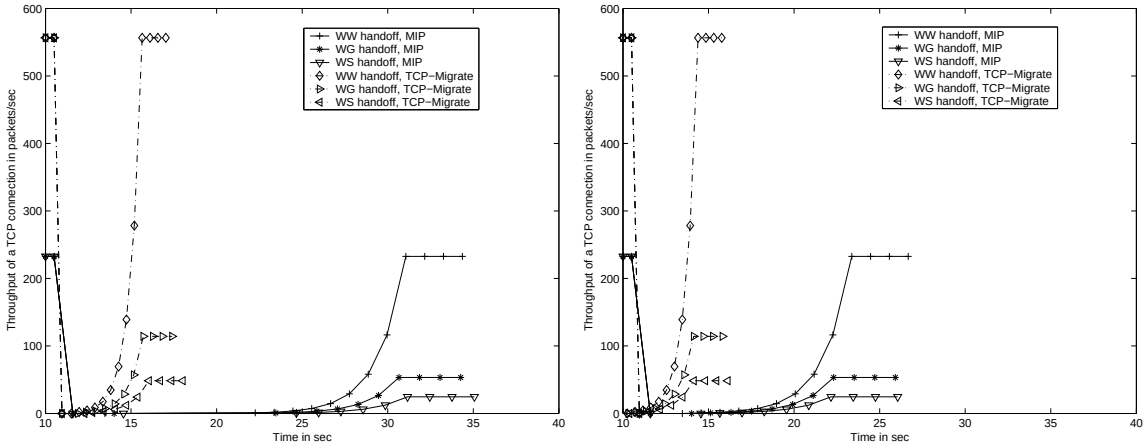


Figure 26: Throughout degradation duration comparison of a TCP connection for Mobile IP and TCP-Migrate (a) no RLP and (b) RLP.

its HA, whereas TCP-Migrate signaling messages are transferred between the MH and the CH. Typically, the distance between the MH and its HA is higher than the distance between the MH and the CH. Second, Mobile IP handoff is not transparent to TCP. Therefore, even after MH's new CoA is registered at the HA, the TCP waits until the retransmission timer to timeout before sending a new packet. On the other hand, when TCP-Migrate is used, CH's TCP resumes the TCP connection as soon as it receives the new IP address. The results also show that the handoff latency for Mobile IP and TCP-Migrate increases as the wireless link FER increases. This can be explained as follows. When no RLP is used in the link layer, a higher value of FER increases the probability of erroneous packet transfer across the link layer. Therefore, the handoff signaling messages have to be retransmitted several times before the successful completion of a handoff. Similarly, when RLP is used in the link layer, a higher value of FER requires more number of link layer retransmissions for successful transfer of handoff messages across the link layer. This increases the link layer packet transfer delay and results in higher handoff signaling delay. During a handoff, the MH is around the boundary of a cell coverage and suffers from higher link layer FER. When no RLP is used, higher FER results in very high Mobile IP handoff latency.

For FER of around 0.2 the Mobile IP handoff latency is around five times higher than the handoff latency of TCP-Migrate. Moreover, as shown in Figure 25 (a) and Figure 25 (b) Mobile IP handoff depends on the delay between the MH and its HA (t_{whn}) as the signaling messages are exchanged between them. On the other hand, as expected the handoff latency of TCP-Migrate depends only on the distance between the MH and the CH.

The throughput of a TCP connection during a handoff is shown in Figure 26 (a) and Figure 26 (b) for no RLP and RLP scenarios, respectively. To investigate the throughput performance of Mobile IP and TCP-Migrate, we use $p_f = 0.2$ and $t_{who} = t_{whn} = 200$ ms. Figure 26 (a) and Figure 26 (b) show the throughput of a TCP connection when the MH previously in a WLAN moves to a WLAN or 3G cellular or Satellite network. We refer to the handoff from a WLAN to another WLAN network as WW handoff. Similarly, WLAN to 3G cellular and WLAN to Satellite network handoffs are referred as WG handoff and WS handoff, respectively. In Figure 26 (a) and Figure 26 (b), the MH moves into the NN at time 10.5 seconds. Therefore, before this time the TCP connection operates in a steady state corresponding to the ON, which is a WLAN in this case. Then, after MH's movement to the NN (either a WLAN or 3G network, or Satellite network) until the handoff process is completed the packets destined for the MH are lost resulting in zero throughput. After, the successful registration of MH's new CoA at the HA, the MH starts to receive packets in the NN. As TCP starts from slow start after the handoff, it takes finite amount of time for TCP to reach its steady state in the NN. Figure 26 (a) and Figure 26 (b) show that this time is minimum in the case of WLAN to WLAN (WW) handoff and maximum for WLAN to Satellite network (WS) handoff. This because the one way access delay of a WLAN network is the lowest and that of the Satellite network is the highest. The dotted lines and solid lines represent the throughput of the TCP connection for TCP-Migrate and Mobile IP, respectively. The results also show that

the throughput degradation of the TCP connection lasts longer for Mobile IP than that of TCP-Migrate. The higher handoff latency of Mobile IP based handoff results in a longer throughput degradation time compared to TCP-Migrate based handoff. Figure 26 (a) and Figure 26 (b) show that for the parameters considered in our analysis, the throughput degradation during a Mobile IP handoff is around twice that of TCP-Migrate. The numerical value of the handoff degradation depends on the handoff latency that depends on the numerical value of FER and the distance between the MH and CH and the MH and its HA. However, Mobile IP always has higher handoff latency and higher throughput degradation time compared to TCP-Migrate.

To summarize, the handoff latency and throughput degradation time of Mobile IP depend on the link layer FER (p_f), the delay between the MH and HA, and wireless access technology. Similarly, the handoff latency and throughput degradation time of TCP-Migrate based handoff depend on the link layer FER (p_f), the delay between the MH and CH, and wireless access technology. TCP-Migrate has lower handoff latency and lower handoff degradation time for *Class B* and *Class C* applications compared to Mobile IP. Therefore, we advocate that TCP-Migrate is suitable for these applications.

4.5 Handoff performance of *Class D* and *Class E* Applications (Mobile IP and SIP)

As *Class D* and *Class E* applications use UDP, we consider a UDP connection between the MH and the CH to investigate their handoff performance. The handoff performance of *Class D* and *Class E* applications is synonymous with the handoff performance of a UDP connection. We consider voice over IP (VoIP) application that uses RTP over UDP. It may be noted that the same analysis is valid for other

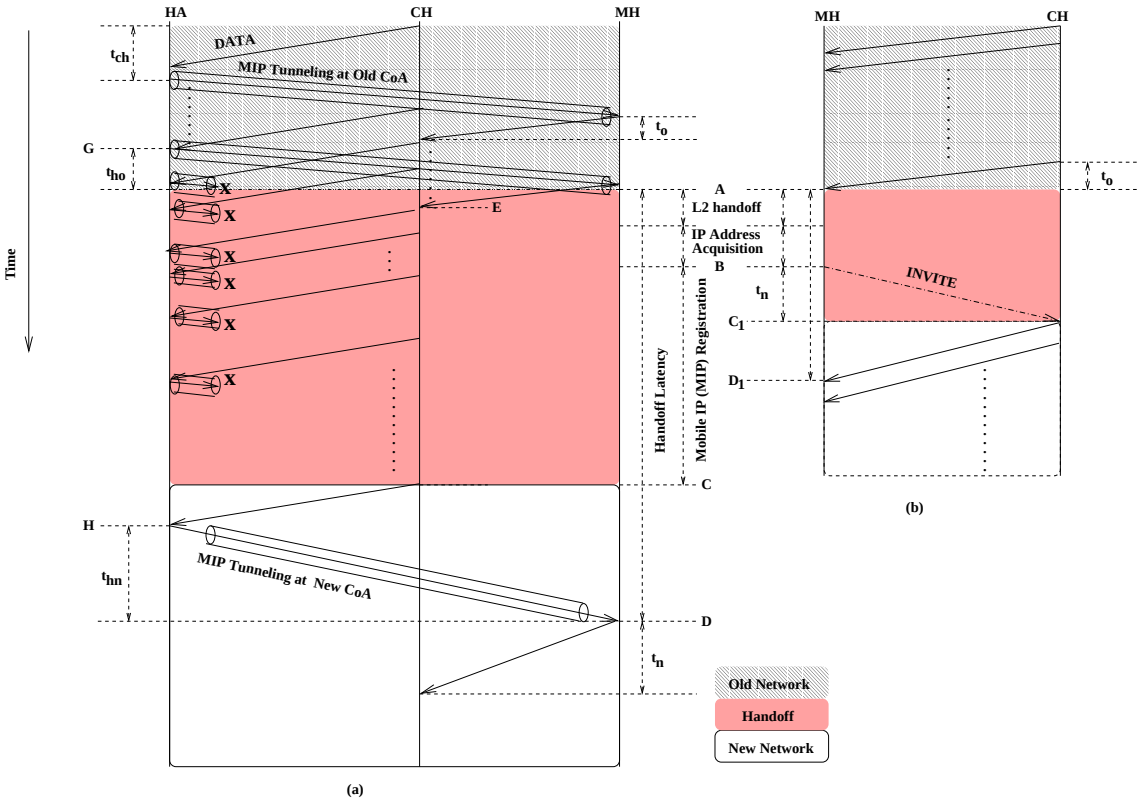


Figure 27: Handoff of a UDP connection using (a) Mobile IP and (b) SIP.

real and non-real time applications using UDP. UDP is not a reliable transport protocol, thus packets lost during a handoff process can not be recovered. Moreover, in this case as we are considering a real-time VoIP application, there is no benefit to buffer packets during a handoff and deliver those after the handoff is completed. Out of the different handoff performance parameters discussed in Section 5.1, since we are considering a UDP connection, we do not consider the *transport-layer transparency*. Mobile IP and SIP support secure handoff. Therefore, we also do not consider *security* in our analysis. As a result, we involve the following three metrics to investigate the performance of Mobile IP and SIP for the VoIP application: *handoff latency*, *packet loss during handoff*, and *end-to-end delay*. The *end-to-end delay* corresponds to the transportation delay of the VoIP data packets.

4.5.1 Handoff performance of a UDP connection when Mobile IP is used

Figure 27 (a) shows the Mobile IP [63] handoff process of a UDP connection when the MH moves from the ON to the NN. In Figure 27 (a) t_{ch} is the one way delay between the CH and the HA, t_{hn} is the one way delay between the MH and the HA when the MH is in the NN, t_n is the one way delay between the MH and the CH while the MH is in the NN, and t_o is the one way delay between the MH and the CH while the MH is in the ON. As shown in Figure 27 (a), the MH starts layer 2 handoff to the NN and IP address acquisition from the NN at time A. These procedures are completed at time B. Then, at time B the MH starts Mobile IP [63] registration with its HA. The new CoA of the MH gets successfully registered at the HA at time instant C. Thus, packets received by the HA after time C are correctly forwarded to the MH in the NN. The packets received by MH's HA from the CH between time G and C are lost as they are forwarded to MH's old CoA. We refer to handoff latency as the time elapsed after the MH receives the last packet in the ON until the MH receives the first packet in the NN. Next, we derive the mathematical formulations for *handoff latency*, *packet loss during handoff*, and *end-to-end delay* of a UDP connection when Mobile IP [63] is used as the mobility management protocol.

4.5.1.1 Handoff Latency

From Figure 27 (a), the handoff latency of the UDP connection is given by

$$T_{h3} = D - A = \tau_{L2} + \tau_a + \tau_m + t_{ch} + t_{hn} \tag{45}$$

where τ_{L2}, τ_a, and τ_m are as defined in Section 4.4.1.

4.5.1.2 Packet Loss

From Figure 27 (a), the packets that are intercepted by the HA between time G and C are lost. Therefore, if the packet transmission rate of the CH is R, the number of

packets that are lost during handoff is given by

$$P_h = R(C - G) = R(\tau_{L2} + \tau_a + \tau_m + t_{ho}) \tag{46}$$

4.5.1.3 End-to-end packet transportation delay

The end-to-end packet transportation delay of the VoIP data packets in the path from the MH to the CH D_{fm} without RLP D_{fmnr} and with RLP D_{fmr} are, respectively, given by

$$D_{fmnr} = D + t_{wcn} \tag{47}$$

and

$$D_{fmr} = D + (K_p - 1)\tau + t_{wcn} \tag{48}$$

D is the link-layer access delay and t_{wcn} is the one way delay in the wired network between the new BS (NBS) and the CH. $K_p = \lceil \frac{L_p}{L_f} \rceil$ is the number of link layer frames per one VoIP data packet, where L_p is the length of one VoIP data packet and L_f is the length of a link-layer frame. Similarly, the end-to-end packet transportation delay from MH to CH path (D_{rm}) without RLP D_{rmnr} and with RLP D_{rmr} are, respectively, given by

$$D_{rmnr} = D + t_{ch} + t_{whn} \tag{49}$$

and

$$D_{rmr} = D + (K_p - 1)\tau + t_{ch} + t_{whn} \tag{50}$$

for no RLP and RLP scenarios, respectively. t_{ch} is the one way delay between the CH and the HA. t_{whn} is the one way delay in the wired network between the NBS and the HA.

4.5.2 Handoff performance of a UDP connection when SIP is used

SIP-based mobility management is proposed to eliminate the drawbacks of Mobile IP [87]. Using SIP-based mobility management, when the MH moves from the ON to the NN, it sends a new *INVITE* [69] message to the CH using the same call identifier as in the original call setup as shown in Figure 27 (b). The MH puts its new IP address in the *contact* field of SIP *INVITE* message [87]. This new IP address informs the CH about MH's change of network. Therefore, after receiving MH's new IP address, CH sends the VoIP data packets to MH's new address. Figure 27 (b) shows the SIP [69] handoff process of when the MH moves from the ON to the NN. At time A, the MH starts the handoff process to the NN. As shown in Figure 27 (b), the MH starts layer 2 handoff to the NN and IP address acquisition from the NN at time A. These procedures are completed at time B. Then, at time B the MH sends the *INVITE* message to the CH that is received by the CH at time C_1. Thus, packets sent by the CH between time $A - t_o$ and C_1 are lost as they were sent to the old IP address of the MH. Next, we derive the mathematical formulations for *handoff latency*, *packet loss during handoff*, and *end-to-end delay* of a VoIP application when SIP [69] is used as the mobility management protocol.

4.5.2.1 Handoff Latency

From Figure 27 (b), the handoff latency when SIP is used is given by

$$T_{h4} = D_1 - A = \tau_{L2} + \tau_a + 2D_{mc}$$

where τ_{L2} and τ_a are the time required for MH's L2 handoff to the NN and IP address acquisition in the NN, respectively. D_{mc} is the average one way delay to transport SIP signaling packets between the MH and the CH. SIP signaling messages can be transferred using either UDP or TCP [69]. For our analysis we consider that SIP signaling messages are transferred over UDP. Using steps similar to the derivation of (29), D_{mc} is given by

$$D_{mc} = (1-q_2)\Bigg\{B_2 + A_2\sum_{i=2}^{m} q_2^{i-1}(\gamma^{i-1}-1) + \sum_{i=m+1}^{\infty} q_2^{i-1}[A_2(\gamma^{m-1}-1) + (i-m)\gamma^{m-2}\Delta_2]\Bigg\} \quad (51)$$

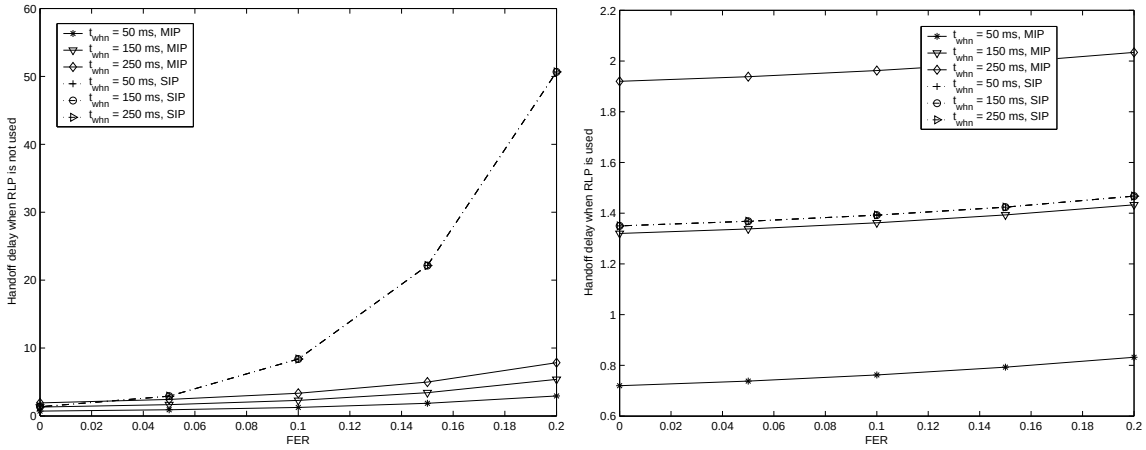
Next, we define each term in (71). B_2 is the end-to-end packet transportation delay between the MH and the CH. $B_2 = B_{2nr}$ when no RLP is used and $B_2 = B_{2r}$ when RLP is used. B_{2nr} is computed from (72) by using $T_{nr} = B_{2nr}$ and $t_w = t_{wcn}$. t_{wcn} is the one way delay in the wired network between the new BS (NBS) and the CH. B_{2r} is computed from (73) by using $T_r = B_{2r}$, $K = K_s$, and $t_w = t_{wcn}$. $K_s = \lceil \frac{L_s}{L_f} \rceil$ is the number of wireless link layer frames per one SIP *INVITE* message, where L_s is the length of a SIP *INVITE* message and L_f is the length of a link-layer frame.

q_2 is the end-to-end packet loss probability between the MH and the CH. $q_2 = q_{2nr}$ when no RLP is used and $q_2 = q_{2r}$ when RLP is used. q_{2nr} is computed from (74) by using $p_{nr} = q_{2nr}$ and $K = K_s$. q_{2r} is computed from (75) by using $p_r = q_{2r}$ and $K = K_s$.

Δ_2 is the initial value of the retransmission timer for SIP signaling messages, which is large enough to account for the size of the SIP signaling messages, twice the round trip time between the MH and the CH, and at least an additional 100 ms to allow for processing the messages at the MH and the CH. γ is the factor by which the retransmission timeout (RTO) duration is incremented after each failed retransmission. Typically, $\gamma = 2$. $A_2 = \frac{\Delta_2}{\gamma - 1}$. m is an integer such that after mth retransmission timeout the retransmission timer is frozen.

4.5.2.2 Packet Loss

From Figure 27 (b), the packets that are transmitted by the CH between time $A - t_o$ and C_1 are lost. Therefore, if the packet transmission rate of the CH is R, the number



(a) (b)

Figure 28: Handoff latency comparison of Mobile IP and SIP (a) no RLP and (b) RLP.

of packets that are lost during handoff is given by

$$P_{h1} = R(C_1 - A + t_o) = R(\tau_{L2} + \tau_a + D_{mc} + t_o).$$

4.5.2.3 End-to-end packet transportation delay

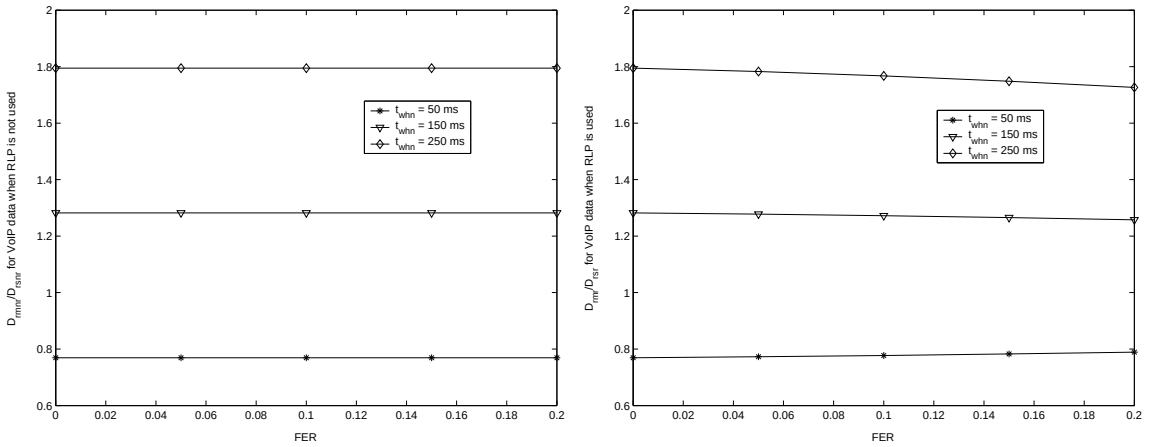
The end-to-end packet transportation delay of the VoIP data packets in the path from MH to the CH and reverse path are same for both no RLP and RLP scenarios and are given by

$$D_{fsnr} = D_{rsnr} = D + t_{wcn} \tag{52}$$

and

$$D_{fsr} = D_{rsr} = D + (K_p - 1)\tau + t_{wcn} \tag{53}$$

where D_{fsnr} and D_{fsr} are the end-to-end packet transportation delay of the VoIP data packets in the path from MH to the CH for no RLP and RLP scenarios, respectively. Similarly, D_{rsnr} and D_{rsr} are the end-to-end packet transportation delay of the VoIP data packets in the reverse path for no RLP and RLP scenarios, respectively. Therefore, when SIP is used, the packet transportation delay is symmetric in both directions. This is because there is no packet redirection when SIP is used.



(a) (b)

Figure 29: End-to-end packet transportation delay comparison of Mobile IP and SIP (a) no RLP and (b) RLP.

4.5.3 Handoff performance comparison of Mobile IP and SIP for a UDP connection

To compare the handoff performance of Mobile IP (MIP) and SIP based mobility management, we assume the following parameters. Length of SIP *INVITE* message (L_s) and *Mobile IP Registration Request/Reply* message (L_m) are 140 bytes and 56 bytes [19], respectively. We consider the numerical values specified in Section 4.4.3 for other parameters. We consider that the length of one VoIP data packet is 87 bytes [72] that includes 20 bytes of IP header, 14 bytes of IP options, 8 bytes of UDP header, and 45 bytes of RTP message (33 bytes of voice data and 12 bytes of RTP header). The 33 bytes of voice data is generated by a GSM codec in every 20 ms. When Mobile IP is used, packets are tunneled from the HA to the MH. This adds another 20 bytes of IP header making the total IP packet of length 107 bytes.

Figure 28 (a) shows the handoff latency comparison of Mobile IP and SIP for different values of FER (p_f) when no RLP is used in the link layer. It shows that for smaller values of p_f the handoff latency of SIP is lower than that of the Mobile IP. On the other hand, for larger value of p_f the handoff latency of SIP is higher than that of Mobile IP. This can be explained as follows. There are two factors that

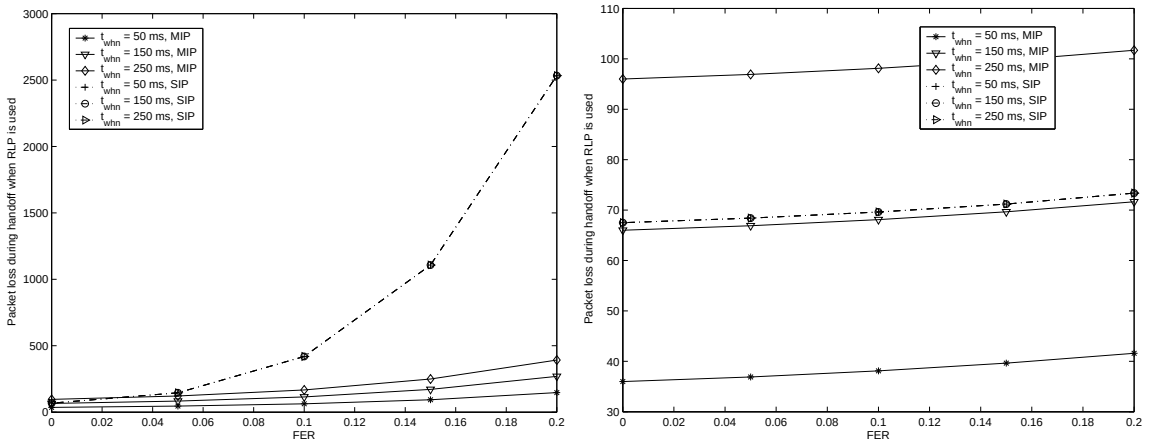


Figure 30: Packet loss during handoff comparison of Mobile IP and SIP (a) no RLP and (b) RLP.

decide the numerical value of handoff delay. One of them is the delay to transfer a handoff signaling message across the link layer and the other is the delay to transfer the handoff signaling message in the wired network. The delay across the wireless link depends on the number of link layer frames and the numerical value of link layer FER. The larger the size of a packet the higher is the probability that it gets erroneous during its transfer over the link layer. Therefore, the number of retransmissions required for the successful transfer of the signaling message increases. This increases the average signaling delay. As the size of a SIP handoff signaling message is larger than that of the Mobile IP handoff signaling message (length of SIP *INVITE* message (L_s) and *Mobile IP Registration Request/Reply* message (L_m) are 140 bytes and 56 bytes [19], respectively.), the SIP messages require more number of retransmissions. This results in higher handoff latency for SIP compared to Mobile IP. Moreover, the difference in the handoff latency between SIP and Mobile IP becomes larger as link layer FER increases. The other part of the handoff latency that incurred because of the handoff signaling transportation delay over the wired network depends on the distance between the entities involved in the handoff process. In case of SIP, the handoff signaling messages are exchanged between the MH and the CH. Whereas in

case of Mobile IP, the handoff signaling messages are exchanged between the MH and the HA. In most cases, the distance between the MH and the HA is larger than the distance between the MH and the CH. Therefore, the wired part of the handoff latency is larger for Mobile IP than SIP. When the wireless link FER is low or its effect is reduced through the use of link layer RLP, the delay in the wired network influences the overall handoff latency. Therefore, for such scenarios Mobile IP has higher handoff latency. On the other hand, for higher value of wireless link FER, the delay over the wireless link plays the major role making the handoff latency of SIP larger than that of Mobile IP. The results shown in Figure 28 (a) and Figure 28 (b) verify this. Figure 28 (a) shows that the handoff latency for SIP is lower than that of Mobile IP for lower values of FER and higher for higher values of FER. On the other hand, when the effect of link layer FER is reduced through the use of link layer RLP, SIP sometimes has lower handoff latency compared to Mobile IP as shown in Figure 28 (b).

The number of packets lost during a handoff is proportional to the handoff latency. Therefore, the number of lost packets for SIP and Mobile IP have similar nature as the handoff delay. These results are shown in Figure 30 (a) and Figure 30 (b). When no RLP is used, SIP suffers from higher packet loss during handoff for higher values of FER. This is because as a SIP *INVITE* message is larger than the *Mobile Registration Request/Reply* message, the probability that a SIP *INVITE* message is lost over the link-layer is higher. This increases the average handoff delay resulting in higher packet loss. However, when RLP is used as this link layer loss is compensated by link layer retransmissions, the longer length of the SIP *INVITE* message does not come into picture. Since, most of the current wireless systems implement RLP, and the distance between the MH and the HA is usually higher, Mobile IP is expected to suffer from higher packet losses.

Figure 29 (a) and Figure 29 (b) show the end-to-end packet transportation delay

comparison of Mobile IP and SIP for different values of FER (p_f). The results show that Mobile IP has higher end-to-end packet transportation delay for all values of FER. This is because the packets follow triangular routes instead of straight routes between the MH and the CH when Mobile IP is used. On the other hand, packets follow the direct path between the MH and the CH when SIP is used. This is one of the major disadvantages of using Mobile IP based mobility management for real-time applications. The results show that the end-to-end packet transportation delay of Mobile IP can be 80% higher than of SIP (the actual value depends on the particular network conditions). This would go higher depending on the distance between the MH and the HA. Since real-time applications such as VoIP require minimum end-to-end delay, SIP based mobility management is preferred over Mobile IP. When RLP is not implemented in the link layer, the packet transportation delay across the link layer remains independent of the numerical value of link layer FER. Therefore, the end-to-end packet transportation delay ratio for Mobile IP and SIP remains the same for all values of link layer FER. This is verified by the results shown in Figure 29 (a). On the other hand, when RLP is implemented at the link layer, the number of retransmissions that are required for the successful transfer of a VoIP data packet across the link layer increases as the link layer FER increases. Therefore, the effect of triangular routing of Mobile IP on the end-to-end packet transportation delay reduces. This reduces the ratio of end-to-end packet transportation delay as the FER increases as shown in Figure 29 (b).

4.6 Summary and Conclusions

To summarize, our analysis shows that the handoff performance of a mobility management protocol depends on the following factors.

- **Type of application:** Different applications use different transport layer protocols. As the operating principles of different transport layer protocols are

different, they react differently to the handoff. Therefore, the performance of a particular mobility management protocol is different for different types of applications. For example, as discussed earlier, the handoff latency of Mobile IP based handoff is larger for applications using TCP than applications using UDP. This is because when packets are lost during the handoff, TCP went through retransmission timeouts before retransmitting the lost packets.

- **Link layer frame error probability:** Our analysis shows that the handoff latency, end-to-end packet transportation delay, and packet loss during handoff depends on the link layer frame error probability (p_f) both when no RLP is used and when RLP is used.

- **Signaling delay:** Handoff latency and packet loss during handoff depend on the signaling delay between the network entities that are involved in a handoff, e.g., MH and HA in case of Mobile IP and MH and CH in case of SIP and TCP-Migrate.

- **Link layer access technologies:** As observed in our analysis, different types of link layer access technologies such as use of RLP also influence the numerical value of handoff parameters. Moreover, the link layer access delay that is different for different access technologies also influence the handoff performance.

Based on our handoff performance investigation, we advocate the use of TCP-Migrate for applications using TCP, i.e., *Class B* and *Class C* applications. SIP is suitable for real-time applications using UDP. However, SIP is standardized only for real-time applications, therefore Mobile IP can be used for non-real time applications that use UDP. In summary, different mobility management protocols operating from different layers of the classical protocol stack are suitable for different classes of applications. The use of application adaptive mobility itself is not enough to support seamless mobility management. This is revealed in our analysis where we observe

that the handoff performance depends heavily on link layer FER, the delay between different network entities that are involved in the handoff, and the wireless access technology. Therefore, we advocate information sharing between different layers to enhance the performance of mobility management. This cross-layering approach will eliminate the negative effects of different parameters such as link layer frame error rate and signaling delay on the handoff performance of mobility management protocols.

CHAPTER V

APPLICATION ADAPTIVE MULTI-LAYER HANDOFF MANAGEMENT IN NEXT-GENERATION WIRELESS SYSTEMS

5.1 Introduction

Different types of applications react differently to handoffs. To understand the effect of handoffs on mobile applications, we classified them into five categories: *Class A* through *Class E*, based on their mobility management requirements in [58]. These application classes are summarized below.

- *Class A Applications*: These include TCP and UDP applications that are short lived and originated by a mobile host (MH) such as Domain Name Service (DNS) resolution [46] [73]. These applications do not require location management or handoff support [58].

- *Class B Applications*: These include TCP applications that are long lived and originated by an MH such as web browsing and telnet sessions. These applications do not require location management support but require handoff support [58].

- *Class C Applications*: TCP applications that are long lived and terminated at an MH such as telnet sessions belong to Class C applications. These applications require both location management and handoff support [58].

- *Class D Applications*: These include UDP applications that are long lived and originated by an MH such as mobile telephony where MH is the calling party. These applications require only handoff support [58].

- *Class E Applications*: UDP applications that are long lived and terminated at an MH such as mobile telephony where MH is the called party constitute Class E applications. These applications require both location management and handoff support [58].

Out of the above five application classes, *Class A* applications do not require any mobility support [58]. Other application classes require support from mobility management protocols to hide the effects of handoffs from mobile users. The effect of handoffs on different application classes can be represented in terms of the following parameters.

- *Handoff latency:* Handoff latency is the time period for which no communication is possible between an MH and a Correspondent Host (CH) during a handoff. Handoff latency influences the performance of different classes of applications in the following ways.

  - *Packet loss:* When a transport layer protocol does not have mechanisms to recover the packets that are lost during handoffs, packet loss occurs for a time period equal to handoff latency. *Class D* and *Class E* applications run over UDP. As UDP is not a reliable protocol, these applications suffer from packet loss during handoffs. On the other hand, *Class B* and *Class C* applications that use TCP do not experience packet loss during handoffs as long as the handoff duration is less than the timeout period of a TCP connection. This is because as TCP is a reliable protocol, the packets that are lost during handoffs are recovered through TCP's retransmission mechanisms. When a mobility management protocol operates from the transport layer, there is no packet loss during handoffs for *Class B* and *Class C* applications even when the handoff duration is larger than the TCP's timeout period. This is because in this case TCP is aware about

the handoff and waits until the handoff is completed instead of closing the connection.

  - *Throughput degradation time:* TCP reacts to packet loss through invocation of its congestion control mechanisms, increment of retransmission timeout, and trigger of slow start. After the completion of a handoff, TCP takes finite amount of time to return to its steady state operation. During this time, TCP achieves throughput that is lower than the maximum achievable throughput. Therefore, *Class B* and *Class C* applications that use TCP experience throughput degradation beyond handoff completion. On the other hand, as UDP is not reactive to packet loss, *Class D* and *Class E* applications that use UDP return to normal operation right after handoff completion.

- *End-to-end delay:* When a mobility management protocol implements redirection of packets as a part of mobility support, the end-to-end delay between the MH and the CH increases when the MH is away from its home domain. This increase in end-to-end delay is detrimental for *Class D* and *Class E* applications that are real-time in nature.

- *Transport-layer transparency:* A TCP connection between the MH and the CH maintains connection states at the MH and the CH. A connection state includes receive and send buffers, congestion control parameters, and sequence and acknowledgement number parameters. Therefore, TCP based applications, e.g., *Class B* and *Class C* applications, require that a mobility management protocol keeps the transport layer connection states transparent to handoffs. On the other hand, as UDP based applications, e.g., *Class D* and *Class E* applications, do not maintain any connection state; they do not require transport layer transparency during handoffs.

- *Security:* The level of security requirements of applications depends whether the MH is in its home network or residing in a foreign network. While the MH is inside a home network domain, applications do not require strict security mechanisms. On the other hand, while in a foreign domain or while communicating with CHs that are in foreign domains, the applications may require strict security mechanisms. In general, security is important for all classes of applications.

Among different handoff performance parameters specified above, *handoff latency* and *security* are important for all mobile application classes. On the other hand, *end-to-end delay* is important for *Class D* and *Class E* applications that are real time in nature while *transport-layer transparency* is important for *Class B* and *Class C* applications [58].

Handoff management protocols operating from different layers of the classical protocol stack (e.g., link, network, transport, and application layers) are proposed in the literature [13]. Mobile IP [63] that operates from the network layer is proposed to support mobility in IP-based networks. It forwards packets to mobile users that are away from their home networks using IP-in-IP tunnels [63]. Transport layer mobility management protocols are proposed to support mobility between networks. These protocols eliminate the need for tunneling of the data streams. TCP-Migrate is proposed in [73] to support end-to-end transport layer mobility management. An architecture called MSOCKS is proposed in [51] for transport layer mobility. MSOCKS implements transport layer mobility using a split-connection proxy architecture and a new technique called TCP Splice that gives split-connection proxy systems the same end-to-end semantics as usual TCP connections [51]. An end-to-end approach for transparent layer mobility across is proposed in [41]. Moreover, work is going on in the IETF to modify the Stream Control Transmission Protocol [76] to allow it to dynamically change endpoint addresses in the midst of a connection [29] [39]. Recently,

application layer mobility using Session Initiation Protocol (SIP) is proposed in [87]. SIP based mobility does not require any changes to the IP stack of the mobile users. In addition, device independent personal mobility and location services are supported by SIP mobility.

However, our analysis in Chapter 4 shows that the above protocols achieve different performance results with respect to different handoff parameters. For example, while Session Initiation Protocol (SIP) [69] based mobility management and TCP-Migrate [73] achieve minimum *end-to-end delay*, TCP-Migrate and Mobile IP [63] achieve *transport-layer transparency*. On the other hand, Mobile IP introduces additional *end-to-end delay*. Similarly, SIP based mobility management does not provide *transport-layer transparency* to the applications.

The above discussion concludes that it is not feasible to support efficient handoff management for all application classes by using only one mobility management protocol. In this Chapter, we propose the use of different mobility management protocols for different application classes. According to our proposed solution, the mobility management protocol that achieves good performance results with respect to handoff parameters that are important for a particular application class is used for that application class. As mobility management protocols operating from different layers are used for different application classes, we call the proposed approach as adaptive multi-layer handoff management framework (AMMF). Although multi-layer handoff management provides best available handoff support to all application classes, this is not enough to support seamless handoff support. (By seamless handoff support, we mean handoff support with minimum or ideally zero handoff latency. This corresponds to minimum or ideally zero packet loss and throughput degradation time during handoffs.) To address this problem, we propose to share information between different layers. This cross-layering approach eliminates the negative effects of different parameters such as link layer frame error rate (FER) and handoff signaling delay

on the handoff performance. Thus, our proposed AMMF has two fold advantages. First, it achieves application adaptive mobility support. Second, it further improves the performance of application adaptive handoff support through cross-layer interactions. Performance evaluation shows that AMMF achieves efficient handoff support for all application classes.

5.2 Design guidelines for application adaptive handoff support

As discussed in the previous section, our proposed application adaptive multi-layer handoff management framework uses different mobility management protocols for different application classes. To answer the question-**What is the suitable mobility management protocol for a particular application class**, we carried out detailed handoff performance investigation of the existing mobility management protocols when they are used for different application classes in [58]. Based on the results of our mathematical analysis, we advocate the use of TCP-Migrate [73] for *Class B* and *Class C* applications, SIP [69] for *Class D* and *Class E* applications that are real time in nature, and Mobile IP [63] for *Class D* and *Class E* applications that are non-real time in nature. In this case, the mobility management protocols operate from only one layer and are agnostic about the dynamics of other layers. Therefore, their handoff performance varies based on the dynamics of the other layers. The parameters that influence the handoff performance are as follows [58].

- **Link layer access technologies:** Different types of link layer access technologies such as the presence or absence of Radio Link Protocol (RLP) [20] in the link layer influence handoff latency.

- **Signaling delay:** Handoff latency depends on the signaling delay between the network entities that are involved during a handoff, e.g., the MH and home

agent (HA) in case of Mobile IP and the MH and the CH in case of SIP and TCP-Migrate.

- **Link layer frame error rate (FER):** Handoff latency also depends on the link layer FER.

Based on the above factors, the applications experience finite amount of handoff latency and the resulting performance degradation. Our analysis in [58] shows that the performance degradation experienced by the applications during handoffs depends solely on handoff latency. Thus, the handoff performance of different mobility management protocols can be improved by reducing their handoff latency. Moreover, this reduction in handoff latency must be achieved irrespective of the dynamics of different layers. Therefore our objective is to achieve minimum handoff latency under all circumstances. We propose to share information between different layers and then use this information to reduce handoff latency.

To provide further insights to our proposed handoff latency reduction technique, we first describe the mechanisms that decide the handoff latency of the mobility management protocols. Towards this, next we describe the different steps for the handoff process of the existing mobility management protocols.

5.2.1 Different Steps for Handoff Process of the Existing Mobility Management Protocols

We consider a scenario where an MH is moving from an Old Network (ON) to a New Network (NN) in the middle of its communication with a CH. The handoff process of the existing mobility management protocols for this scenario can be divided into the following steps.

- *Step 1*: Detection of an MH's movement to the NN. In case of Mobile IP, this is done using the *Agent Advertisement* messages [63]. In case of SIP and TCP-Migrate this functionality is assumed to be supported by the underlaying link

layer. Link layer algorithms detect MH's movement into a different network using received signal strength (RSS) information from the neighboring base stations (BSs).

- *Step 2*: Once the MH's movement to the NN is detected, the MH performs layer 2 (L2) handoff to the NN.

- *Step 3*: After completing L2 handoff, the MH acquires a new IP address from the NN. This can be done in several ways, e.g., by using Dynamic Host Configuration Protocol (DHCP). Step 2 and Step 3 are similar for all the mobility management protocols.

- *Step 4*: Once the MH obtains a new IP address from the NN, the next step is to register MH's new IP address with the appropriate network entities, e.g., the HA in case of Mobile IP and the CH in case of TCP-Migrate and SIP. The address registration procedures are different for different mobility management protocols and can be summarized as follows.

  - In case of Mobile IP, the MH registers its new IP address with its HA as its new care-of-address (CoA). Then, the HA tunnels the packets destined for the MH to MH's new CoA.

  - In case of TCP Migrate, the MH sends its new IP address to the CH. Then, the CH modifies the identifier of the TCP connection to reflect the change of network address as defined in [73] and resumes its operation to MH's new address.

  - In case of SIP [87], the MH sends an *INVITE* message [69] to the CH upon whose reception the CH re-establishes the connection to MH's new IP address.

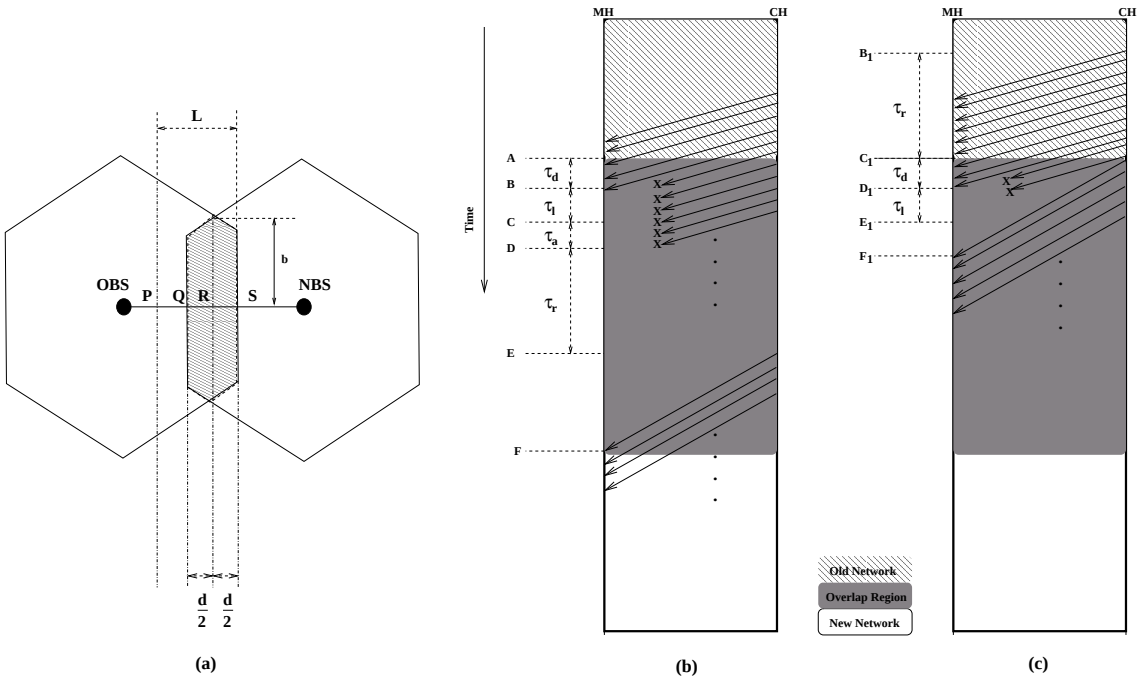


Figure 31: Coverage area of the OBS and the NBS.

Every mobility management protocol executes the above steps during a handoff. Therefore, the time required to complete a handoff is the sum of time required to carry out each of these steps. We denote this time by T, which is given by

$$T = \tau_d + \tau_l + \tau_a + \tau_r \tag{54}$$

where τ_d, τ_l, τ_a, and τ_r are the time required by the MH for the detection of movement to the NN, L2 handoff to the NN, IP address acquisition from the NN, and IP address registration, respectively. Based on the implementation procedures, some of the handoff steps can be carried out without interrupting MH's ongoing communications, whereas to carry out other steps, the MH is required to halt its ongoing communications. Handoff latency is the time for which the MH can not communicate during a handoff. To determine the handoff latency of the existing mobility management protocols, next we present their implementation details using Figure 31 (a) that shows two cells between which the MH is moving. We assume that these two cells belong to two different network domains. Therefore, when the MH moves from the

old BS (OBS) of the ON to the new BS (NBS) of the NN, it gets a new IP address from the NN. The coverage areas of the OBS and the NBS overlap in the hatched region in Figure 31 (a). The MH detects the presence of the NN when it crosses the point Q, i.e., at time A as shown in Figure 31 (b). The movement detection process is completed at time B. It may be noted that during the movement detection, the MH continues its communication through the OBS. At time B, the MH starts L2 handoff to the NBS. After time B, the packets that are destined to MH's old address can not be delivered to it as shown in Figure 31 (b). Therefore, they are dropped starting from time B. Once L2 handoff is completed at time C, the MH obtains a new IP address from the NN, which it registers with the appropriate network entity at time E (the network entity is HA for Mobile IP and CH for TCP-Migrate and SIP). After the address registration, the MH receives the packets through the NBS. Therefore, the MH can not communicate from time B to time E. Thus, while *step 1* does not contribute towards handoff latency, *step 2* through *step 4* contribute to handoff latency. The expression for handoff latency of the existing mobility management protocols is given by

$$T_h = \tau_l + \tau_a + \tau_r. \tag{55}$$

τ_l and τ_a are same for all the existing mobility management protocols. On the other hand, τ_r is different for different mobility management protocols and depends on their specific address registration procedures. We denote the τ_r for Mobile IP, TCP-Migrate, and SIP by τ_{rm}, τ_{rt}, and τ_{rs}, respectively. The handoff latency of these protocols can be computed from (55) by using the corresponding value of τ_r. The handoff latency of Mobile IP (T_{hm}), TCP-Migrate (T_{ht}), and SIP (T_{hs}) are respectively given by

$$T_{hm} = \tau_l + \tau_a + \tau_{rm}, \tag{56}$$

$$T_{ht} = \tau_l + \tau_a + \tau_{rt}, \tag{57}$$

and

$$T_{hs} = \tau_l + \tau_a + \tau_{rs}. \tag{58}$$

5.2.2 Proposed Handoff Latency Reduction Approach

Out of *step 2* through to *step 4* of the handoff process discussed earlier, while *step 2* can be performed only after the MH enters the coverage area of the NBS, *step 3* and *step 4* can be performed before MH's movement into the coverage area of the NBS. However, for this, the MH is required to learn the NN in advance. Therefore, MH's movement has to be predicted. Based on this prediction, the MH obtains an IP address from the NN and performs the address registration before it moves into the NN. When the MH detects its movement into the NN, it needs to perform only L2 handoff. In this case, MH's ongoing communications are interrupted for the time duration of L2 handoff. Therefore, the handoff latency is reduced to

$$\hat{T_h} = \tau_l. \tag{59}$$

To summarize, to reduce handoff latency, we propose to predict the NN in advance, carry out address acquisition and address registration before MH's movement into the NN, and carry out only L2 handoff when the MH moves into the NN. The approach of carrying out some of the handoff tasks before MH's movement into the NN is proposed in [58] and [50] to reduce Mobile IP handoff latency. However, these approaches are specific to Mobile IP and can not be used for our proposed multi-layer mobility management framework.

As address acquisition does not interrupt MH's ongoing communications, it can be carried out once the NN is predicted. Now, the next question is, **when should the MH start the address registration procedures?** We answer this question below. Let us assume that the MH starts the address registration at time B_1 as shown

in Figure 31 (c). Ideally, the address registration should be started such that it is completed when the first packet sent to MH's new address reaches the MH between the time the MH detects its movement to the NN and completes its L2 handoff to the NN, i.e., between time D_1 and E_1 in Figure 31 (c). (The packets that reach the NBS while the MH carries out its L2 handoff to the NN are lost. However, as L2 handoff time is very small (of the order 10-20 ms), these packet losses are not very severe.) The MH starts to receive packets at its new IP address after its L2 handoff. Thus, if the address registration with the NN are carried out in such a way that they are completed when the MH completes its movement detection to the NN, the handoff latency is limited to τ_l. Therefore, B_1 is the right time for the MH to start address registration. B_1 depends on the time required for address registration τ_r. If the MH is located at point P at time B_1, then the distance between P and S (S is the point located in the boundary of the cell covered by the OBS), L, is given by

$$L = v\tau_r + d \tag{60}$$

where v is the speed of the MH and d is the length of the overlap region. When the MH is at a distance L from the boundary of the cell served by the OBS, it starts the address registration. The instant when the MH is at a distance L from the boundary of the OBS is determined as described in Section 5.2.3. To compute L, v and τ_r are required. We estimate v as described in Section 5.3.1.2. We estimate τ_r for SIP, TCP-Migrate, SIP, i.e., τ_{rs}, τ_{rt}, τ_{rm} in Section 5.3.1.6, Section 5.3.1.7, and Section 5.3.1.8, respectively.

We denote the value of L for Mobile IP, TCP-Migrate, and SIP by L_m, L_t, and L_s, respectively. The expression for L_m can be derived from (60) by using $L = L_m$ and $\tau_r = \tau_{rm}$. Similarly, expressions for L_t and L_s can be derived from (60) by using $L = L_t$, $\tau_r = \tau_{rt}$ and $L = L_s$, $\tau_r = \tau_{rs}$, respectively.

5.2.3 Estimation of time for address acquisition

To determine the instant at which the MH is at a distance L from the boundary of the cell served by the OBS, i.e., to determine when the MH crosses the point P in Figure 31 (a), we first estimate the expected RSS when the MH is at point P. We denote this RSS as S. Then, we monitor the RSS of the MH and when the RSS become equal to or less than S, we learn that the MH has crossed the point P. The numerical value of S is calculated as follows.

$$S = 10 \log_{10}[P_r(a - L)] \tag{61}$$

where $P_r(a - L)$ is the RSS when the MH is at a distance of $a - L$ from the serving OBS, where a is the distance between the OBS and S in Figure 31 (a). To determine the RSS at the distance of $a - L$ from the OBS, we use the path loss model given by [77]

$$P_r(x) = P_r(d_0) \left(\frac{d_0}{x} \right)^{\alpha} + \epsilon \tag{62}$$

where x is the distance between the base station and MH, $P_r(d_0)$ is the received power at a known reference distance, which is in the far field of the transmitting antenna. Typical value of d_0 is 1 km for macrocells, 100 m for outdoor microcells, and 1 m for indoor pico cells [77]. The numerical value of $P_r(d_0)$ depends on different factors such as frequency, antenna heights, and antenna gains. α is the path loss exponent. The numerical value of α is dependent on the cell size and local terrain characteristics. The typical value of α ranges from 3 to 4 and 2 to 8 for a typical macro-cellular and micro-cellular environment, respectively. ϵ is a zero-mean Gaussian random variable that represents the statistical variation in $P_r(x)$ caused by shadowing. Typical standard deviation of ϵ is 8 dB [77]. Its actual value depends on the cell size.

5.3 Application Adaptive Multi-layer handoff management framework

The adaptive multi-layer handoff management framework (AMMF) operates on the following two step approach. First, it selects a mobility management protocol that is best suitable for that application. Then, it estimates the handoff signaling delay in advance and initiates the handoff procedures of the selected mobility management protocol at an appropriate time so that the handoff latency is minimized. We describe the architecture of AMMF followed by its operating principles in the subsequent subsections.

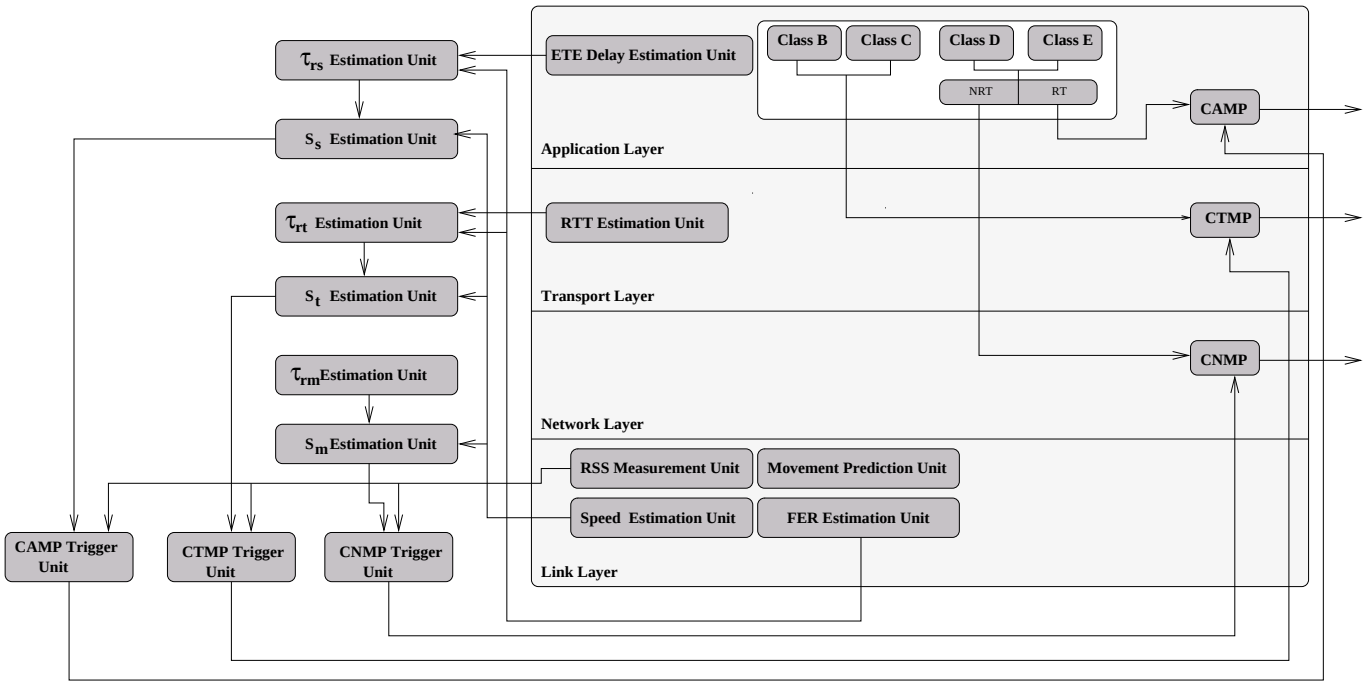


Figure 32: Architecture of the multi-layer mobility management framework.

5.3.1 Architecture of AMMF

The architecture of our proposed adaptive multi-layer handoff management framework (AMMF) is shown in Figure A.1. NRT in Figure A.1 refers to non-real time Class D and Class E applications. Similarly, RT refers to real time Class D and Class E applications. As shown in Figure A.1, the use of information from different

layers enables the cross-operation of the handoff management protocols. We refer to the cross-layer application, transport, and network layer mobility management as cross-layer application layer mobility protocol (CAMP), cross-layer transport layer mobility protocol (CTMP), and cross-layer network layer mobility protocol (CNMP), respectively. CAMP, CTMP, and CNMP use SIP, TCP-Migrate, and Mobile IP, respectively. Moreover, they use the proposed handoff reduction technique to improve the handoff performance. These protocols are summarized below.

- CAMP: CAMP is used for real time Class D and Class E applications. τ_{rs} *estimation unit* estimates the τ_{rs} (address registration delay of SIP) using FER, link layer access technology, and end-to-end (ETE) delay information as discussed in Section 5.3.1.6 . Then, the S_s *estimation unit* computes S_s using τ_{rs} and v, and informs *CAMP handoff trigger unit* about it. S_s is computed from (61) using $S = S_s$ and $L = L_s$, where L_s is computed from (60) using $L = L_s$ and $\tau_r = \tau_{rs}$. The *CAMP handoff trigger unit* initiates the handoff procedures of CAMP when the RSS of MH's serving BS drops below S_s.

- CTMP: CTMP is used for Class B and Class C applications. τ_{rt} *estimation unit* estimates the τ_{rt} (address registration delay of TCP-Migrate) using FER, link layer access technology, and TCP's RTT information as discussed in Section 5.3.1.7 . Then, the S_t *estimation unit* computes S_t using τ_{rt} and v, and informs *CTMP handoff trigger unit* about it. S_t is computed from (61) using $S = S_t$ and $L = L_t$, where L_t is computed from (60) using $L = L_t$ and $\tau_r = \tau_{rt}$. The *CTMP handoff trigger unit* initiates the handoff procedures of CTMP when the RSS of MH's serving BS drops below S_t.

- CNMP: CNMP is used for non-real time Class D and Class E applications. τ_{rm} *estimation unit* estimates the τ_{rm} (address registration delay of Mobile IP) as discussed in Section 5.3.1.8. Then, the S_m *estimation unit* determines S_m using

τ_{rm} and v, and informs *CNMP handoff trigger unit* about it. S_m is computed from (61) using $S = S_m$ and $L = L_m$, where L_m is computed from (60) using $L = L_m$ and $\tau_r = \tau_{rm}$. The *CTMP handoff trigger unit* initiates the handoff procedures of CNMP when the RSS of MH's serving BS drops below S_m.

The different modules of AMMF and their functionalities are as follows.

5.3.1.1 Movement prediction unit

It predicts the NN using the movement prediction algorithm proposed in [12].

5.3.1.2 Speed estimation unit

It estimates the speed of the MH. We use our own algorithm, VEPSD (velocity estimation using the power spectral density of the received signal envelope), proposed in [56] to estimate MH's speed. In [56], we estimate MH's speed using the information about the maximum Doppler frequency (f_m) that we capture from the received signal envelope. f_m is related to v, speed of light in free space (c), and the carrier frequency of the received signal (f_c) through

$$v = \left(\frac{c}{f_c}\right) f_m. \tag{63}$$

VEPSD estimates f_m using the slope of the power spectral density (PSD) of the received signal envelope. The slope of PSD of the receive signal envelope has maximum values at frequencies $f_c \pm f_m$ in mobile environments [56]. VEPSD detects the maximum value of received signal envelope's PSD that corresponds to the highest frequency component ($f_c + f_m$) to estimate f_m. We select this algorithm over other speed estimation algorithms such as [16] [40] because the latter suffer from larger estimation errors [56].

5.3.1.3 FER estimation unit

It estimates the link layer frame error rate (FER). In practice, the wireless MAC protocols have information about FER [7]. *FER estimation unit* collects FER

information from the MAC layer.

5.3.1.4 RTT estimation unit

It learns the RTT of a TCP connection using the state variables of TCP.

5.3.1.5 ETE delay estimation unit

It estimates the application layer end-to-end (ETE) delay of real time application when SIP is used. The ETE delay depends on the presence or absence of Radio Link Protocol (RLP) in the link layer. The ETE delay without RLP ED_{nr} and with RLP ED_r are, respectively, given by [58]

$$ED_{nr} = D + t_{wco} \tag{64}$$

and

$$ED_r = T_f + (K-1)\tau + t_{wco} \tag{65}$$

where D is the link layer access delay, t_{wco} is the delay in the wired link between the OBS and the CH, and $K = \lceil \frac{L_p}{L_f} \rceil$ is the number of link layer frames per one RTP data packet. L_p is the length of one RTP packet and L_f is the length of one link layer frame.

The formulation for T_f is given by [20]

$$T_f = D(1-p_f) + \sum_{i=1}^{n}\sum_{j=1}^{i} P(C_{i,j})(2iD + 2(j-1)\tau) \tag{66}$$

where p_f is the FER and τ is the link layer inter-frame interval, which is typically around 20 ms [20]. $P(C_{i,j})$ is the probability that the first frame transmitted by the MH is received correctly by the BS, being the ith retransmitted frame at the jth retransmission trial. The expression for $P(C_{i,j})$ is given by [20]

$$P(C_{i,j}) = p_f(1-p_f)^2((2-p_f)p_f)^{(\frac{i^2-i}{2}+j-1)} \quad \text{for } i = 1, 2, ..., n \text{ and } j = 1, 2, .., i \tag{67}$$

We estimate ED_{nr} and ED_r using the time stamp information carried in the real-time protocol (RTP) header. When the MH receives a new RTP packet it obtains a new sample value for the end-to-end delay by comparing the time stamp field of the RTP packet to its own clock. Then, it updates the estimated end-to-end delay ED_{nr} and ED_r as follows

$$ED_{nr} = (1 - x)ED_{nr} + xT_{nr} \tag{68}$$

$$ED_r = (1 - x)ED_r + xT_r \tag{69}$$

where T_{nr} and T_{nr} are the instantaneous sampled value of ETE delay without RLP and with RLP, respectively. We consider typical value of $x = 0.125$.

5.3.1.6 τ_{rs} estimation unit

It estimates the address registration delay of SIP (τ_{rs}). The expression for τ_{rs} is derived in [58] and is given

$$\tau_{rs} = 2D_{mc} \tag{70}$$

where D_{mc} is the average one way delay to transport SIP signaling packets between the MH and the CH. SIP signaling messages can be transferred using either UDP or TCP [69]. For our analysis we consider that SIP signaling messages are transferred over UDP. Then, D_{mc} is given by [58]

$$\begin{aligned} D_{mc} &= (1-q)\Bigg\{B + A\sum_{i=2}^{m} q^{i-1}(\gamma^{i-1} - 1) + \\ &\qquad \sum_{i=m+1}^{\infty} q^{i-1}[A(\gamma^{m-1} - 1) + (i-m)\gamma^{m-2}\Delta]\Bigg\} \end{aligned} \tag{71}$$

Next, we define each term in (71). B is the end-to-end packet transportation delay between the MH and the CH. $B = B_{nr}$ when no RLP is used and $B = B_r$ when RLP

is used. The expressions for B_{nr} and B_r are, respectively, given by [58]

$$B_{nr} = D + t_{wco} \tag{72}$$

and

$$B_r = T_f + (K_s - 1)\tau + t_{wco} \tag{73}$$

where t_{wco} is the delay in the wired link between the OBS and the CH. D is the link-layer access delay. $K_s = \lceil \frac{L_s}{L_f} \rceil$ is the number of wireless link layer frames per one SIP *INVITE* message, where L_s is the length of a SIP *INVITE* message. T_f is given by (66).

q is the end-to-end packet loss probability between the MH and the CH. $q = q_{nr}$ when no RLP is used and $q = q_r$ when RLP is used. The expressions for q_{nr} and q_r are, respectively, given by [58]

$$p_{nr} = 1 - (1 - p_f)^{K_s}(1 - p_c) \tag{74}$$

$$p_r = 1 - \left[1 - p_f((2 - p_f)p_f)^{\frac{(n^2+n)}{2}}\right]^{K_s}(1 - p_c) \tag{75}$$

where p_f is the FER, p_c is the packet loss probability in the wired network between the MH and the CH, and n is the maximum number of trials that the RLP carries out before aborting the attempt to transmit a frame over the link layer. Typically, $n = 3$ for RLP [20].

Δ is the initial value of the retransmission timer for SIP signaling messages, which is large enough to account for the size of the SIP signaling messages, twice the round trip time between the MH and the CH, and at least an additional 100 ms to allow for processing the messages at the MH and the CH. γ is the factor by which the

retransmission timeout (RTO) duration is incremented after each failed retransmission. Typically, $\gamma = 2$. $A = \frac{\Delta}{\gamma - 1}$. m is an integer such that after mth retransmission timeout the retransmission timer is frozen.

Therefore, the numerical value of τ_{rs} can be estimated using the following information.

- Whether or not RLP is implemented at the link layer.
- The link layer FER.
- B

Out of these, the first information is already known to the MH. FER (p_f) is collected from *FER estimation unit*. We estimate B using the ETE delay information that is estimated by the *ETE delay estimation unit*. Using (72), (73), (64) and (65), the expression for B_{nr} and B_r are, respectively, given by

$$B_{nr} = ED_{nr} \tag{76}$$

and

$$B_r = ED_r - (K-1)\tau + (K_s - 1)\tau \tag{77}$$

Now, τ_{rs} is estimated using (70).

5.3.1.7 τ_{rt} estimation unit

It estimates the address registration delay of TCP-Migrate (τ_{rt}). The expression for average value of τ_{rt} is derived in [58] and is given by

$$\tau_{rt} = \sum_{i=0}^{N_m - 1} \sum_{j=0}^{N_m - 1} \sum_{k=0}^{N_m - 1} P_h(i,j,k) L_h(i,j,k) \tag{78}$$

where N_m is such that TCP abort a connection establishment attempt after N_m number of retransmissions. $P_h(i,j,k)$ is given by [58]

$$P_h(i,j,k) = p_1^i (1-p_1) p_2^j (1-p_2) p_2^k (1-p_2) \quad \text{for } i,\ j,\ k = 0,\ 1,\ 2,\ ...,N_m\text{-}1 \tag{79}$$

where p_1 is the end-to-end packet loss probability between the MH and the CH for a SYN packet and p_2 is the end-to-end packet loss probability between the MH and the CH for a SYN/ACK or ACK packet. $p_1 = p_{1nr}$ when no RLP is used and $p_1 = p_{1r}$ when RLP is used. p_{1nr} is computed from (74) by using $q_{nr} = p_{1nr}$ and $K_s = K_1$. p_{1r} is computed from (75) by using $q_r = p_{1r}$ and $K_s = K_1$. $K_1 = \lceil \frac{L_1}{L_f} \rceil$ is the number of link layer frames per one SYN packet. L_1 is the length of the SYN packet and L_f is the length of a link-layer frame. Similarly, $p_2 = p_{2nr}$ when no RLP is used and $p_2 = p_{2r}$ when RLP is used. p_{2nr} is computed from (74) by using $q_{nr} = p_{2nr}$ and $K_s = K_2$. p_{2r} is computed from (75) by using $q_r = p_{2r}$ and $K_s = K_2$. $K_2 = \lceil \frac{L_2}{L_f} \rceil$ is the number of link layer frames per one SYN/ACK or ACK packet. L_2 is the length of the SYN/ACK or ACK packet.

The expression for $L_h(i, j, k)$ is given by [58]

$$\begin{aligned} L_h(i,j,k) &= 1.5RTT_o + \sum_{m=0}^{i-1} 2^m RTO + \sum_{m=0}^{j-1} 2^m RTO + \sum_{m=0}^{k-1} 2^m RTO \\ &= 1.5RTT_o + (2^i + 2^j + 2^k - 3)RTO \\ &\qquad \text{for } i,\ j,\ k = 0,\ 1,\ 2,\ ...,N_m\text{-}1 \end{aligned} \tag{80}$$

where RTO is the initial retransmission time out for the TCP connection and $RTO = \xi RTT_o$. ξ is a constant weighting factor. RTT_o is TCP round trip time (RTT) in the ON. Therefore, the numerical value of τ_{rt} can be estimated using the following information.

- Whether or not RLP is implemented at the link layer.
- The link layer FER.
- RTT_o

Out of these, the first information is already known to the MH. It obtains the FER information from the *FER estimation unit*. RTT_o is collected from the *RTT estimation*

unit. Then, τ_{rt} is estimated using (78).

5.3.1.8 τ_{rm} estimation unit

It estimates the address registration delay of Mobile IP (τ_{rm}). τ_{rm} is equal to the time required for Mobile IP registration process. The numerical value of τ_{rm} depends on the delay between the MH and its HA. We propose a simple technique that uses the Mobile IP protocol to estimate τ_{rm}. The MH sends the Mobile IP registration messages to the HA with an invalid *Mobile-HA Authentication Extension*. The objective of using invalid *Authentication Extension* is to just learn the address registration signaling delay without changing the mobility binding at the HA. When the HA receives the Mobile IP registration messages and learns the presence of the invalid *Authentication Extension*, it returns the Mobile IP *Registration Reply* with appropriate code [63] that signifies MH failed authentication. Then τ_{rm} is estimated by comparing the time difference between the transmission time of Mobile IP registration request and the reception time of Mobile IP registration reply. This technique introduces extra signaling overhead to the system. However, we advocate its use because of its simplicity. Moreover, this technique can be implemented using the existing Mobile IP protocol, hence no extra implementation is required.

5.3.1.9 CAMP trigger unit

It collects τ_{rs} and v information from the τ_{rs} *estimation unit* and *speed estimation unit*. It determines the value of L_s using $L = L_s$ and $\tau = \tau_{rs}$ in (60). Then, it calculates the dynamic RSS threshold for SIP address registration, S_s by using $S = S_s$ and $L = L_s$ in (61). When the RSS of MH's serving BS drops below S_s, the *CAMP trigger unit* sends the trigger to *CAMP* for handoff execution.

5.3.1.10 CTMP trigger unit

It collects τ_{rt} and v information from the τ_{rt} *estimation unit* and *speed estimation unit*. It determines the value of L_t using $L = L_t$ and $\tau = \tau_{rt}$ in (60). Then, it calculates the dynamic RSS threshold for TCP-Migrate address registration, S_t by using $S = S_t$ and $L = L_t$ in (61). When the RSS of MH's serving BS drops below S_t, the *CTMP trigger unit* sends the trigger to *CTMP* for handoff execution.

5.3.1.11 CNMP trigger unit

It collects τ_{rm} and v information from the τ_{rm} *estimation unit* and *speed estimation unit*. It determines the value of L_m using $L = L_m$ and $\tau = \tau_{rm}$ in (60). Then, it calculates the dynamic RSS threshold for SIP address registration, S_m by using $S = S_m$ and $L = L_m$ in (61). When the RSS of MH's serving BS drops below S_m, the *CNMP trigger unit* sends the trigger to *CNMP* for handoff execution.

The operation of AMMF is shown in the flow chart in Figure 33. First, the MH anticipates a handoff and predicts the NN. Then, based on the type of applications, it selects one of the mobility management protocols. This is followed by the estimation of handoff signaling delay for that mobility management protocol. Then, the handoff initiation time is determined. The MH initiates the address acquisition from the NN at the handoff initiation time.

5.4 Analytical Modeling for the Performance Evaluation of AMMF

We illustrate the handoff process in AMMF using Figure 31 that shows two cells between which the MH is moving. When the existing mobility protocols are used, the MH moving from the OBS to the NBS initiates a handoff to the NBS when it crosses the point Q in Figure 31 (a). In this case, the handoff latency of the exiting mobility management protocols is given by (55). In AMMF, we start the handoff

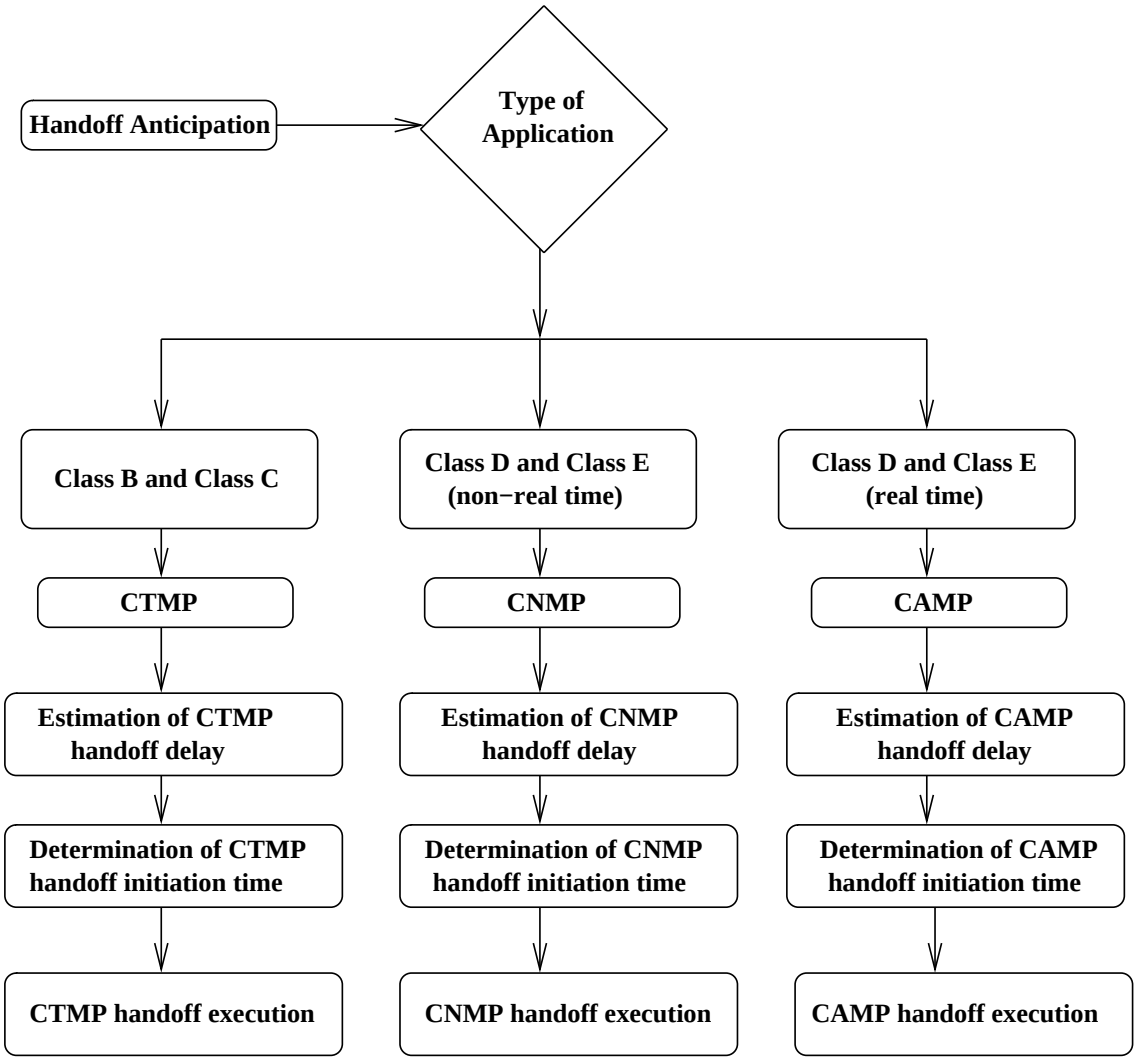


Figure 33: Flow chart showing the operation of multi-layer mobility management framework.

process before the MH enters the overlap region between the OBS and the NBS. We consider that in AMMF the handoff procedures are initiated when the MH crosses the point P as shown in Figure 31 (a). Here, we assume that the MH is going to move to the predicted BS. In this case, one of the following scenarios may occur:

- The NN prediction is incorrect, i.e., the MH moves to a BS other than the predicted BS. It may be noted that the MH learns about the unsuccessful prediction of the BS by comparing the ID of the NBS (the MH learns about the ID of NBS from the BS advertisement messages that it receives after it enters

the overlap region between the OBS and NBS) with the ID of the predicted BS. In this case, IP address acquisition and address registration that were carried out for the predicted BS are wasted and the MH initiates handoff to the NBS after crossing the point Q shown in Figure 31 (a). This scenario is similar to the handoff process when the existing mobility management protocols are used. In this case the handoff delay is given by (55). We denote the probability of the occurrence of this case as p_{c1}, where p_{c1} = probability of unsuccessful prediction of the New Network (NN). We denote the handoff latency of this event by T_{hc1}, which is given by (55).

- **Case 2:** The MH moves to the BS predicted by the *movement prediction unit*. We denote the probability of the occurrence of this case as p_{c2}, where p_{c2} = probability of successful prediction of the NN. $p_{c2} = 1 - p_{c1}$. We denote the handoff latency of this event by T_{hc2}, which we derive below. We carry out the following analysis for CAMP. Therefore, the address registration time is τ_{rs}. It may be noted that the similar analysis can be used for CTMP and CNMP by using $\tau_{rs} = \tau_{rt}$ and $\tau_{rs} = \tau_{rm}$, respectively, in the following formulations.

In Case 2, one of the following scenarios may occur.

1. MH's address acquisition and address registration processes of CAMP are completed before the MH enters the overlap region. We refer to this situation as the early completion of address registration and denote the probability of this event by p_e. The expression for p_e is given by

$$p_e = p(t < \tau_{rs}) \tag{81}$$

where τ_{rs} is the time required for address registration and t is the time taken

by the MH to go from P to Q. The expression for the pdf of t is given by [57]

$$f_t(t) = \begin{cases} \frac{L-d}{\theta t\sqrt{v^2t^2-(L-d)^2}}, & \frac{L-d}{v} < t < \frac{\sqrt{b^2+(L-d)^2}}{v} \\ 0 & \text{otherwise} \end{cases} \quad (82)$$

where $\theta = \arctan(\frac{b}{L-d})$. $b = \frac{a}{2} + \frac{d}{\sqrt{3}}$ using Figure 31 (a). L is the distance of the point P from the point S in Figure 31 (a). d is the length of the overlap region in Figure 31 (a). v is the speed with which the MH is moving. Therefore,

$$\begin{aligned} p_e &= \int_0^{\tau_{rs}} f_t(t)dt \\ &= \int_{\frac{L-d}{v}}^{\tau_{rs}} \frac{L-d}{\pi t\sqrt{v^2t^2-(L-d)^2}}dt \\ &\approx \frac{1}{\theta}\arccos\left(\frac{L-d}{v\tau_{rs}}\right) \end{aligned} \quad (83)$$

In this case, the packets destined for the MH start to arrive in the NN before the MH moves to the NN. Therefore, if no further action is taken these packets are dropped by the NN until the MH enters the overlap area shown in Figure 31 (a). In this case, the handoff latency is given by

$$T_{he} = \frac{L}{v} - \tau_{rs} + \tau_d + \tau_l \quad (84)$$

where τ_d is the time required for the detection of NN and τ_l is the time required for L2 handoff to the NN.

2. Address registration is completed after the MH enters the overlap area. We denote the probability of this event by p_u and the corresponding handoff latency by T_{hu}. In this case, one of the following three situations may occur.

   - Address registration is completed during the movement detection. We denote the probability of this scenario as p_{u1}. The expression for p_{u1} is given by

$$p_{u1} = p\left(\frac{L-d}{v} < t < \frac{L-d}{v} + \tau_d\right) \tag{85}$$

In this case, the handoff latency is given by

$$T_{u1} = \tau_l + \frac{L-d}{v} + \tau_d - \tau_{rs} \tag{86}$$

- Address registration is completed after the movement detection but before the completion of L2 handoff. We denote the probability of this scenario as p_{u2}. The expression for p_{u2} is given by

$$p_{u2} = p\left(\frac{L-d}{v} + \tau_d < t < \frac{L-d}{v} + \tau_d + \tau_l\right) \tag{87}$$

In this case, the handoff latency is given by

$$T_{u2} = \tau_l \tag{88}$$

- Address registration is completed after the L2 handoff. We denote the probability of this scenario as p_{u3}. The expression for p_{u3} is given by

$$p_{u3} = p\left(\frac{L-d}{v} + \tau_d + \tau_l < t < \frac{L}{v}\right) \tag{89}$$

In this case, the handoff latency is given by

$$T_{u3} = \tau_{rs} - \frac{L-d}{v} - \tau_d \tag{90}$$

Therefore, p_u and T_{hu} are, respectively, given by

$$p_u = p_{u1} + p_{u2} + p_{u3} \tag{91}$$

and

$$T_{hu} = p_{u1}T_{u1} + p_{u2}T_{u2} + p_{u3}T_{u3} \tag{92}$$

The numerical value of p_{u1}, p_{u2}, and p_{u3} can be calculated using the procedure used for the derivation of (83).

Now, using (83) and (91), the expression for p_{c2} is given by

$$p_{c2} = p_e + p_u \tag{93}$$

The handoff latency when the NN is predicted successfully T_{hc2} is given by

$$T_{hc2} = p_e T_{he} + p_u T_{hu} \tag{94}$$

Then, the average handoff latency of CAMP is

$$\hat{T_{hs}} = p_{c1}T_{hc1} + (1 - p_{c1})T_{hc2} \tag{95}$$

where T_{hc1} is given by (55) as discussed earlier and T_{hc2} is given by (94). Similarly, we compute the average handoff latency of CTMP $\hat{T_{ht}}$ and CNMP $\hat{T_{hm}}$.

The packet losses of CNMP and CAMP and throughput degradation time of CTMP depends on the handoff latency as follows.

5.4.1 Packet Loss of Mobile IP and SIP

The number of packets that are lost during CNMP (P_{hm}) and CAMP (P_{hs}) based handoff in AMMF are, respectively, given by

$$P_{hm} = R\hat{T_{hm}} \tag{96}$$

and

$$P_{hs} = R\hat{T_{hs}} \tag{97}$$

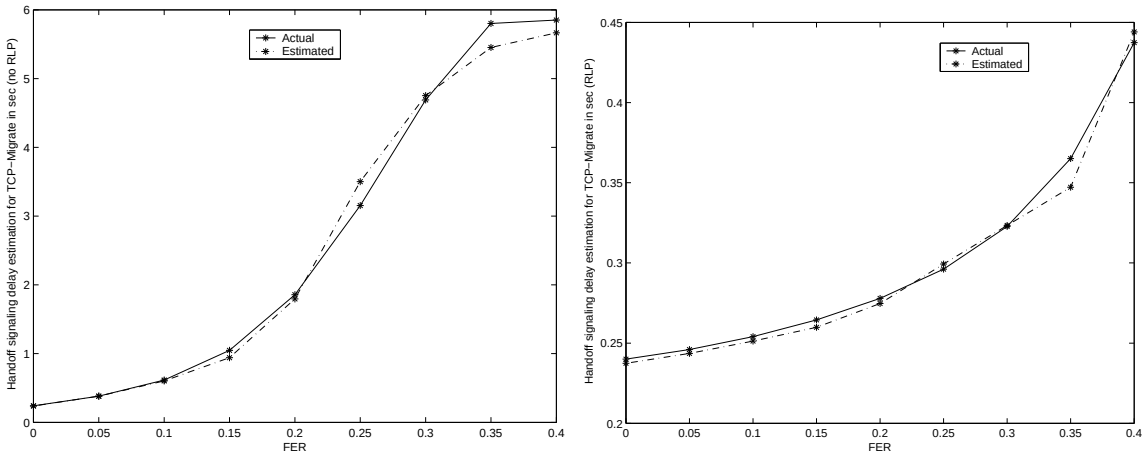
where R is the data rate of the connection between the CH and the CH.

5.4.2 Throughput Degradation Time of TCP-Migrate

The expression for the throughput degradation time of CTMP is given by

$$T_t = \hat{T_{ht}} + [1 + log_2 CW_n]RTT_n \tag{98}$$

where CW_n is the steady state congestion size of TCP in the NN and RTT_n is the TCP's RTT in the NN.

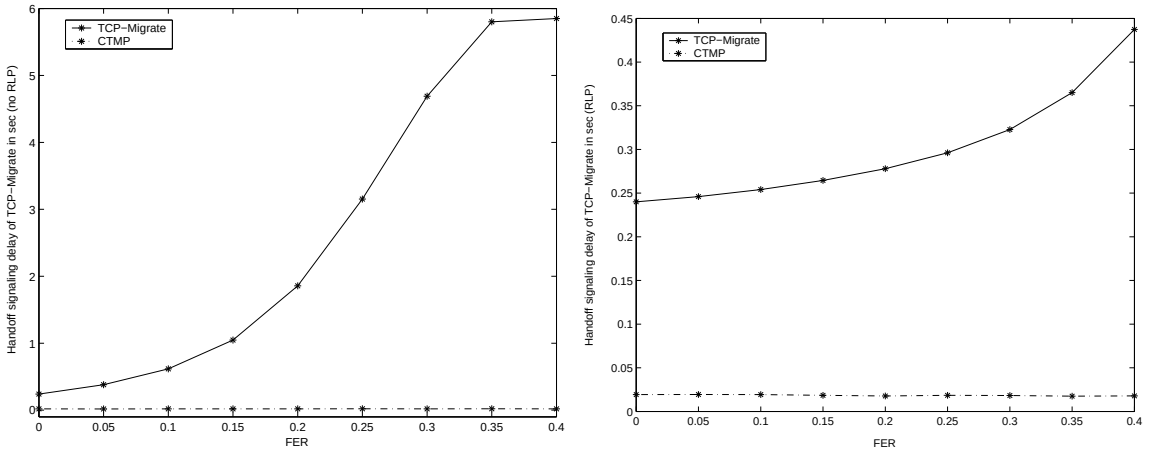


(a) (b)

Figure 34: Handoff signaling delay estimation for TCP-Migrate (a) no RLP and (b) RLP.

5.5 Performance Evaluation of AMMF

To investigate the performance of our proposed AMMF, we consider the scenario shown in Figure 31 (a). We then compare the handoff performance of the existing mobility management protocols with our proposed AMMF for different classes of applications. We first present the results for *Class B* and *Class C* applications that



(a) (b)

Figure 35: Handoff signaling delay comparison of CTMP and TCP-Migrate (a) no RLP and (b) RLP.

use CTMP. Then, we present the results for non-real time *Class D* and *Class E* applications that use CNMP followed by the results for real time *Class D* and *Class E* applications that use CAMP. We also compare the handoff performance of CTMP with TCP-Migrate, CNMP with Mobile IP, and CAMP with SIP.

In our simulation experiments, we assume the following values for different parameters: the time required for MH's L2 handoff to the NN $\tau_{L2} = 10$ ms, the time required for IP address acquisition in the NN $\tau_a = 20$ ms, the time required for the detection of the NN $\tau_d = 10$ ms, the time required for L2 handoff $\tau_l = 10$ ms, one way delay between the CH and the HA $t_{ch} = 50$ ms, link layer access delay $D = 10, 50, 150$ ms for WLAN, 3G cellular, and satellite networks, respectively [7], length of link-layer frame $L_f = 19$ bytes, link-layer inter-frame interval $\tau = 20$ ms, one way delay in the wired network between the old BS (OBS) and the CH $t_{wco} = 100$ ms, one way delay in the wired network between the new BS (NBS) and the CH $t_{wcn} = 100$ ms, packet loss probability in the wired network $p_c = 1e - 5$.

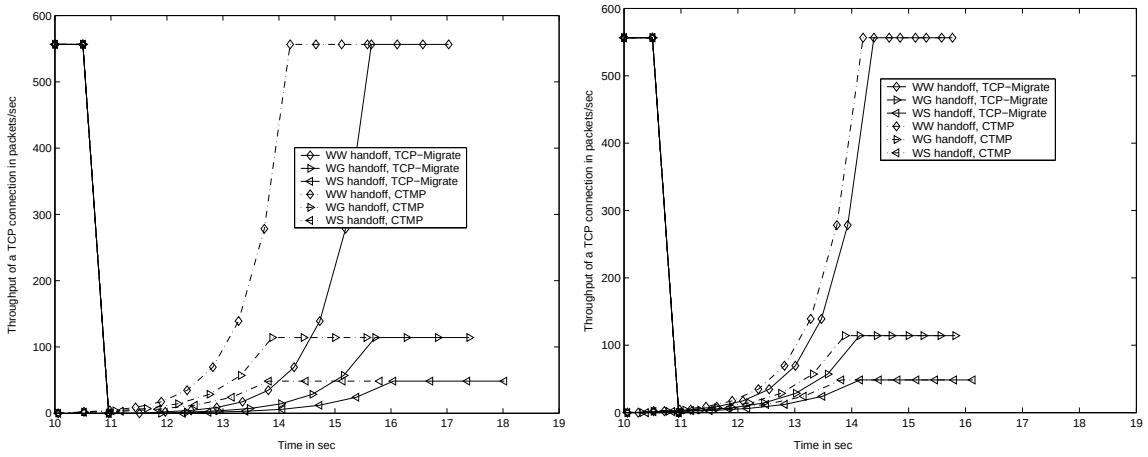


Figure 36: Throughout degradation time comparison of CTMP and TCP-Migrate (a) no RLP and (b) RLP.

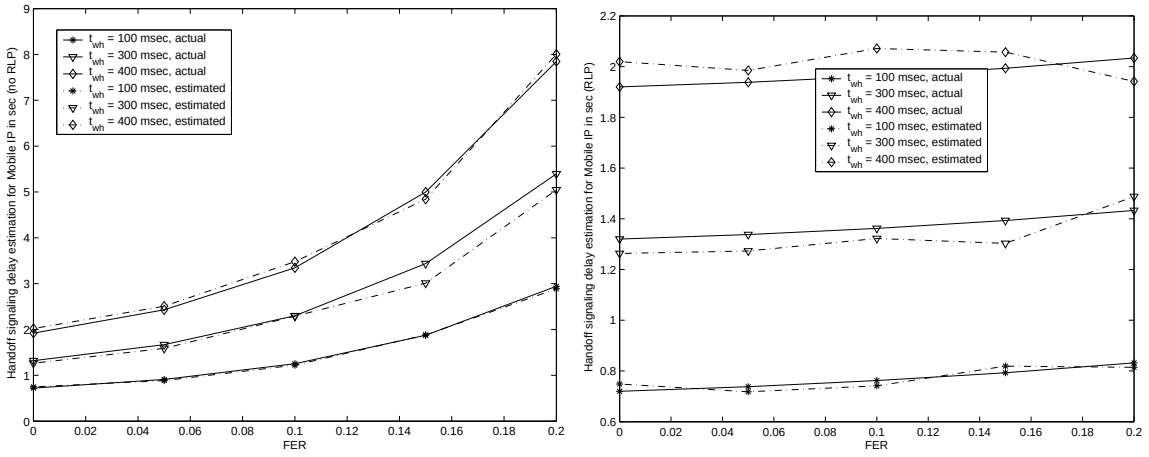
5.5.1 *Class B* and *Class C* Applications (CTMP)

We collect the estimated value of RTT from the TCP state variables. Then, using RTT and the link layer FER, we determine the handoff signaling delay for TCP-Migrate using (78). Figure 34 (a) and Figure 34 (b) show the actual and estimated value of TCP-Migrate handoff signaling delay for no RLP and RLP scenarios, respectively. The results show that for both RLP and no RLP scenarios, TCP-Migrate's handoff signaling delay is estimated close to its actual value for different values of link layer FER. This estimated handoff signaling delay is used in CTMP for pre-handoff execution of TCP-Migrate. We compare the handoff latency of TCP-Migrate and CTMP in Figure 35 (a) and Figure 35 (b) for no RLP and RLP scenarios, respectively. The results show that handoff latency of CTMP is significantly lower than that of TCP-Migrate. Moreover, The handoff latency does not depend on link layer FER or the access technology (i.e., the presence or absence of RLP in the link layer). Therefore, CTMP eliminates the effect of these parameters on the handoff performance making the handoff performance independent of the access technology and FER. To compare the throughput degradation time of CTMP and TCP-Migrate,

we consider a scenario where an MH previously in a WLAN moves to a WLAN or 3G cellular or Satellite network. We refer to the handoff from a WLAN to another WLAN network as WW handoff. Similarly, WLAN to 3G cellular and WLAN to Satellite network handoffs are referred as WG handoff and WS handoff, respectively. We consider link layer access delay $D = 10, 50, 150$ ms for WLAN, 3G cellular, and satellite networks, respectively [7] for this simulation. We further consider that the MH moves to the NN at time 10.5 seconds. Therefore, before this time the TCP connection operates in the steady state corresponding to the ON, which is a WLAN in this case. Then, after MH's movement to the NN (either a WLAN or 3G network, or Satellite network) until the handoff process is completed the packets destined for the MH are lost resulting in zero throughput. As TCP starts from slow start after the handoff, it takes finite amount of time for TCP to reach its steady state in the NN resulting in throughput degradation even after the completion of handoff. We compare the throughput degradation time of TCP-Migrate and CTMP in Figure 36 (a) and Figure 36 (b) for no RLP and RLP scenarios, respectively. The results show that there is upto 30 % reduction in the throughput degradation time in CTMP compared to TCP-Migrate. This is a direct consequence of the previous results that show that the handoff latency of CTMP is less than that of TCP-Migrate. Moreover, comparison of Figure 36 (a) and Figure 36 (b) show that this throughput degradation time of CTMP is independent of whether or not RLP is used in the link layer.

5.5.2 Non-real Time *Class D* and *Class E* Applications (CNMP)

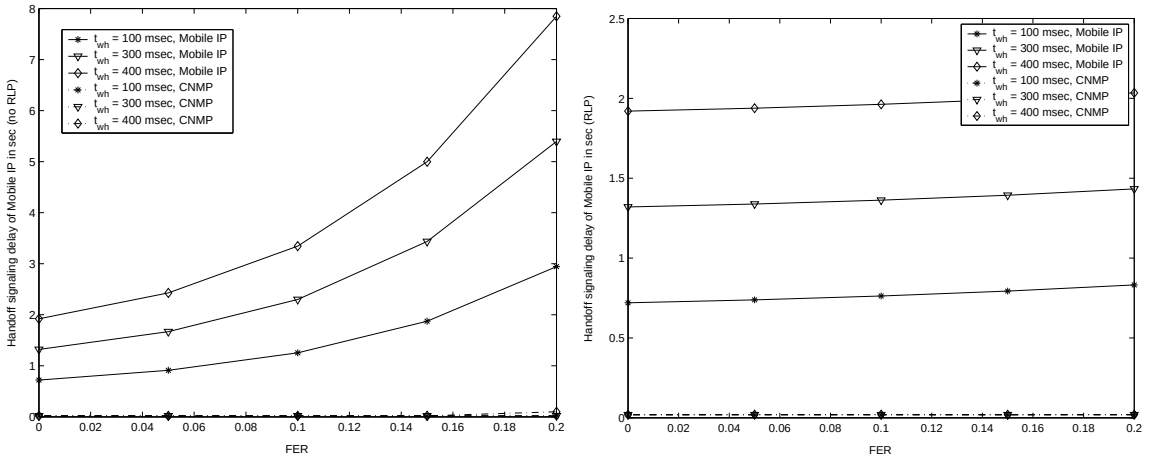
Figure 37 (a) and Figure 37 (b) show the actual and estimated value of Mobile IP handoff signaling delay for no RLP and RLP scenarios, respectively. We assume that there is 20 % error in the estimated handoff signaling delay. Based on the estimated value of handoff signaling delay, in AMMF we carry out Mobile IP pre-handoff execution to reduce the effective handoff latency. Figure 38 (a) and Figure 38



(a) (b)

Figure 37: Handoff signaling delay estimation for Mobile IP (a) no RLP and (b) RLP.

(b) show the handoff latency comparison of Mobile IP and CNMP. The results show that when CNMP is used the handoff latency is reduced significantly. In fact, the handoff latency CNMP is negligible compared to that of Mobile IP. As we can observe from Figure 38 (a) and Figure 38 (b), the handoff latency of CNMP is independent of the link layer FER. This is because the effect of FER is already captured during our handoff signaling delay estimation. Similarly, the effect of signaling delay between the MH and its HA, t_{wh}, on the handoff latency is also eliminated in CNMP through prior estimation of Mobile IP signaling delay. Moreover, as we can observe from Figure 38 (a) and Figure 38 (b) that the handoff latency is no more dependent on whether or not RLP is used in the link layer. To summarize, by using the information about link layer access technology, link layer FER, and the signaling delay between the MH and its HA (t_{wh}), CNMP eliminates the negative effect of these parameters on the handoff latency. In addition, when we use this information to estimate the handoff signaling delay in advance and accordingly initiate the pre-handoff execution processes, the numerical value of handoff latency is reduced significantly. To quantify this reduction of handoff latency on applications' performance, we show the packet loss comparison of Mobile IP and CNMP in Figure 39 (a) and Figure 39 (b) for no



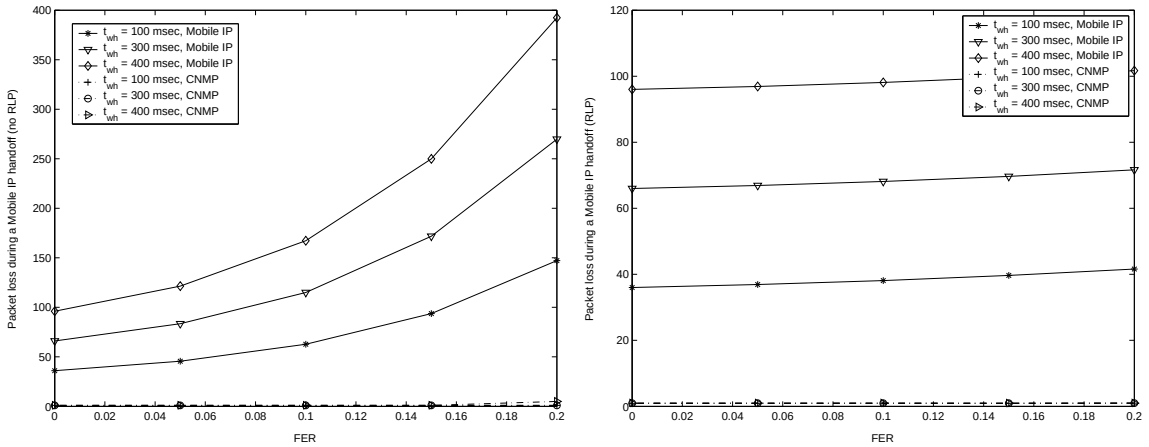
(a) (b)

Figure 38: Handoff signaling delay comparison of CNMP and Mobile IP (a) no RLP and (b) RLP.

RLP and RLP scenarios, respectively. The results show that the packet loss during handoff in CNMP is negligible compared to the packet loss when base Mobile IP is used.

5.5.3 Real Time *Class D* and *Class E* Applications (CAMP)

Figure 40 (a) and Figure 40 (b) show the actual and estimated value of SIP based handoff signaling delay for no RLP and RLP scenarios, respectively. The results are shown for different values of signaling delay between the MH and the CH, t_{wc}. Based on the estimated value of handoff signaling delay, in AMMF we carry out the procedures of pre-handoff execution of SIP based handoff to reduce the effective handoff latency. Figure 41 (a) and Figure 41 (b) show the handoff latency comparison of SIP and CAMP. The results show that there is up to 50% reduction in the handoff latency of SIP when CAMP is used. To quantify this reduction of handoff latency on application performance, we show the packet loss comparison of SIP and CAMP in Figure 42 (a) and Figure 42 (b) for no RLP and RLP scenarios, respectively. For this we considered a VoIP application. We consider that the length of one VoIP data packet is 87 bytes [72] that includes 20 bytes of IP header, 14 bytes of IP options,



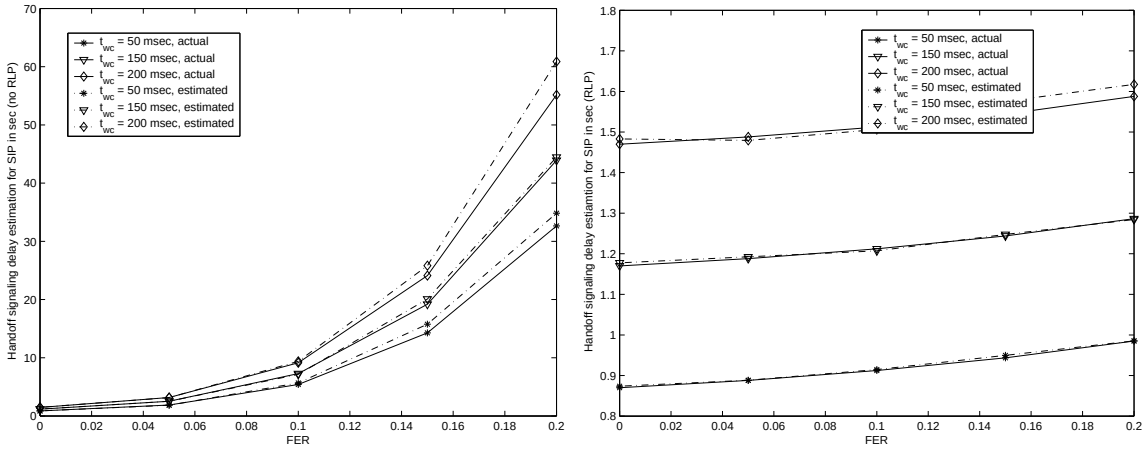
(a) (b)

Figure 39: Packet loss during handoff comparison of CNMP and Mobile IP (a) no RLP and (b) RLP.

8 bytes of UDP header, and 45 bytes of RTP message (33 bytes of voice data and 12 bytes of RTP header). The 33 bytes of voice data is generated by a GSM codec in every 20 ms. The results show that when CAMP is used, the packet loss during handoff is up to 50 % less compared to SIP.

5.6 Summary

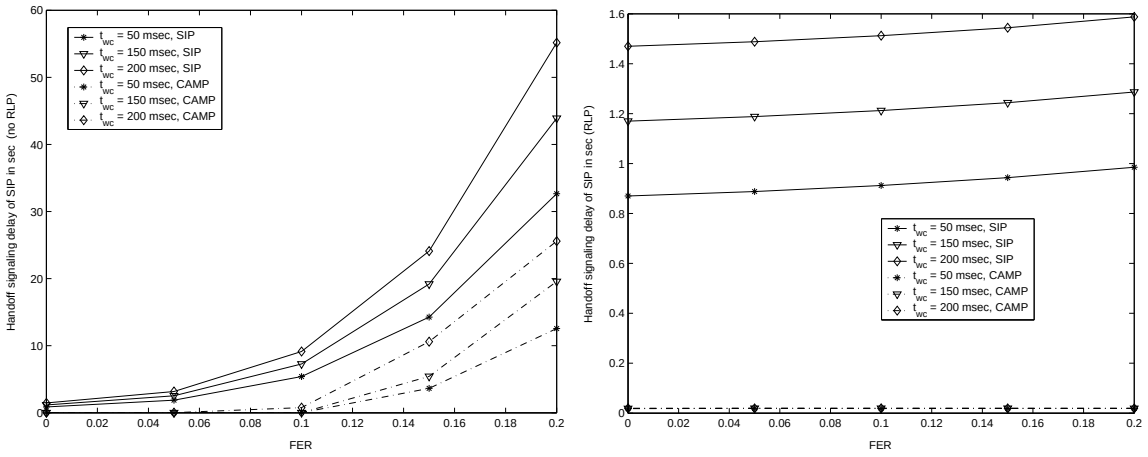
As none of the existing mobility management protocols offer handoff support suitable to different types of applications, we advocate the use of application adaptive handoff management. However, this application adaptive handoff management itself is not enough to support seamless handoff management. This is because when mobility management protocols operate from a particular layer of the network protocol stack and are unaware of the dynamics of the other layers, the handoff latency is higher, especially when wireless links suffer from higher error rate. Moreover, the signaling delay between the networking entities that are involved in handoffs also influences the handoff latency. This large value of handoff latency results in severe performance degradation during handoffs. Therefore, we propose to share information between



(a) (b)

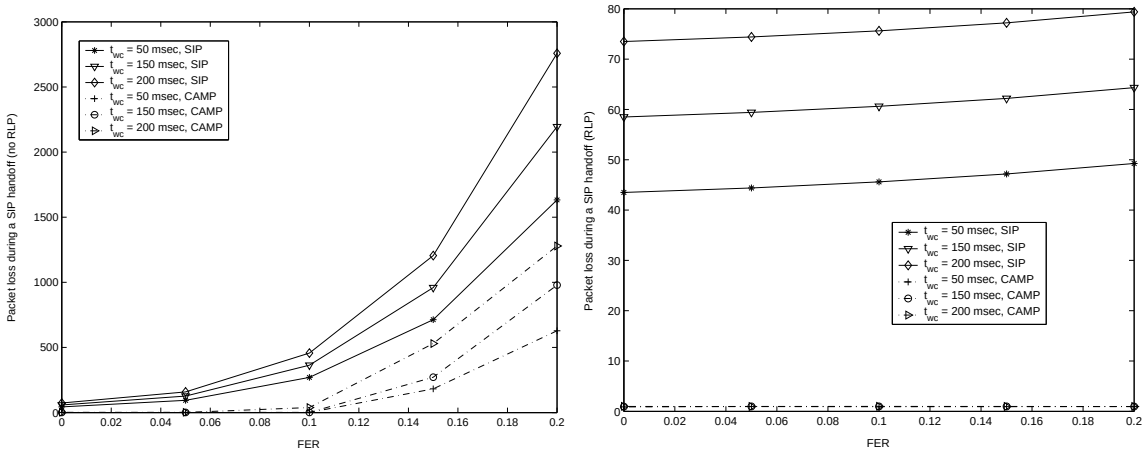
Figure 40: Handoff signaling delay estimation for SIP (a) no RLP and (b) RLP.

different layers to enhance the performance of application adaptive handoff management. This cross-layer approach eliminates the negative effects of different parameters (such as link layer frame error rate, wireless access technology, and signaling delay between the entities that are involved in the handoff process) on the handoff performance. Therefore, we propose to estimate link layer FER and signaling delay in advance and use this information to enhance the handoff performance. The basic idea is to use this information to estimate the handoff signaling delay and then decide about the appropriate time to initiate the handoff. Instead of initiating the handoff when the MH arrives in the NN, we propose that mechanisms other than L2 handoff should be completed while the MH is in the ON. This is because other procedures such as IP address acquisition in the NN, registration of new CoA with the HA, and transmission of SIP *INVITE* message to the CH can be done while the MH in the ON. Therefore, when the MH moves to the NN, these procedures need not be carried out. This significantly reduces handoff latency.



(a) (b)

Figure 41: Handoff signaling delay comparison of CAMP and SIP (a) no RLP and (b) RLP.



(a) (b)

Figure 42: Packet loss during handoff comparison of CAMP and SIP (a) no RLP and (b) RLP.

CHAPTER VI

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

6.1 Research Contributions

In this thesis, new architecture is proposed to integrate the heterogeneous wireless systems to realize a scalable architecture for NGWS. Moreover, cross-layer mobility management protocols are proposed to support seamless handoff management in NGWS. Research contributions have been made in the following areas:

1. Architecture to integrate the existing heterogeneous wireless systems.
2. Cross-layer (Layer 2 + 3) handoff management protocol for NGWS.
3. Performance analysis of handoff techniques based on Mobile IP, TCP-Migrate, and SIP.
4. Application adaptive multi-layer handoff management in NGWS.

6.1.1 AMC: A Ubiquitous Mobile Communication Architecture for Next Generation Wireless Systems

Various heterogeneous systems exist in the current wireless world. They adopt different radio technologies and have different network architectures and protocols, such as Bluetooth for personal areas, IEEE 802.11 for local areas, Universal Mobile Telecommunication System (UMTS) for wide areas, and satellite networks for global areas. These systems are designed for specific service needs and vary widely in terms of bandwidth, area of coverage, cost, and quality of service (QoS) provisioning. However, none of them can simultaneously satisfy the low-latency, high-bandwidth, and ubiquitous-coverage needs of mobile users at low cost. Since different wireless systems, each of which is optimized for some specific service demands and coverage area,

are complementary to each other, they can co-operate to provide ubiquitous "always best connection" [34] to mobile users. This necessitates the design of intelligently integrating the existing wireless systems so that the users may receive their services via the best available wireless network anytime anywhere.

In Chapter 2, a third-party-based integrated architecture, AMC, is proposed to integrate the heterogeneous wireless networks. AMC reduces the cost of architecture deployment by using the access and core network infrastructures of the existing wireless systems. AMC integrates heterogeneous wireless systems of different operators who may not necessarily have direct SLAs among them. Therefore, it is scalable. Furthermore, security equivalent to the existing wireless systems is achieved under AMC. Finally, advanced link layer sensing algorithms and neighbor discovery protocols are developed to achieve seamless inter-system handoff by reducing the connection interruption and handoff failure during inter-system handover. Performance evaluation results shows that AMC achieves significant reduction in the number of required SLAs compared to the existing bilateral SLA based architectures. As AMC is a centralized third-party-based architecture, it can afford greater control over heterogeneous networks for providing authentication, service agreement, mobility management, etc. It avoids problems of distributed coordination among individual networks. However, it may create a single point of failure and the third-party may become a bottleneck, reducing performance. Advanced solutions are needed to take care of the reliability and scalability issues of AMC. The hierarchical NIA architecture as discussed in Section 2.7 can resolve the bottleneck problem and still maintain the benefits of centralized control.

6.1.2 A Cross-Layer (Layer 2 + 3) Handoff Management Protocol for Next Generation Wireless Systems

In the integrated NGWS, users are always connected to the best available networks and switch between different networks based on their service needs. It is an important

and challenging issue to support seamless mobility management in NGWS. Mobility management contains two components: location management and handoff management. Location management enables the system to track the locations of mobile users between consecutive communications. On the other hand, handoff management is the process by which a user keeps its connection active when it moves from one base station (BS) to another. There exist efficient location management techniques in the literature for NGWS. However, seamless support of handoff management in NGWS is still an open issue.

In Chapter 3, a cross-layer mobility management protocol called CMP is proposed. CMP estimates users' speed and predicts the handoff signaling delay of possible handoffs. CMP uses this information to estimate the appropriate instance for handoff initiation. Performance analysis and simulation results show that CMP significantly enhances the performance of both intra- and inter-system handoffs. CMP also significantly reduces the cost associated with the false handoff initiation because it achieves lower false handoff initiation probability.

6.1.3 Performance Analysis of Handoff Techniques based on Mobile IP, TCP-Migrate, and SIP

The cross-layer mobility management protocol, CMP, proposed in Chapter 3 uses Mobile IP to support mobility management. However, Mobile IP suffers from different performance issues such as triangular routing, higher global signaling load, and the requirement of new network entities. Therefore, mobility management protocols operating from transport layer and application layer are proposed to eliminate the limitations of Mobile IP. However, all these mobility management protocols achieve different performance results with respect to different handoff parameters. In parallel, different types of applications have different requirements in terms of handoff performance parameters. Therefore, none of the existing mobility management protocols can support efficient handoff management for all types mobile application.

This necessitate the need to develop novel approaches to support seamless handoff management to all types of applications.

In Chapter 4, to understand the effect of handoffs on mobile applications, different applications are classified into five categories: *Class A* through *Class E*, based on their mobility management requirements. Analytical models are developed to investigate the performance of the existing mobility management protocols for these classes of applications. The analysis shows that applications of a particular class experience different handoff performance when different mobility management protocols are used. Based on this observation, handoff performance comparison of different mobility management protocols are carried out to decide on the suitable mobility management protocol for a particular class of application. Moreover, through analytical modeling the parameters that influence the handoff performance of mobility management protocols are identified. These parameters can be used to design new application adaptive techniques to enhance the handoff performance of the existing mobility management protocols.

6.1.4 A Framework for Adaptive Multi-Layer Mobility Management in Next-Generation Wireless Systems

In Chapter 4, the mobile applications are classified different types of mobile applications into five categories: *Class A* through *Class E*, based on their mobility management requirements. Among different handoff performance parameters, *handoff latency* and *security* are important for all mobile application classes. On the other hand, *end-to-end delay* is important for *Class D* and *Class E* applications that are real time in nature while *transport-layer transparency* is important for *Class B* and *Class C* applications. In parallel, the results of qualitative analysis in 4 shows that mobility management protocols operating from different layers of the classical TCP/IP protocol stack achieve different performance results with respect to different handoff parameters. Therefore, instead of using one mobility management protocol for all

application classes, we propose application adaptive mobility management.

In Chapter 5, an application adaptive mobility management framework, AMMF, is proposed to support seamless handoff support to all application classes. In AMMF, a particular application class uses the mobility management protocol that achieves good performance results for the parameters that are relevant to this application class. However, this application dependent mobility management itself will not be enough to support seamless mobility management. To address this problem, AMMF proposes to share information between different layers. This cross-layering approach eliminates the negative effects of different parameters such as link layer frame error rate and handoff signaling delay on the performance. First, the working principles of AMMF are developed. This is followed by the design of architectural components of AMMF. Then, analytical modeling is developed to investigate the handoff performance of AMMF. Finally, simulation experiments are carried out using the analytical modeling to evaluate the handoff performance of AMMF for different types of applications. The results show that AMMF significantly enhances the handoff performance for different classes of applications.

6.2 Future Research Directions

In future NGWS that integrates the heterogeneous wireless networks will be important to deliver best possible services to the mobile users. There are many challenging research issues related to NGWS.

- **Resource management and user profile based service provisioning in NGWS:** In the NGWS, the ultimate objective is to achieve efficient utilization of the resources of the individual networks to serve the users. In this context, new resource management techniques are required to operate the individual networks at their optimum capacity. User profile based service provisioning

in wireless networks is becoming more and more important. User profile consists of users' locations and personal preferences, in terms of cost (choosing the cheapest access network), applications (with requirement of capacity, delay, and security), or bandwidth (choosing the fastest network). By taking into account the user profile and best available network information, user profile based service provisioning can be supported in NGWS.

- **Integration of ad-hoc and sensor networks in NGWS:** Ad-hoc and sensor networks are the major areas of research and represent the future of wireless communications. These networks are going to play an important role in the success of NGWS. QoS is of great importance in these networks since it can improve performance and allow critical information to flow even under difficult conditions. Variable link conditions and mobility are intrinsic characteristics in ad-hoc and sensor networks. These factors result in frequent rerouting among the mobile nodes; hence, network topology and traffic load conditions change dynamically. Therefore, it is difficult to support appropriate QoS in these networks. Efficient QoS support in ad-hoc and sensor networks can be achieved by using the wireless link condition and mobility prediction. This area needs more exploration to develop efficient QoS provisioning under dynamic network conditions.

- **Cross-layer protocol design:** Existing communication systems are designed by dividing the entire communication process into independent layers. This approach has reached its maximum capacity. To further enhance the performance of existing communication systems, recently, an increased interest in protocols that rely on interactions between different layers of the protocol stack has occurred. This approach, known as cross-layer protocol design, is in its infancy; hence, open research areas exist. Cross-layer approach is studied in this research

in the context of mobility management. The results of the analysis developed in this research can be used to develop efficient cross-layer routing and transport layer protocols.

APPENDIX A

ARCHITECTURE AND INTER-SYSTEM HANDOVER MANAGEMENT PROTOCOLS FOR 2G/3G AND WLAN INTEGRATION

A.1 Introduction

Third generation (3G) wireless systems and WLAN technologies are becoming the integral part of the wireless communications. Currently, both technologies are operating independently within their inherent limitations. For example, 3G with ubiquitous coverage supports maximum data rate of only 2 Mbps at a higher cost and WLAN provides data rate up to 100 Mbps at extremely low cost, but only for low mobility users and has local coverage.

The complementary nature of 3G and WLAN [14] has attracted industry, academia, and standard bodies [1], [6] for their integration. The integrated 3G/WLAN system keeps the best features of both 3G and WLAN, i.e., global coverage of 3G, and high-speed and low-cost of WLAN [71]. At the same time, it eliminates the weaknesses of either system. For example, the low data rate limitation of 3G can be overcome when a WLAN coverage is available, through handover of the user to the WLAN. Similarly, when the user moves out of WLAN coverage area, it can be handed over to the overlaying 3G system. The basic idea is to use small-coverage area high-bandwidth WLAN whenever possible else to use 3G.

In the literature several architectures have been proposed to interconnect 3G and WLAN. These can be broadly classified into *tight coupling* (also known as *emulator approach*), *loose coupling* (also known as *Mobile IP approach*), and *no coupling* (also known as *gateway approach*) [79], [88]. In the *tight coupling* architecture, the WLAN network appears to the 3G system as either a radio access network (RAN)

in case of GPRS [1], [71], [79] or as a Packet Control Function (PCF) in case of cdma2000 [23]. This approach has the advantages of low handover delay and reduced packet loss [1], [79]. However, because both systems are tightly coupled, it is not flexible and also independently operated WLANs cannot be integrated [23], [79]. Moreover, in this architecture all the packets go through the 3G network. Hence, the 3G network becomes a bottleneck [79], and needs to be redesigned to sustain the increased load [23].

In case of *loose coupling* architecture, mobility management in the integrated 3G/WLAN system is handled using Mobile IP (MIP) protocols [1], [23], [71]. This approach has several advantages such as independent data path for WLAN and 3G traffic, and independent deployment and traffic engineering of WLAN and 3G [23]. However it suffers from many shortcomings including triangular routing if route optimization is not performed [79], high handover delay [54], packet loss, high update latency. Multi-tunnel technology in Mobile IP is used in [54] to reduce the handoff delay and packets loss. Loose coupling architecture requires the authentication, billing, and mobility management mechanisms of 3G and WLAN to inter-operate [23]. It also requires that the 3G and WLAN systems have roaming agreement [23].

The *No coupling* architecture treats 3G and WLAN as peer-to-peer networks. In this case, the legacy mobility management schemes are used to handle intra-system roaming, whereas, the inter-system roaming between two networks having roaming agreement, is performed by a gateway. The gateway converts control signals and routes data packets between two networks for roaming users [80]. In [80] a gateway called virtual GPRS support node (VGSN) is used to integrate 3G and WLAN.

All of the above architectures require the existence of bilateral service level agreement (SLA) between the 3G and WLAN operators. However, architectures requiring bilateral SLA between different 3G and WLAN providers are not feasible because of the following reasons. First, operators have reservations to open their network

interfaces to every other operators. Secondly, each time a new operator deploys its WLAN service, it has to be integrated to every other existing operators networks separately. This requires changes to the network infrastructures of all the existing operators. Moreover, schemes requiring bilateral SLA are not scalable [33].

Therefore, a new architecture is required to integrate the 3G systems and WLANs of different providers who may not necessarily have bilateral SLA among them. Once such an architecture is designed the next challenge is to support seamless roaming between the 3G and WLAN networks.

In this research, a novel architecture is proposed to integrate the 3G and WLANs of different providers with or without bilateral SLA among them. We propose the use of a third party, Network Inter-operating Agent (NIA), to integrate these networks. The proposed architecture is scalable, i.e., can incorporate any number of 3G and WLANs of different service providers. We analyze both client assisted and network assisted approaches to provide seamless roaming between 3G and WLAN; and advocate the latter as the preferred choice. Then a novel network assisted seamless roaming algorithm is proposed using the concept of dynamic boundary area. We define steps for both WLAN to 3G and 3G to WLAN inter-system handover (ISHO), and design the associated protocols. In addition, the mathematical formulation of the dynamic boundary area size is derived. Furthermore, performance evaluation of the proposed network assisted ISHO algorithm is carried out.

A.2 The NIA based 3G/WLAN Integrated Architecture

The proposed 3G/WLAN integrated architecture is shown in Fig. 43, consisting of cdma2000[1] networks of two different providers (A and B), their WLANs, and WLAN

[1]cdma2000 is used as the reference 3G network to explain our architecture. The proposed architecture can also integrate other 3G networks such as UMTS. The terms 3G and cdma2000 are used interchangeably in the rest part of this paper.

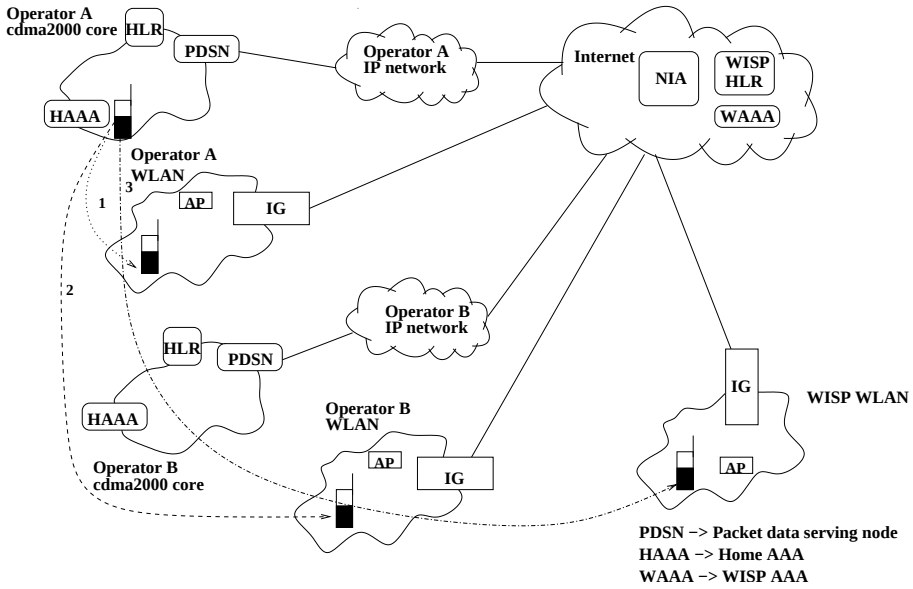


Figure 43: NIA based integrated 3G/WLAN architecture.

deployed by a wireless Internet service provider (WISP). It may be noted that our architecture can integrate any number of cdma2000 networks of different providers and their WLANs; other 3G networks of different operators and their WLANs; and also any number of WLANs of different WISPs. Two new entities **Network Interoperating Agent** (NIA) and **Interworking Gateway** (IG) that are shown in Figure 43 are proposed to integrate the 3G networks and WLANs of different service providers.

Architectures requiring bilateral SLA among different 3G and WLAN providers are not feasible because of the reasons mentioned earlier. Therefore, the use of a third party to integrate the 3G and WLANs of different service providers is proposed in this research. The NIA in the proposed architecture is the third party and it resides in the Internet. A WLAN provider does not have to create separate bilateral SLA with every other 3G operators. Instead it offers roaming service to users of several 3G operators with only one SLA with the NIA. The NIA handles the authentication, billing and mobility management issues of inter-system roaming. Currently, the Authentication, Authorization, and Accounting (AAA) broker networks support authentication and

billing for users belonging to different service providers. But they can not handle the mobility management issues, and hence, can not be used as the third party. NIA has service level agreements (SLAs) with 3G and WLAN operators. The sub-systems of the NIA are shown in Figure 44 (a) and are described below.

- The *authentication unit* is used to authenticate the users moving between 3G and WLAN belonging to two different service providers (refer to Section A.2.1.1).
- The *accounting unit* handles the billing issues between 3G and WLAN as discussed in Section A.2.1.2.
- The *operators database* stores information about the 3G and WLAN operators who have SLAs with the NIA.
- The *handover management unit* decides if the MT's [2] ISHO request should be granted or not. For this, it derives the *Network Access Identifier* (NAI) from the Mobile IP *Registration Request* message and verifies with the *operators database* for the existence of SLA with the home operator of the MT. When applicable it also acts as the mediator between 3G and WLAN, e.g., for transfer of user service profile from the 3G to WLAN. Moreover, it stores the locations of the WLANs of various providers and assists the MTs to learn about the available WLANs in their vicinity.

The **Integration Gateway** (IG) functions as the gateway between the WLAN domain and the Internet. Its sub-systms are as follows (refer to Figure 44 (b)).

- The *mobility management unit* implements the mobile IP [63] (MIP) functionalities using the MIP foreign agent (FA). It also has a *seamless roaming module* which implements the network based mobility management for seamless roaming of users between 3G and WLAN networks as discussed in Section A.3.

[2]We use the terms mobile user and mobile terminal (MT) interchangeably in this paper.

- IG implements traffic monitoring function in its *traffic management unit* by discarding the packets coming from unauthorized users.
- The *authentication unit and accounting unit* provide authentication service and billing support, respectively, to the roaming users (refer to Section A.2.1).

The sub-systems of IG other than the *seamless roaming module*, shares functionalities of IOTA gateway proposed in [23].

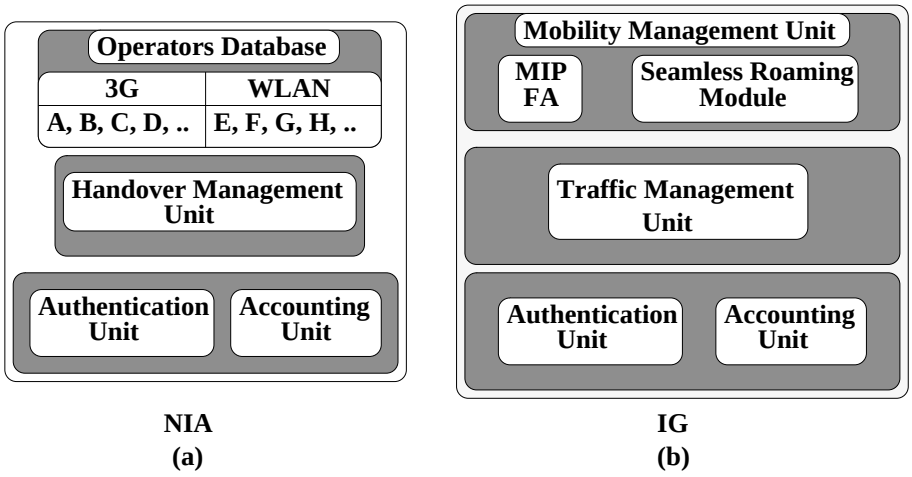


Figure 44: Logical diagram showing the subsystems of NIA and IG.

A.2.1 Security and billing

The proposed 3G/WLAN integrated architecture provides WLAN operators means to verify the legitimacy of the roaming users. It also provides the operators with suitable billing mechanisms.

A.2.1.1 Security

The proposed security architecture for the third party based 3G/WLAN integration is shown in Fig. 45, where the Foreign Network (FN) is a WLAN network and MT's Home Network (HN) is a 3G network. This architecture glues the security architectures of WLAN and 3G through *Authentication Unit* (AU) of NIA (AU_NIA). The use of AU_NIA eliminates the need for any direct security association/agreement between

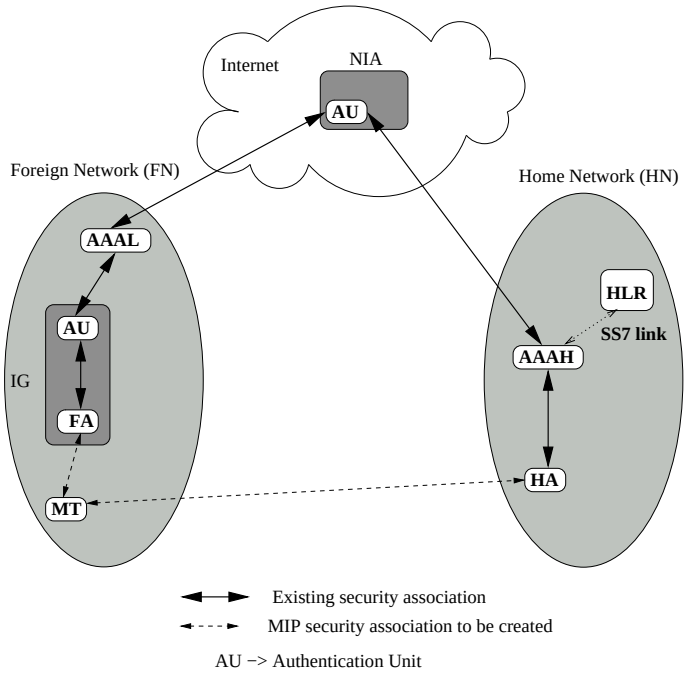


Figure 45: The proposed security architecture for NIA based 3G/WLAN integrated architecture.

WLAN and 3G networks. Both WLAN and 3G networks have separate security association/agreement with AU_NIA. Thus, AU_NIA functions, in essence, as a trusted third party for authentication dialogs between WLAN and 3G, which do not have security agreement with each other. The working principle of this third party based security architecture is as follows. When a mobile user requests service from a foreign WLAN network and the WLAN determines that it has no SLA with user's home 3G provider, it forwards the request to AU_NIA to authenticate the user. Then, AU_NIA talks to user's home 3G provider and mediates between 3G and WLAN for authentication message exchanges. Once the user is authenticated, AU_NIA also creates security associations/keys required between different network entities. Finally the 3G and WLAN networks will be mutually authenticated, and will have session keys for secured data transfer.

The authentication and Mobile IP registration processes are integrated in the proposed architecture using the procedures defined in [33]. The architecture in Fig. 45

shows the existing security associations along with the required MIP security associations so that the Foreign Network (FN) will be able to deliver services to the roaming MT. IEEE 802.1x port access control standard [3] is used for end-to-end mutual authentication between a MT and its home AAA server (AAAH). IEEE 802.1x uses a special frame format known as Extensible Authentication Protocol (EAP) over LAN (EAPOL) for transportation of authentication messages between a MT and an access point (AP). EAP [22] over RADIUS [68] or Diameter [26] is used for the transportation of authentication messages between other entities. When the MT roams into a foreign WLAN domain, the authentication and MIP registration are carried out as described below. The signaling messages for this are shown in Fig. 46. Here, EAP-SIM [37] is used to illustrate the authentication process. Note that any other authentication schemes, e.g. EAP-AKA [15], EAP-SKE [70], EAP-TLS [5] etc. can also be used.

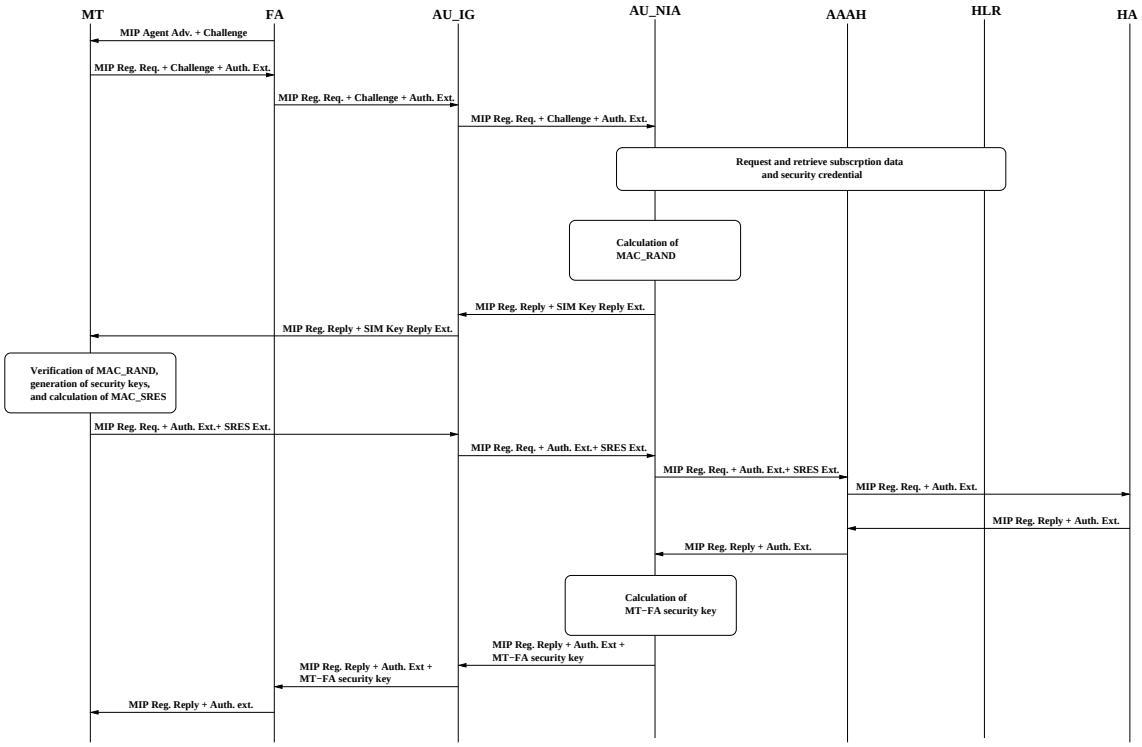


Figure 46: The authentication signaling messages for 3G/WLAN integrated architecture.

1. When the MT hears Mobile IP (MIP) *Agent Advertisement* containing *Mobile IP Challenge/Response* extension [24], it sends MIP *Registration Request* including *Mobile IP Challenge/Response* extension and *Mobile-AAA Authentication* extension (as defined in [24]) to the FA located in IG. The MT also includes a *SIM Key Request* extension [36] and a *Network Access Identifier* (NAI) [25], e.g. *MT@relam*, in its MIP *Registration Request*. The *SIM Key Request* extension contains a random number (NONCE_MT) picked up by the MT, which is used for new authentication key generation as discussed later in this section.

2. When the FA receives the MIP *Registration Request* and finds the *Mobile-AAA Authentication* extension, it learns that the MT is a roaming user and forwards the MIP *Registration Request* to the Authentication Unit of IG (AU_IG). Based on the NAI in the MIP *Registration Request*, the AU_IG recognizes that the WLAN operator does not have direct SLA with the MT's Home Network (HN) and forwards the MIP *Registration Request* to the Authentication Unit of NIA (AU_NIA), either directly or through other AAA proxies.

3. The AU_NIA examines the NAI of the received MIP *Registration Request* message and forwards it to MT's Home AAA server (AAAH). Once, AAAH receives the MIP *Registration Request* containing the *SIM Key Request* extension, first it verifies the *Mobile-AAA authentication* extension. If the authentication is successful, it contacts MT's home 3G network elements over SS7 network and obtains n number of triplets (RAND, SRES, Kc). Then it forwards a copy of these triplets to AU_NIA.

4. When AU_NIA receives n triplets it derives a MT_AAAH key (K_{MT_AAAH}) and calculates message authentication code (MAC) for the RANDs (MAC_RAND) using [36]

$$K_{MT\_AAAH} = h(n * Kc | NONCE\_MT) \ \ and$$

$$MAC\_RAND = PRF(K_{MT\_AAAH},\ \alpha) \tag{99}$$

where α is $n * RAND|$key lifetime; and $h()$ and $PRF()$ denotes a one-way hash function and a keyed pseudo-random function, respectively. Then, AU_NIA sends the RANDs and MAC_RAND to AU_IG, which forwards those to FA. The FA sends a MIP *Registration Reply* message to the MT containing a *SIM Key Reply* extension. The MIP *Registration Reply* reply message also contains the RANDs, MAC_RAND, and the remaining key lifetime. The MT derives the corresponding SRES and Kc values using its SIM card and the received RANDs. It also calculates (K_{MT_AAAH}) and MAC_RAND using (99). It validates the authenticity of RANDs by comparing the calculated MAC_RAND with the received MAC_RAND. Thus, confirming that the RANDs are generated by its HN. If the MAC_RAND is valid, the MT calculates a MAC for its SRES values using [36]

$$MAC\_SRES = PRF(K_{MT\_AAAH}, n * SRES) \tag{100}$$

The MAC_SRES is used by AU_NIA to know if the SRES values are fresh and authentic. The MT also generates security association keys; (K_{MT_FA}) for the FA and (K_{MT_HA}) for the HA using [36]

$$K_{MT\_FA} = PRF(K_{MT\_AAAH},\ Add_{FA})\ \ and$$
$$K_{MT\_HA} = PRF(K_{MT\_AAAH},\ Add_{HA}) \tag{101}$$

where Add_{FA} and Add_{HA} are the IP address of FA and HA, respectively. These keys are used to authenticate subsequent Mobile IP registrations until the key lifetime expires.

5. Now, the MT resends MIP *Registration Request* message to the FA containing *SRES* extension [36] and *Mobile-AAA Authentication* extension. When FA detects the presence of *Mobile-AAA Authentication* extension, it forwards the MIP

Registration Request message to AU_IG, which forwards it to AU_NIA. AU_NIA calculates MAC_SRES and compares that with the received MAC_SRES. If valid, it forwards the MIP *Registration Request* message to the AAAH. After successful authentication AAAH forwards the MIP *Registration Request* containing K_{MT_HA} (calculated using (101)) to the HA. The HA carries out the registration for the MT as defined in [63] and sends MIP *Registration Reply* to AAAH, who forwards it to AU_NIA. AU_NIA calculates MT-FA security key, K_{MT_FA}, and forwards the MIP *Registration Reply* (containing K_{MT_FA} and the Kc keys) to AU_IG. AU_IG forwards it to FA. FA extracts K_{MT_FA}) and the Kc keys and send a MIP *Registration Reply* to the MT. The Kc keys are used for secure data transfer between the MT and FA providing confidentiality and integrity to the data traffic. If necessary a FA-HA security association key can be generated by AU_NIA using (102) and distributed to the FA and HA as a part of authentication process.

$$K_{FA\_HA} = PRF(K_{MT\_AAAH}, Add_{FA}, Add_{HA}) \tag{102}$$

A.2.1.2 **Billing**

Once the MT is authorized by the WLAN, *Accounting Unit* of Integration Gateway (IG) (ACU_IG) maintains a per user accounting record based on the charging policy of the WLAN provider (e.g., connection duration, amount of data transfered etc.). It transfers the accounting information either on per session basis or in real-time to the AAAL server of the WLAN domain. The AAAL server collects and consolidates the accounting information for the MT and forwards it as WLAN access call detail records (WLAN CDRs) to the *Accounting Unit* of NIA (ACU_NIA), which converts it to the CDR format supported by MT's home network and forwards the final CDRs to the AAAH for billing the user.

A.2.2 Hierarchical NIA

In the architecture, the NIA is involved only during the ISHO process and transfers the control signals between 3G and WLAN. Once the ISHO is over, the data traffic of the roaming users do not go through NIA as discussed in Section A.4. Therefore, the load on NIA is limited. We propose hierarchical NIA structure to integrate the 3G and WLAN networks globally. In this hierarchical structure, first the 3G and WLAN networks of various providers are integrated at the regional (e.g. city) level through first tier NIAs. These regional NIAs of a particular country or several countries are then integrated through second tier NIAs, followed by the integration of second tier NIAs through third tier NIAs to realize global 3G and WLAN integration. Exact number of tiers and number of NIAs at each tier depend on several factors, such as number of 3G and WLAN providers in that tier, number of roaming user etc. In this hierarchical NIA structure, a 3G or WLAN operator only need to have SLA with the nearest first tier (also known as regional) NIA operator to be able to provide its subscribers with global WLAN access.

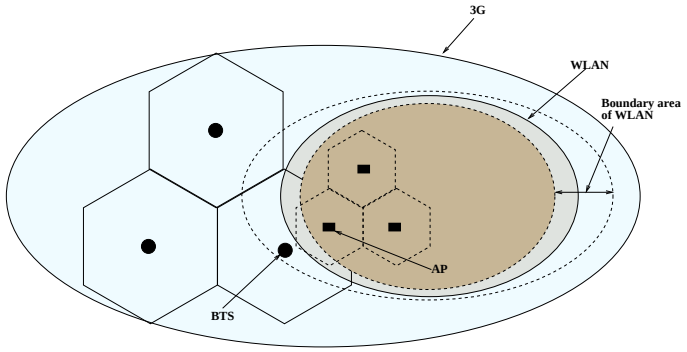


Figure 47: Dynamic boundary area between WLAN and 3G.

A.3 Network Assisted Algorithm for Seamless Roaming from WLAN to 3G

3G coverage overlaps the coverage area of WLANs. This means that there is no possibility of connection loss during a 3G to WLAN ISHO. On the other hand, during

a WLAN to 3G ISHO if the MT moves out of WLAN coverage before the successful completion of ISHO procedures, it will encounter a connection loss. A client assisted algorithm using the received signal strength (RSS) and the priority of the 3G/WLAN interfaces, is proposed in [23] to implement seamless roaming from WLAN to 3G. The mobile client monitors the RSS of WLAN and switch to 3G when it goes below a threshold. A FFT-based technique is proposed in [97] to trigger a handover from WLAN to 3G when the RSS goes below a threshold value. In these approaches the ISHO procedures must be completed before the WLAN RSS goes from the threshold value to RSS_{min}, i.e., the minimum RSS required for successful communication with a WLAN AP. Else the ISHO process will be unsuccessful and MT will loose its connections. This can happen when the MT is near the boundary of a WLAN and drives out the WLAN coverage area very fast before it is handed over to the 3G system (for example when a user is driving away from the office/airport/campus parking lot, while accessing the WLAN over there). Moreover, in these algorithms, the mobile client always monitors the RSS of the WLAN and 3G interfaces to decide about a possible ISHO. This adds significant unnecessary processing especially when the MT stays inside the WLAN for a long time. This extra processing is costly for power constrained devices such as PDAs, 802.11 phones etc. A typical WLAN user stays inside a WLAN for a long time, especially in offices, universities, airports, shopping malls etc. Therefore, it is unnecessary to carry out the extra processing of monitoring the RSS of the available 3G interface unless the MT is anticipated to move out of the serving WLAN.

In this research, a network assisted approach is proposed to carry out seamless roaming between 3G and WLAN that eliminates the shortcomings of the above client assisted approaches. The proposed algorithm is implemented in the *seamless roaming module* of IG. When the *seamless roaming module* learns that the MT starts getting served by a boundary access point (AP), an AP serving a boundary cell of WLAN, it

anticipates that the MT may move out of the WLAN coverage area in the near future. Then it estimates the right time to initiate the ISHO process to ensure a successful handoff from WLAN to 3G. The IG determines the right time for WLAN to 3G ISHO initiation using the concept of dynamic boundary area as shown in Fig. 47. The WLAN to 3G ISHO is initiated when an MT enters the boundary area. The size of the boundary area (L_{BA}) is a function of MT's QoS requirement (q), speed (v) and network state (s) as shown in (103).

$$L_{BA} = f(q, v, s) \tag{103}$$

For simplicity, in this research only MT's speed is considered to estimate the size of the boundary area. It may be noted that this estimate can be easily extended to incorporate the QoS and network state information. L_{BA} is estimated such that ISHO procedures are completed before the MT crosses the WLAN coverage area. Let the time required to complete the ISHO process be τ. During this time an MT with high speed, will travel more distance compared to a slow moving MT. Hence, the ISHO process must start from a farther distance from the boundary of WLAN for a fast moving MT compared to a slow moving one. Therefore, when the speed of the MT is higher the size of boundary area is larger compared to a lower speed case. Detailed procedure to calculate the size of this boundary area is described in Section A.5.1. The boundary area is extended beyond the WLAN coverage, such that it is symmetric around the WLAN coverage area. The boundary area beyond the WLAN coverage area is used to avoid the ping-pong effect as described in Section A.4.

Inter-Access Point Protocol (IAPP) (IEEE Std 802.11f/D5) [2] is used to detect the association of a MT with a boundary AP. When a MT enters the coverage area of a new AP, it initiates a handover to the new AP. Then, the Access Point management entity (APME) sends the *IAPP-ADD.request* message to the IAPP entity of that AP. When the IAPP receives an *IAPP-ADD.request* message, it sends an IAPP *ADD-notify* packet and a *Layer 2 Update* Frame to the IAPP IP multicast address. This

multicast group consists of APs and Layer 2 interworking devices, e.g. bridges and switches of the WLAN domain. The *Integration Gateway* (IG) is a part of this multicast group and hence, it receives the IAPP *ADD-notify* packet and the *Layer 2 Update* Frame. Upon their receipt, in addition to the functions defined in IAPP, the IG also determines if the new AP to which the MT moved, is a boundary AP. For this, the IG maintains a table containing the BSSIDs of the boundary BSSs and the IP address of the corresponding APs. The BSSIDs of the boundary BSSs are available during the WLAN deployment. The mapping between the boundary BSSIDs and the IP addresses of the corresponding APs is done using a RADIUS exchange or locally configured information as defined in [2]. When the IG receives an IAPP *ADD-notify* packet, it checks its Boundary BSSIDs Table. If there is an BSSID in this table with the IP address of the AP received in the IAPP *ADD-notify* packet, then learns that the MT has moved to a boundary AP.

A.4 Inter-system Handover Protocols

In 3G/WLAN integrated system ISHO can be from WLAN to 3G (henceforth referred as WG_ISHO) or from 3G to WLAN (henceforth referred as GW_ISHO). The entire ISHO process is divided into four phases: *Initiation, Preparation, Start, and Completion*. In the *Initiation* phase, the ISHO process is initiated. Once initiated, the *Preparation* phase prepares the MT for a possible ISHO. Resource allocation in the next system, and alternative route set up are carried out in the *Preparation* phase. Finally, the network decides when to begin the handover and executes the *Start* phase, which is followed by the *Completion* phase. The ISHO protocols are described in reference to Fig. 43. The ISHO protocols for GW_ISHO are described first followed by those for WG_ISHO.

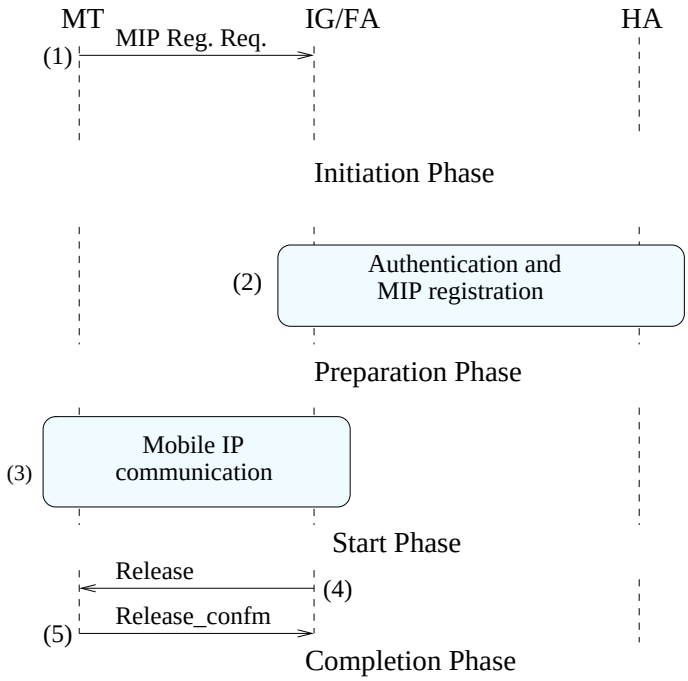


Figure 48: Signaling messages for GW_ISHO.

A.4.1 ISHO protocols for 3G to WLAN handover

When the MT is served by 3G (e.g. cdma2000), its WLAN interface goes to passive scan mode (where the MT spends only little power) to search for an available WLAN coverage. The MT can avoid the use of passive scan mode to save power and learn about the available WLANs in its vicinity using the *handover management unit* of the NIA. When an MT served by 3G detects the presence of a WLAN, it initiates a handover to the WLAN, i.e., GW_ISHO. The GW_ISHO protocols are illustrated in Figure 48. These are explained below.

1. *Initiation*

   When an MT detects the presence of a WLAN, it listens for Mobile IP (MIP) *Agent Advertisement* message or sends an MIP *Agent Solicitation* message [63]. It initiates a GW_ISHO by sending an MIP *Registration Request* message to the FA, located in the IG of the corresponding WLAN domain.

2. *Preparation*

The *Preparation* phase starts once FA receives the *Registration Request*. It carries out the mobile IP registration along with the authentication and authorization operations as discussed in Section A.2.1.

3. *Start*

   The *Start* phase is started after the successful registration of the MT with the WLAN. In this phase the MT maintains simultaneous registrations [63] with 3G and WLAN networks as long as it is in a boundary cell of the WLAN. The MT starts receiving packets from its CNs through both 3G and WLAN. But it sends all its traffic through WLAN to take advantage of the higher data rate of WLAN [83].The corresponding nodes (CNs) communicate with the MT using the MIP procedures [63]. The packets from the CNs are first intercepted by the HA. The HA encapsulates the packets destined for the MT and tunnels those at its care-of-addresses (CoAs). When the MT moves into a non-boundary cell of WLAN, it deregisters from the 3G network. The simultaneous registration during MT's stay in the boundary WLAN cell eliminates the need for a WLAN to 3G handover if the MT moves back to 3G network. Hence, ping-pong effect during ISHO is reduced.

4. *Completion*

   When the IG learns that the MT is no longer in a boundary cell of WLAN, it sends a *release* message (Release) to the MT. The MT acknowledges to this using Release_confm message and deregisters from the 3G network. Then MT's 3G interface goes to the off state.

A.4.2 ISHO protocols for WLAN to 3G handover

Different phases of WLAN to 3G handover (WG_ISHO) are described below using Figure 49.

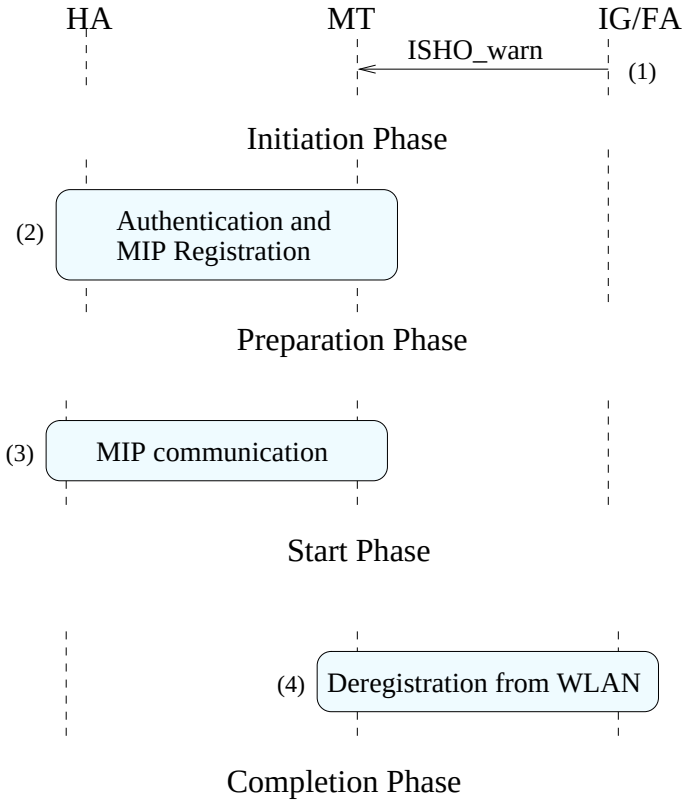


Figure 49: Signaling messages for WG_ISHO.

1. *Initiation*

   When an MT moves into a WLAN boundary cell, the *seamless roaming module* of IG anticipates a possible ISHO of the MT into the overlaying 3G system. It estimates the boundary area length (L_{BA}) for the MT as discussed in Section A.5.1 and starts to monitor the RSS on both 3G and WLAN interfaces. When the WLAN RSS goes below a dynamically selected threshold value (S_{dth}) (as discussed in Section A.5.1), the IG sends an *Inter-system handover warning* (ISHO_warn) message to the MT. Upon the receipt of this message the MT starts the ISHO procedures for its possible handover to 3G while continuing its ongoing connections with the WLAN.

2. *Preparation*

   In the *Preparation* phase, the MT registers with the 3G network using MIP

registration procedures. If the 3G network does not belong to MT's home provider, then 3G roaming protocols are used for this registration. MT also maintains its registration with the WLAN using simultaneous mobility binding to both 3G and WLAN networks.

3. *Start*

   After successful registration with 3G, the MT starts receiving packets from its CNs through both 3G and WLAN. But it sends all its traffic through WLAN as long as it is within the WLAN coverage area. As the MT is registered with the 3G, its ongoing communications can be immediately switched to 3G when it moves out of WLAN. This ensures a seamless ISHO.

4. *Completion*

   Once the MT moves out of the WLAN coverage it uses 3G. The IG keeps MT's registration with WLAN active for a timeout duration equal to $\frac{L_{BA}}{v}$, where L_{BA} is the boundary area length and v is the user speed. In this way IG virtually extends the boundary area beyond WLAN's coverage area. If the MT moves back to the WLAN coverage within this time, there is no need to call for GW_ISHO procedures.

A.5 Performance Evaluation

In this section, first the mathematical formulation for the length of the dynamic boundary area is derived as a function of users' speed. Then, the WG_ISHO failure probability of the proposed dynamic boundary area based ISHO algorithm is compared with that of the fixed RSS based ISHO algorithm that monitors the RSS of both WLAN and 3G interfaces and initiates a WG_ISHO when the difference of RSS of the interfaces goes below a threshold value. Finally, power consumption and the cost associated with false WG_ISHO initiation are investigated for these algorithms.

The boundary area based ISHO algorithm and fixed RSS based algorithm are referred by BA_ISHO and FRSS_ISHO algorithm, respectively.

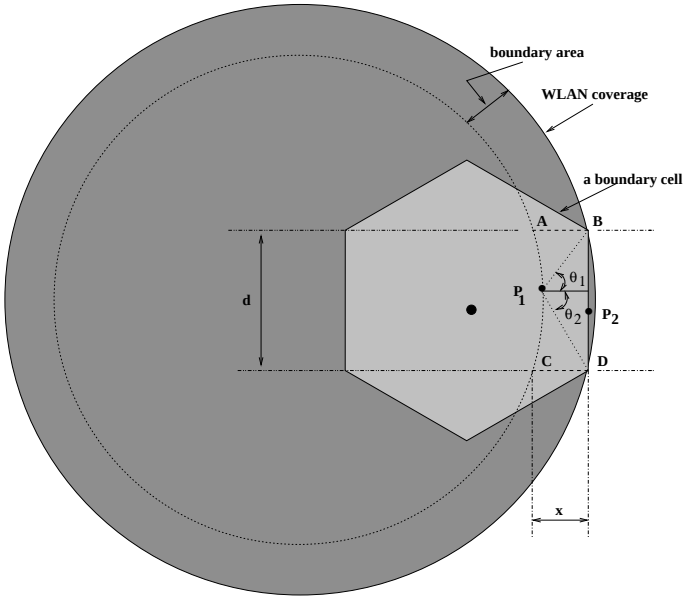


Figure 50: The boundary region of a WLAN network.

A.5.1 Dynamic Boundary Area Length Estimation

While being served by a boundary AP, an MT may enter the boundary area at any point P_1 along the line AC (as shown in Fig. 50) with equal probability. It is assumed that user's speed (v) and direction of motion (θ) are uniformly distributed in [v_{min}, v_{max}] and [$-\pi$, π], respectively. As the WLAN coverage area is usually much larger than the size of a WLAN cell, the shape of the region ABCD is close to a rectangle. It may be noted that this assumption does not introduce any noticeable error to our analytical model. As θ is uniformly distributed the user may move out of the WLAN coverage at any point P_2 along the cell boundary BD (as shown in Fig. 50) with equal probability. Therefore, probability density function (pdf) of the locations of P_1 and P_2 are given, respectively, by

$$f_{P_1}(y_1) = \begin{cases} \frac{1}{d} & \text{for } 0 \leq y_1 \leq d \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad f_{P_2}(y_2) = \begin{cases} \frac{1}{d} & \text{for } 0 \leq y_2 \leq d \\ 0 & \text{otherwise.} \end{cases} \tag{104}$$

where d is the length of WLAN cell as shown in Fig. 50. Since the locations of P_1 and P_2 are independent from each other, their joint pdf is given by,

$$f_{P_1} f_{P_2}(y_1, y_2) = \begin{cases} \frac{1}{d^2} & \text{for } 0 \leq y_1, y_2 \leq d \\ 0 & \text{otherwise.} \end{cases} \tag{105}$$

The distance between two random locations of P_1 and P_2 is denoted by $L = |P_{y_1} - P_{y_2}|$. The probability that $L \leq l$ can be derived using the following integral [21],

$$P(L \leq l) = \iint_{\Omega} f_{P_1} f_{P_2}(y_1,\, y_2)\, dy_2\, dy_1 \tag{106}$$

where Ω is the space of locations of P_1 and P_2 such that $L \leq l$ and $x \leq l \leq \sqrt{x^2 + d^2}$, where x is the length of the boundary area. $P(L \leq l) = 0$ for $l < x$ and $P(L \leq l) = 1$ for $l > \sqrt{x^2 + d^2}$. (106) can be rewritten as

$$\begin{aligned} P(L \leq l) &= \frac{1}{d^2}\Bigg[\int_0^{\sqrt{l^2-x^2}} \int_0^{\sqrt{l^2-x^2}+y_1} + \int_{\sqrt{l^2-x^2}}^{d-\sqrt{l^2-x^2}} \int_{-\sqrt{l^2-x^2}+y_1}^{\sqrt{l^2-x^2}+y_1} + \\ &\quad \int_{d-\sqrt{l^2-x^2}}^{d} \int_{-\sqrt{l^2-x^2}+y_1}^{d} \Bigg] dy_2\, dy_1 \\ &= \frac{2}{d}\sqrt{l^2 - x^2} - \frac{l^2 - x^2}{d^2} \quad \text{for} \quad x \leq l \leq \sqrt{x^2 + d^2} \end{aligned} \tag{107}$$

The pdf of l can be derived by taking the derivative of (107) and is given by

$$f_L(l) = \begin{cases} \frac{2l}{d^2}\left(\frac{d}{\sqrt{l^2-x^2}} - 1\right) & \text{for } x \leq l \leq \sqrt{x^2 + d^2} \\ 0 & \text{otherwise} \end{cases} \tag{108}$$

The amount of time a user will take to travel the distance between the points P_1 and P_2 is $T = \frac{L}{v}$. The pdf of T is given by

$$f_T(t) = v f_L(vt) = \begin{cases} \frac{2v^2 t}{d^2}\left(\frac{d}{\sqrt{v^2t^2-x^2}} - 1\right) & \text{for } \frac{x}{v} \leq t \leq \frac{\sqrt{x^2+d^2}}{v} \\ 0 & \text{otherwise} \end{cases} \tag{109}$$

Using (109), the WG_ISHO failure probability is given by

$$p_f = \begin{cases} 1 & \text{for } \tau > \frac{\sqrt{x^2+d^2}}{v} \\ p_T(t < \tau) = \frac{\sqrt{v^2\tau^2-x^2}}{d}\left(2 - \frac{\sqrt{v^2\tau^2-x^2}}{d}\right) & \text{for } \frac{x}{v} \leq \tau \leq \frac{\sqrt{x^2+d^2}}{v} \\ 0 & \text{for } \tau < \frac{x}{v} \end{cases} \tag{110}$$

where τ is the WG_ISHO signaling delay and $p_T(t < \tau)$ is the probability that $t < \tau$. The equation (110) shows that zero probability of WG_ISHO failure is achieved for $x > v\tau$. Moreover, to guarantee a non-zero WG_ISHO failure ($0 < p_f < 1$) the required value of x can be estimated using,

$$x = [\tau^2 v^2 + d^2(p_f - 2 + 2\sqrt{1 - p_f})]^{\frac{1}{2}} \tag{111}$$

(111) is derived from (110) for a particular value of p_f such as $0 < p_f < 1$.

The value of x that is estimated in (111) is the required size of the boundary area, L_{BA}. The WLAN RSS at the entrance of the boundary area, i.e., the RSS at a distance L_{BA} from the boundary of the WLAN coverage is determined using the path loss model given by [77]

$$RSS(x)\ [dBm] = RSS(x_0)\ [dBm] - 10\beta \log_{10}\left(\frac{x}{x_0}\right) + \epsilon\ [dB] \tag{112}$$

where β is the path loss co-efficient, $RSS(x)$ and $RSS(x_0)$ are the RSS at distance of x and a reference distance (x_0), respectively, from an AP. $\epsilon\ [dB]$ is a zero-mean Gaussian random variable with standard deviation σ (typical value of σ is 6 to 8 dB) that represents the statistical variation in $RSS(x)$ caused by shadowing . Using (112) the RSS at the entrance of boundary area that we refer as dynamic RSS threshold (S_{dth}) is given by

$$S_{dth}\ [dBm] = RSS_{min}\ [dBm] + 10\beta \log_{10}\left(\frac{d}{d - L_{BA}}\right) + \epsilon\ [dB] \tag{113}$$

where RSS_{min} is the minimum RSS required for the MT to communicate with an AP, i.e., the RSS at the boundary of a WLAN cell. In BA_ISHO the WG_ISHO is initiated when WLAN RSS goes below S_{dth}. We use 914 MHz Lucent WaveLAN DSSS radio interface for which RSS_{min} (i.e., RXThresh) is -64 dBm and $\beta = 4$ for our simulation. Figure 51 shows that the value of S_{dth} increases as the speed increases for a particular value of τ. This is because ISHO must be started earlier for the fast moving users. Moreover, for a particular speed value S_{dth} is increases as τ increases

as shown in Fig. 51 as the WG_ISHO procedures must be initiated earlier for higher values of τ.

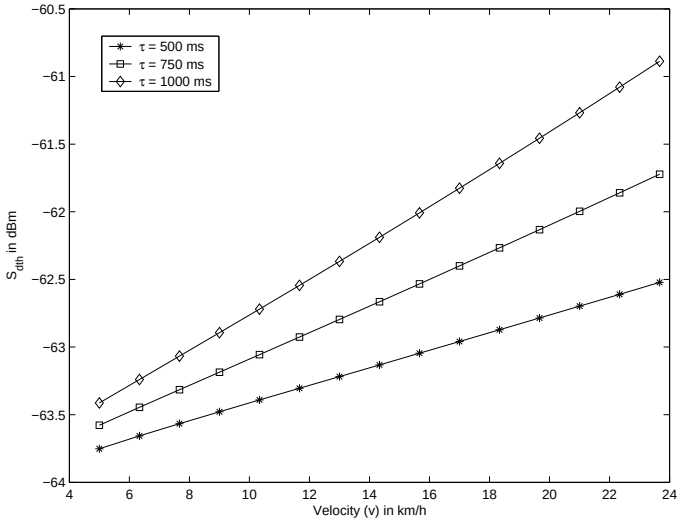


Figure 51: The value of S_{dth} vs. speed for different value of WG_ISHO signaling delay.

A.5.2 WG_ISHO Failure Probability

To analyze the WG_ISHO failure probability, we assume that the target WG_ISHO failure probability is $p_f = 0.02$. The WG_ISHO failure probability for the BA_ISHO is given by (110) for different values of speed. FRSS_ISHO uses a fixed value of RSS threshold (RSS_f). Therefore, the WG_ISHO is initiated effectively at a distance L_f from the boundary of the WLAN coverage where L_f is given by

$$L_f = d\Big[1 - 10^{-\left(\frac{\Delta RSS_f + \epsilon}{10\beta}\right)}\Big] \tag{114}$$

where $\Delta RSS_f = RSS_f - RSS_{min}$. Therefore, the WG_ISHO failure probability for FRSS_ISHO algorithm can be calculated using $x = L_f$ in (110). The WG_ISHO failure probability for BA_ISHO and FRSS_ISHO algorithms is shown in Fig. 52 for τ = 0.5 sec. The results show that for FRSS_ISHO algorithm p_f depends on the speed and the value of the RSS_f used. Therefore, the target p_f is achieved only for certain speed values. On the other hand, for BA_ISHO algorithm, p_f is always limited to the

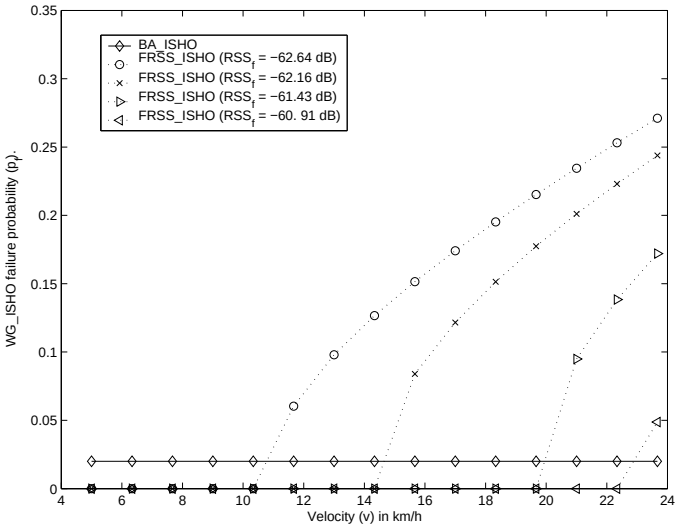


Figure 52: WG_ISHO failure probability of BA_ISHO algorithm vs FRSS_ISHO algorithm.

target p_f of 0.02 and is independent of speed. Figure 52 shows that for FRSS_ISHO algorithm a higher value of RSS_f reduces p_f. However, in this case the false handoff initiation probability (p_a) increases as discussed in next.

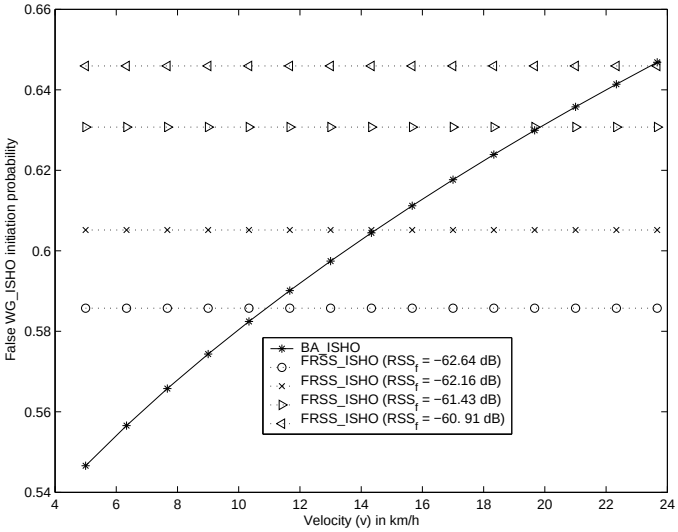


Figure 53: WG_ISHO false initiation probability.

A.5.3 False WG_ISHO Initiation Probability

In BA_ISHO and FRSS_ISHO algorithms the WG_ISHO is initiated when the RSS goes below a certain threshold value (either fixed threshold or dynamically selected threshold). This implies that the WG_ISHO is initiated from a particular distance from the boundary of the WLAN coverage. The distance is dynamically chosen for BA_ISHO algorithm and is a fixed value for FRSS_ISHO algorithm. From Fig. 50, it is clear that when started from a distance x from the boundary of the WLAN the need for WG_ISHO arises only if the MT's direction of motion from P_1 is in the range $[-\theta_2, \theta_1]$. Otherwise, the WG_ISHO initiation is a false one. Therefore, the probability of false WG_ISHO (p_a) is given by

$$p_a = 1 - \frac{1}{d}\int_0^d \left(\frac{\theta_1 + \theta_2}{2\pi}\right)dy = 1 - \frac{1}{\pi}\arctan\left(\frac{d}{x}\right) + \frac{x}{2\pi d}\ln\left(1 + \frac{d^2}{x^2}\right) \qquad (115)$$

The value of p_a can be calculated for BA_ISHO and FRSS_ISHO algorithms by using $x = L_{BA}$ and $x = L_f$, respectively. Figure 53 shows that for FRSS_ISHO p_a depends on the value of RSS_f. p_a is higher for larger value of RSS_f. Therefore, it is not a good idea to use a unnecessarily large value of RSS_f in a hope to reduce p_f for higher speed values as this will increase the value of p_a for lower speed. This is because for higher RSS_f threshold, the ISHO is initiated too early even when the speed of the user is low. This leads to the wastage of limited wireless network resources. Moreover, this increases the load on the network that arise because of the handoff initiation. On the other hand BA_ISHO algorithm initiates the WG_ISHO in such a way that just enough time is there for successful execution of ISHO procedures for a particular speed. Therefore, the ISHO is neither started too early nor too late. The former limits the high cost associated with unnecessarily large value of false handoff initiation for low speed value. The later ensures that ISHO procedures are smooth even for high speed. Thus, the BA_ISHO algorithm optimizes the WG_ISHO false initiation probability through the dynamic selection of S_{dth} as shown in Fig. 53.

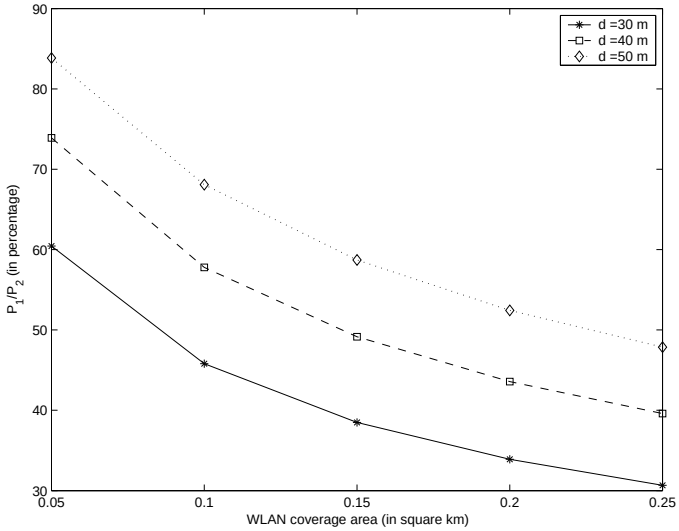


Figure 54: Comparison of power consumption for RSS monitoring on the 3G interface.

A.5.4 WG_ISHO Power Consumption

In case of the existing FRSS_ISHO algorithms while inside a WLAN an MT always monitors the RSS on the 3G interface to decide about a possible WG_ISHO initiation. On the other hand in BA_ISHO the terminal monitors the RSS on the 3G interface only after moved into a boundary WLAN cell. Therefore, if we assume that the WLAN has a coverage area of A_w and hexagonal cell size of d, then ratio of power consumption because of the RSS monitoring on the 3G interface for BA_ISHO (P_1) and FRSS_ISHO (P_2) is given by

$$\begin{aligned} \frac{P_1}{P_2} &= \frac{(\text{number of boundary WLAN cells})\ t_r P}{(\text{total number of WLAN cells})\ t_r P} \\ &= \left(\frac{A_c}{A_w}\right)\left[-3+3\sqrt{1+\frac{4}{3}\left(\frac{A_w}{A_c}-1\right)}\right] \end{aligned} \tag{116}$$

where A_c is the area of each WLAN cell, t_r is the mean WLAN residence time, and P is the power required to monitor the RSS on 3G interface. Figure 54 shows that BA_ISHO achieves significant power saving. The amount of power saving is more for larger WLAN coverage areas.

A.6 Summary

In this research, a novel 3G/WLAN integrated architecture is proposed using the third party, Network Inter-operating Agent (NIA), to integrate 3G and WLANs of different providers. The proposed architecture does not require the existence of direct SLAs among the network providers. Therefore it is scalable. A novel algorithm using the concept of dynamic boundary area is proposed to support seamless ISHO between the 3G and WLAN. In addition, signaling protocols were designed for WG_ISHO and GW_ISHO. The boundary area based 3G/WLAN ISHO algorithm selects a dynamic RSS threshold to initiate the WG_ISHO in such a way that the ISHO procedures are completed before the MT moves out of the WLAN coverage area. Thereby always ensures a successful handoff from WLAN to 3G. Moreover, it optimizes the cost associated with the false handoff initiation. In addition as it does not require the MT to monitor the RSS of the 3G system interface when the MT is served by a WLAN unless the need for WLAN to 3G ISHO arises. Thus it reduces the power consumption associated with the monitoring of 3G interface significantly.

APPENDIX B

VEPSD: AN ACCURATE VELOCITY ESTIMATION ALGORITHM

B.1 Introduction

In current and next generation wireless systems (NGWS), the estimation of users' velocity[1] is important to improve the network performance. In hierarchical cellular systems, the velocity information can be used to assign slow moving users to micro/pico cells and fast moving users to macro cells to reduce the handoff rate for the fast moving users. This increases the system capacity and reduces the number of dropped calls [89]. Moreover, velocity information can be used to ensure successful handoff in a cellular system. For example, when the position and the velocity of a mobile user (MU) is known, MU's arrival time in the next cell can be estimated and accordingly resources can be reserved in advance to ensure a successful handoff.

Several techniques are proposed in the literature for velocity estimation. The algorithm proposed in [89] using the normalized auto-correlation values of the received signal is efficient in classifying the velocity into slow, medium, or fast. However, a better resolution of the velocity is not achievable. In [16], the level crossing rate (LCR) based velocity estimator is proposed. However, in the presence of additive white Gaussian noise (AWGN), this estimator suffers from severe estimation error when the velocity is low. Wavelets are used in [60] for velocity estimation. Switching rate of diversity branches is used for the velocity estimation in [47], but it is shown in [30] that this method is sensitive to the fading scenarios. Hence, it is not practical. In [85], velocity estimation algorithms are proposed that use pattern recognition. However,

[1]Velocity is a vector with both magnitude and direction. However, we refer to the magnitude as the velocity throughout this paper.

these algorithms are computationally intensive [89]. In [59], a velocity estimator based on the statistical analysis of the channel power variations is proposed. In [40], the squared deviations of the received signal envelope is used for velocity estimation. Adaptive array antennas are used for the velocity estimator proposed in [28]. The first moment of the instantaneous frequency of the received signal is used in [17] for the velocity estimation. However, this study is limited only to the Rayleigh fading channels.

A coarse estimation that classifies velocity to slow, medium, or fast is sufficient when the velocity information is used to assign an MU to a macro, micro, or pico cell. On the other hand, accurate velocity estimation is required for seamless mobility support [86] in NGWS. Hence, the desired accuracy of velocity estimation depends on the application. Moreover, the accuracy of velocity estimation should be independent of the fading types (e.g., Rayleigh and Rician fading). Finally, the algorithm should not be computationally extensive. To our knowledge none of the existing velocity estimation techniques mentioned above satisfy all these requirements simultaneously.

In this research, we propose a novel algorithm called VEPSD for velocity estimation. In VEPSD, we first estimate the maximum Doppler spreading frequency (f_m). Then, we use f_m for velocity estimation. VEPSD satisfies the above mentioned requirements of an efficient velocity estimation algorithm.

B.2 VEPSD

The maximum Doppler frequency (f_m) is related to the velocity (v) of a mobile user, speed of light in free space (c), and the carrier frequency (f_c) through

$$v = \left(\frac{c}{f_c}\right) f_m. \tag{117}$$

In case of a narrow band multi-path fading channel, the received low pass signal is given by

$$r(t) = \sum_{n=1}^{N} \alpha_n e^{j\beta_n} e^{(j2\pi f_m \cos\theta_n)t} + w(t) \tag{118}$$

where α_n is the amplitude of the nth arriving wave, β_ns are uniformly distributed over $[-\pi, \pi]$, θ_n is the angle of arrival of the nth arriving wave, and $w(t)$ is AWGN. For large N, the envelope of $r(t)$, i.e., $|r(t)|$ is Rayleigh distributed if no line of sight (LOS) component is present, else it is Rician distributed.

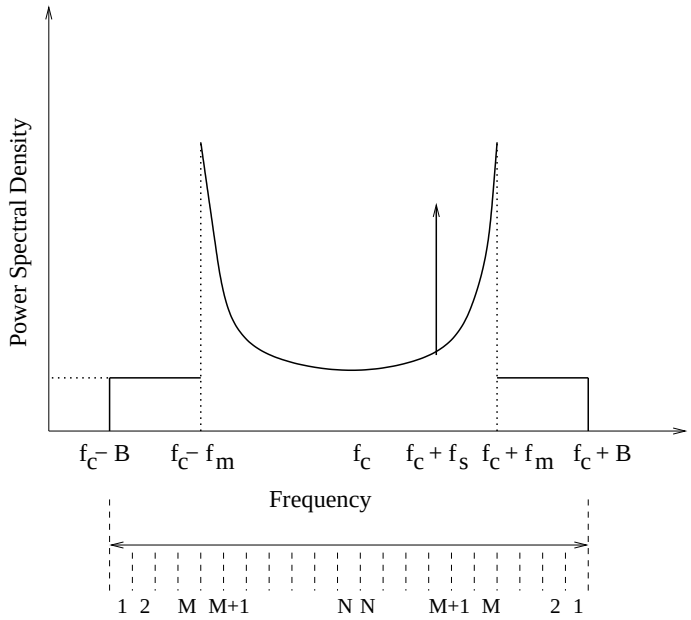


Figure 55: PSD of the received signal envelope in a Rician fading channel in the presence of AWGN.

In a Rician fading environment, the power spectral density (PSD) of $|r(t)|$, $S(f)$,

is expressed as [16]

$$S(f) = \begin{cases} \frac{\Omega}{4(K+1)\pi f_m \sqrt{1-(\frac{f-f_c}{f_m})^2}} + N_0, \\ \qquad |f - f_c| \leq f_m, f - f_c \neq f_s \\ \frac{\Omega}{4(K+1)\pi f_m \sqrt{1-(\frac{f-f_c}{f_m})^2}} + \frac{K\Omega}{4(K+1)} + N_0, \\ \qquad f = f_c + f_s \\ N_0, \\ \qquad f_m < |f - f_c| < B. \end{cases} \tag{119}$$

where Ω is the total received envelope power, K is the Rice factor, N_0 is the PSD of AWGN, and $2B$ is the receiver bandwidth. $f_s = f_c + f_m \cos\theta_0$ is the frequency of the LOS component where θ_0 is the angle of arrival of the LOS component. The plot of (119) is shown in Fig. 55. For a Rayleigh fading channel, i.e., a channel with no LOS component, $S(f)$ can be derived from (119) when $K = 0$ and its plot is similar to Fig. 55, except that there is no LOS component. We differentiate (119) to get the slope of $S(f)$ in a Rician fading environment, which can be expressed as

$$\frac{dS(f)}{df} = \begin{cases} \frac{\Omega(f-f_c)}{4(K+1)\pi f_m^3 [1-(\frac{f-f_c}{f_m})^2]^{\frac{3}{2}}}, \\ \qquad |f - f_c| \leq f_m, f - f_c \neq f_s \\ \frac{\Omega(f-f_c)}{4(K+1)\pi f_m^3 [1-(\frac{f-f_c}{f_m})^2]^{\frac{3}{2}}} + \frac{K\Omega}{4(K+1)}\delta(f_c + f_s), \\ \qquad f = f_c + f_s \\ 0, \\ \qquad f_m < |f - f_c| < f_B. \end{cases} \tag{120}$$

In (120), the slope has three maxima: at $f = f_c + f_m$, $f = f_c - f_m$, and $f = f_c + f_s$. The maximum at $f = f_c - f_m$ and $f = f_c + f_m$ are due to the maximum Doppler frequency. The maximum at $f = f_c + f_s$ is because of the LOS component. When the angle of LOS component $\theta_0 = 0$, the maximum of (120) due to f_m and f_s coincide with each other. Note that when no LOS component is present (Rayleigh fading

channel) (120) has only two maxima at $f = f_c + f_m$ and $f = f_c - f_m$. Therefore, for both Rayleigh and Rician fading, the slope of PSD of the received signal envelope has maximum values at frequencies $f_c \pm f_m$. The frequency component, $f = f_c + f_m$, is always greater than or equal to $f = f_c + f_s$ and greater than $f = f_c - f_m$. We detect the maximum value of (120) which corresponds to the highest frequency component $(f_c + f_m)$ to estimate f_m.

For practical implementation, we use discrete slope calculation. We divide the entire receiver bandwidth $(2B)$ into $2N$ equally spaced intervals as shown in Fig. 55. Each interval has a mirror image about f_c. The discrete frequency value associated with the i^{th} interval is $f_c + B - (\frac{B}{N})i$. We calculate the slope as

$$S(k) = \frac{\sum_{i=1}^{k+1} P(i) - \sum_{i=1}^{k} P(i)}{2\Delta B} = \frac{P(k+1)}{2\Delta B} \tag{121}$$

where $P(i)$ is the sum of the power of i^{th} interval and its mirror image interval. $\Delta B = \frac{B}{N}$ is the width of one interval. Using (121) and Fig. 55, it is clear that, $S(i), i = 1, 2, \ldots, (M-2)$, have the same value and equal to PSD of AWGN, N_0. In a real scenario, N_0 is not flat. Hence, $S(i), i = 1, 2, \ldots, (M-2)$, are not exactly equal to each other. Their values are close to N_0 and different from each other. When noise PSD (N_0) is insignificant compared to the power in the interval containing frequency $f = f_c + f_m$, $S(M-1)$ will be dominant among all the slopes in a Rayleigh fading scenario. On the other hand, in a Rician fading scenario the slopes corresponding to the intervals containing $f_c + f_m$ and $f_c + f_s$ are both dominant. But the slope corresponding to the interval containing $f_c + f_m$ is of lower order compared to the one corresponding to the interval containing $f_c + f_s$, where the order of the slope is given by k in (121). When both $f_c + f_m$ and $f_c + f_s$ belong to the same interval, there is only one dominant slope in case of Rician fading scenario. We detect the lowest order dominant slope of the received signal envelope's PSD to estimate f_m. This ensures that our algorithm is independent of the fading environment.

The estimation of the lowest order dominant slope can be carried out in two ways:

(1) one way is to calculate all the slopes and then detect the peak slope of lowest order and (2) the other way is to calculate $S(1)$, then $S(2)$, and so on until the first dominant slope is detected. In this approach, initially the values of the slopes ($S(1), S(2)$ etc.) are close to N_0 up to the slope corresponding to the interval containing $f_c + f_m$. The slope corresponding to interval containing $f_c + f_m$ is significantly higher than N_0. This is the lowest order dominant slope. There is no need to calculate the other slopes.

For the second approach, there is no need to calculate all the slopes and no sorting is required. Therefore, it has less computational complexity. However, it requires the knowledge about N_0. This requirement can be eliminated if the worst case SNR for a mobile system is known. If the value of N_0 corresponding to worst case SNR is $N_{0(worst)}$, then in the second approach initially the values of slopes corresponding to the intervals before the interval containing $f_c + f_m$ are less than or equal to $N_{0(worst)}$. And the slope corresponding to the interval containing $f_c + f_m$ is significantly higher than $N_{0(worst)}$. Therefore, with the knowledge of $N_{0(worst)}$ in the second approach, when a particular slope is significantly greater than $N_{0(worst)}$, we consider that as the lowest order peak slope. We refer to $N_{0(worst)}$ as slope threshold, S_{th}, in the rest part of the paper. We use the second approach because of its low computational complexity. If the lowest order peak slope corresponds to $k = k_{min}$, then

$$f_m = B - k_{min}(\Delta B) \tag{122}$$

To further reduce the computational complexity, we use a two-step approach to estimate f_m.

- First, we carry out a coarse estimation of $f_m = f_m^1$ using interval width of ΔB_{coarse} for slope calculation in (121). If we denote the index of slope corresponding to lowest order peak as k_{coarse}, f_m^1 is expressed as

$$f_m^1 = B - k_{coarse}\Delta B_{coarse} \tag{123}$$

- Then, we carry out a finer estimate of $f_m = \hat{f_m}$ using interval of ΔB for slope calculation in (121). In this step, we calculate the slope of the received signal envelope's PSD in the frequency range over $f_m^1 - x$ to $f_m^1 + x$. Our choice of $2x$ Hz over which the slope is calculated is arbitrary. Any value for x can be used as long as $2x$ is greater than $2\Delta B_{coarse}$ (which is the granularity of the previous step). If we denote the index of the peak slope, which has the lowest order as k_{finer}, then $\hat{f_m}$ is given by

$$\hat{f_m} = f_m^1 + x - k_{finer}\Delta B. \tag{124}$$

Finally, we estimate the velocity using $f_m = \hat{f_m}$ in (117). (124) shows that, in VEPSD the maximum error in velocity estimation (Δv) is equal to the velocity corresponding to the Doppler spread of ΔB, i.e., $\Delta v = \Delta B \, \frac{c}{f_c}$. Thus, the error in estimation reduces as carrier frequency increases. Hence, our algorithm provides better estimation accuracy for next generation of wireless systems that are expected to operate at higher carrier frequencies around 5 GHz. Another advantage of VEPSD is its scalability to estimate the velocity up to the desired level of accuracy through proper selection of the number of intervals (N) for slope calculation. For example, to determine if the velocity of the mobile is slow, medium or fast, we just need three intervals. On the other hand, using more number of intervals an accurate estimation of the velocity can be achieved.

So far we have discussed VEPSD for narrow band wireless communication systems. In case of CDMA systems, the mobile channel can be represented by the impulse response model [65]

$$h(\tau; t) = \sum_{k=1}^{l} h_k(t)\delta(\tau - kT_c), \tag{125}$$

where l is the number of resolvable paths, T_c is the chip interval, and $h_k(t)$ is the complex channel gain of the kth multipath. $h_k(t)$ has the form in (118) and $|h_k(t)|$ is Rayleigh distributed when no LOS component is present, else it is Rician distributed.

The RAKE receiver can resolve each of the paths in (125) [81]. Then, the channel gain for the kth path, $h_k(t)$, can be obtained by the help of pilot channel or other means [53]. VEPSD uses this $|h_k(t)|$ that contains the Doppler spreading information for velocity estimation. Hence, VEPSD works for CDMA systems as well.

In case of multi-carrier systems the channel across each sub-carrier is a narrow band channel. Hence, the received signal envelope across any sub-carrier can be used in VEPSD for velocity estimation.

Up until now, we discussed the algorithm for isotropic scattering environments. Non-isotropic scattering is usually modeled using von Mises/Tikhonov distribution [78]. Also, in this case the PSD of the received envelope has maximum values at frequencies $f_c \pm f_m$. Hence, f_m can be detected using our VEPSD algorithm. This ensures that the VEPSD algorithm is applicable to both isotropic and non-isotropic scattering environments.

B.3 Performance Evaluation

We carried out the performance evaluation of the VEPSD algorithm through simulation using the fading model proposed by Jakes [77]. We selected the receiver bandwidth B such that it is just greater than the maximum Doppler spread for the highest vehicular velocity to minimize the effect of noise on estimation [16]. We consider B = 325 Hz, which allows velocity up to 175 Km/h at f_c = 2 GHz. We use f_c = 2 GHz, as this frequency band is widely used for cellular systems. We consider ΔB = 5Hz that can estimate velocity to an accuracy of 2.7 km/h. We use ΔB_{coarse} = 27 Hz. To determine the value of slope threshold, S_{th}, we assume the worst case SNR to be 15 dB. This assumption is realistic as the typical SNR in cellular systems is in the order of 20 dB [40]. The threshold value is determined in such a way that, $S_{th} >> N_0$. This ensures that slight variation in N_0 is not identified as a dominant slope. Through simulations, the value of S_{th} for coarse estimation (123)

is found to be 20 and that for finer estimation (124) is 4.5. The estimation interval (T_{est}) corresponds to the time interval over which the received signal envelope samples are collected for the velocity estimation. Hence, $T_{est} = N * \tau$, where N is the number of samples and τ is the sampling period. We use $T_{est} = 1$ s and $\tau = 1$ ms, because these values give a good estimate of the velocity.

B.3.1 Simulation Results for a Rayleigh Fading Channel

We start with the investigation of the effect of AWGN on the estimation error in a Rayleigh fading channel. Then, we investigate the estimation accuracy for various ranges of velocity and analyze the response of the algorithm to changes in the velocity.

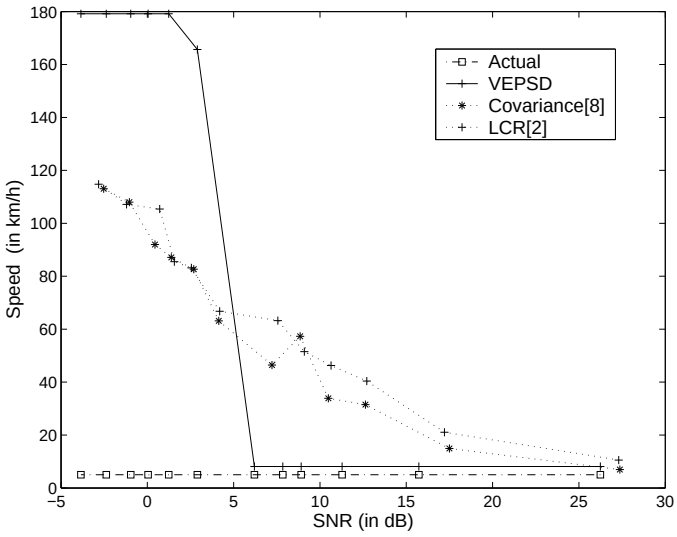


Figure 56: Estimated velocity vs. SNR in a Rayleigh fading channel for $v = 5$ km/h with $\tau = 1$ ms and $T_{est} = 1$ s.

B.3.1.1 Effect of AWGN on Accuracy of Velocity Estimation

Figure 56 shows the performance of the velocity estimation algorithms versus SNR for velocity of 5 km/h. For VEPSD estimator, the performance is degraded when the SNR is below 7 dB in Figure 56. This is because below these values, the relation $S_{th} >> N_0$ does not hold. Hence, the clear existence of the dominant value of the slope corresponding to frequency f_m is lost. From Figures 56 it is clear that at very

low SNR, the VEPSD algorithm always estimates the velocity to be 179 km/h. This is because for very low SNR, $N_0 > S_{th}$. Therefore, the VEPSD algorithm detects interval (1) as the interval corresponding to peak slope, both in coarse and fine estimation steps. Now, using (123) and (124), f_m is estimated as 331 Hz and the corresponding velocity is 179 km/h. Figure 56 shows that the error in velocity estimation increases as the SNR decreases for both covariance and LCR based methods. Also, estimation error is severe for lower velocity compared to that for higher velocity for both LCR [16] and co-variance [40]. Interestingly, the estimation error for VEPSD is independent of SNR when SNR is more than 10dB for both low and high velocity values.

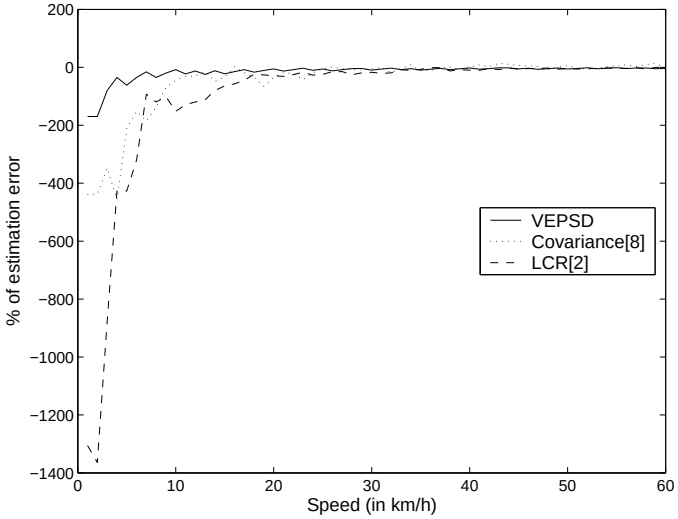


Figure 57: Estimation accuracy in a Rayleigh fading channel for $\tau = 1$ ms, $T_{est} = 1$ s, and SNR=20 dB.

B.3.1.2 Velocity Estimation Accuracy

We carried out the simulation study for all the three estimation algorithms over the velocity range of $1 - 60$ km/h. Figure 57 shows that the proposed VEPSD algorithm can estimate the velocity to a very good accuracy over the entire range (1-60 km/h). This is in contrast to the LCR and the co-variance based estimators, where the error introduced by AWGN is severe for low velocity values.

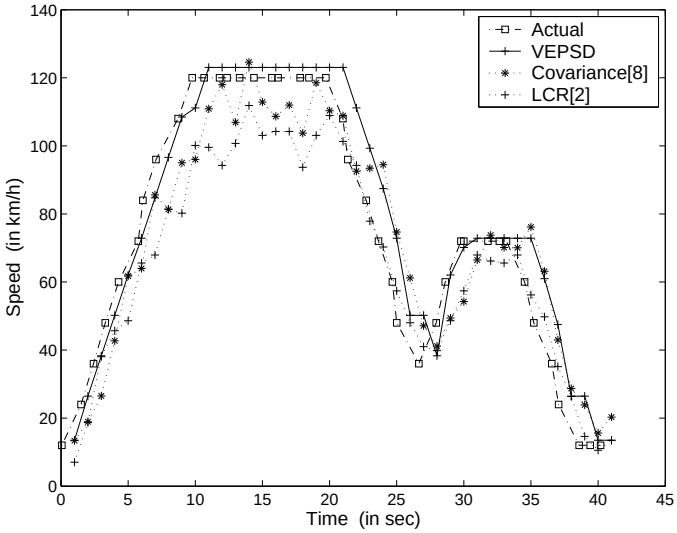


Figure 58: Velocity tracking in a Rayleigh fading channel for $\tau = 1$ ms, $T_{est} = 1$ s, and SNR=20 dB.

B.3.1.3 Velocity tracking

Figure 58 shows that the tracking performance of all the three algorithms are comparable when the mobile is either accelerating or decelerating. When the mobile stays at a constant velocity, VEPSD has better accuracy of estimation than those of LCR [16] and co-variance [40] based estimators. This is because of the randomness of the received envelope, which varies the LCR count and also the variance from time to time. The VEPSD algorithm performs better in this case because even for the randomly varying received envelope, the maximum Doppler frequency used by the VEPSD remains constant during each observation interval.

B.3.2 Simulation Results for a Rician Fading Channel

Figures 59 (a) and 59 (b), show that for VEPSD algorithm the velocity estimation accuracy is independent of the angle of arrival of the LOS component (θ_0) and Rice factor (K). This is in contrast to the co-variance based algorithm, where the accuracy of velocity estimation depend on K and θ_0 as shown in Fig. 59 (b). The LCR based estimator is robust to Rice factor (K), when the level is chosen as the root mean

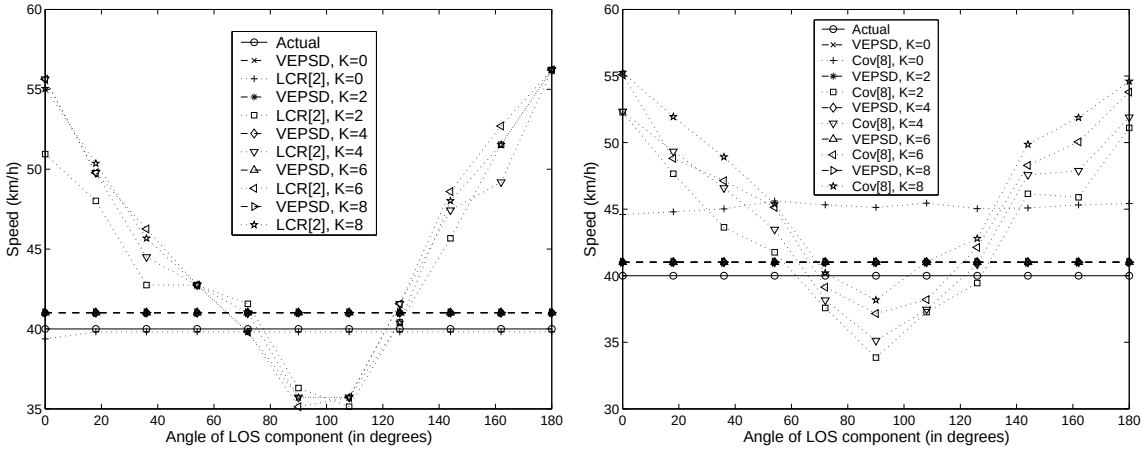


Figure 59: Comparison of VEPSD estimator for $v = 40$ km/h with $\tau = 1$ ms, $T_{est} = 1$ s, and SNR=20 dB: (a) with LCR based estimator and (b) with covariance based estimator.

square (rms) value of the received envelope samples. This also is clear from Fig. 59(a), where the velocity estimation based on LCR depends only on the angle of arrival of the LOS component (θ_0) and is independent of the Rice factor. The robustness of VEPSD algorithm to K can be explained as follows. As K increases, the power of the LOS component increases and that of the scattered components decreases. But still the nature of the PSD plot and hence its slope remains unchanged. Just that the value of slope decreases. For SNR value more than 15dB, this value of slope is much more than S_{th}. So the VEPSD algorithm still detects the peak corresponding to f_m. The value of θ_0 determines the position of the LOS frequency component ($f_c \pm f_s$) with respect to $f_c + f_m$. But our VEPSD algorithm always discards the peak value of slope at $f_c \pm f_s$, as discussed in Sec. B.2. Hence, it is insensitive to θ_0.

B.4 Summary

In this paper, we presented VEPSD, a novel velocity estimation algorithm. We carried out a detailed performance analysis of the VEPSD algorithm and also compared it with two other existing algorithms: LCR [16] and covariance [40] based velocity

estimation. The results show that VEPSD algorithm works very well in both Rayleigh and Rician fading environments. The VEPSD algorithm is robust to both Rice factor and angle of arrival of the LOS component. This is the key advantage of VEPSD compared to LCR and co-variance based velocity estimators. We investigated the effect of AWGN on the accuracy of velocity estimation. The results show that VEPSD algorithm works significantly better in the SNR range typical of cellular systems. In addition, the tracking performance of the VEPSD estimator is comparable to other estimators. VEPSD algorithm works very well for wide range of velocities and is well suited for next generation of wireless systems operating at higher frequencies. VEPSD algorithm can be used to estimate velocity up to the desired level of accuracy. Hence, it is scalable.

REFERENCES

[1] "3GPP System to WLAN Interworking: Functional and Architectural Definition." *Tech. rep. 3GPP TR 23.934 v0.3.0.* 3GPP.

[2] "Draft 4 Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11 Operation." *IEEE Draft 802.11f/D5.* IEEE 802.11f.

[3] "IEEE Standard for Local and metropolitan area networks - Port-Based Network Access Control." IEEE Std 802.1X-2001.

[4] "Inter-PLMN backbone guidelines." GSM association classifications, version 3.4.0, March 2003.

[5] Aboba, B. and Simon, D., "PPP EAP TLS Authentication Protocol," *RFC 2716, IETF*, 1999.

[6] Ahmavaara, K., Haverinen, H., and Pichna, R., "Interworking Architecture Between 3GPP and WLAN Systems," *IEEE Communications Magazine*, vol. 41, no. 11, 2003.

[7] Akan, O. B. and Akyildiz, I. F., "ATL: An Adaptive Transport Layer for Next Generation Wireless Internet," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 22, no. 5, pp. 802–817, 2004.

[8] Akyildiz, I. F., McNair, J., Ho, J. S. M., Uzunalioglu, H., and Wang, W., "Mobility management for next generation wireless systems," *Proceedings of IEEE*, vol. 87, no. 8, pp. 1347–1384, 1999.

[9] Akyildiz, I. F., Mohanty, S., and Xie, J., "A ubiquitous mobile communication architecture for next generation heterogeneous wireless systems," *IEEE Communications Magazine*, vol. 43, no. 6, pp. 29–36, 2005.

[10] Akyildiz, I. F., Morabito, G., and Palazzo, S., "TCP-Peach: a new congestion control scheme for satellite IP networks," *IEEE/ACM Trans. Networking*, vol. 9, no. 3, pp. 307–321, 2001.

[11] Akyildiz, I. F. and Wang, W., "A dynamic location management scheme for next generation multitier PCS systems," *IEEE Trans. Wireless Communications*, vol. 1, no. 1, pp. 178–189, 2002.

[12] Akyildiz, I. F. and Wang, W., "A Predictive User Mobility Profile for Wireless Multimedia Networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 6, pp. 1021–1035, 2004.

[13] Akyildiz, I. F., Xie, J., and Mohanty, S., "A survey on mobility management in next generation all-IP based wireless systems," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 16–28, 2004.

[14] Ala-Laurila, J., Mikkonen, J., and Rinnemaa, J., "Wireless LAN Access Network Architecture for Mobile Operators," *IEEE Communications Magazine*, 2001.

[15] Arkko, J. and Haverinen, H., "EAP AKA Authentication," *Internet Draft, draft-arkko-pppest-eap-aka-09.txt, work in progress*, 2003.

[16] Austin, M. D. and Stuber, G. L., "Velocity Adaptive Handoff Algorithms for Microcellular Systems," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, pp. 549–561, 1994.

[17] Azemi, G., Senadji, B., and Boashash, B., "Velocity estimation in cellular systems based on the time-frequency characteristics of the received signal," in *Proc. ISSPA*, pp. 509–512.

[18] Banerjee, N., Basu, K., and Das, S., "Hand-off delay analysis in SIP-based mobility management in wireless networks," in *Proc. of the International Workshop on Wireless, Mobile and Ad Hoc Networks (WMAN).*

[19] Banerjee, N., Wu, W., Das, S., Dawkins, S., and Pathak, J., "Mobility support in wireless Internet," *IEEE Wireless Communications*, vol. 10, no. 5, 2003.

[20] Bao, G., "Performance evaluation of TCP/RLP protocol stack over CDMA wireless links," *ACM/Springer Journal of Wireless Networks (WINET)*, 1996.

[21] Bettstetter, C., Hartenstein, H., and Pérez-Costa, X., "Stochastic Properties of the Random Waypoint Mobility Model," *ACM/Kluwer Wireless Networks, Special Issue on Modeling and Analysis of Mobile Networks*, vol. 10, no. 5, pp. 555–567, 2004.

[22] Blunk, L. and Vollbrecht, J., "PPP Extensible Authentication Protocol (EAP)," *RFC 2284, IETF*, 1998.

[23] Buddhikot, M. M., Chandranmenon, G., S. Han, Y. L., Miller, S., and Salgarelli, L., "Design and implementation of a WLAN/CDMA2000 interworking architecture," *IEEE Communications Magazine*, vol. 41, no. 11, pp. 90–100, 2003.

[24] C. Perkins, P. R. C. and Bharatia, J., "Mobile IPv4 challenge/response extensions (revised)," *Internet Draft, draft-ietf-mobileip-rfc3012bis-05.txt, work in progress*, May 2003.

[25] Calhoun, P. and Perkins, C., "Mobile IP network access identifier extension for IPv4," *RFC 2290, IETF*, 2000.

[26] Calhoun, P. R., "Diameter Mobile IPv4 application," *Internet Draft, draft-ietf-aaa-diameter-mobileip-16.txt, work in progress*, 2004.

[27] Cheng, A. and *et al*, "Secure transparent Mobile IP for intelligent transportation systems," in *Proc. of the 2004 IEEE International Conference on Networking, Sensing and Controls*.

[28] Chung, Y. and Cho, D., "Velocity estimation using adaptive array antennas," in *Proc. IEEE Veh. Tech. Conf.*, pp. 2565–2569.

[29] Conrad, P. and *et al*, "SCTP in battlefield networks," in *Proc. of IEEE MILCOM*, pp. 289–95.

[30] Doumi, T. L. and Gardiner, J. G., "Use of base station antenna diversity for mobile speed estimation," *Electron. Lett.*, vol. 30, no. 22, pp. 1835–1836, 1999.

[31] Garg, V. and Wilkes, J. E., "Interworking and interoperability issues for North American PCS," *IEEE Communications Magazine*, vol. 34, no. 3, pp. 94–99, 1996.

[32] Ghaheri-Niri, S. and Tafazolli, R., "Cordless-cellular network integration for the 3rd generation personal communication systems," in *Proc. IEEE Vehicular Technology Conference (VTC'98)*, pp. 402–408.

[33] Glass, S., Hiller, T., Jacobs, S., and Perkins, C., "Mobile IP authentication, authorization, and accounting requirements," *RFC 2977, IETF*, 2000.

[34] Gustafsson, E. and Jonsson, A., "Always best connected," *IEEE Wireless Communications*, vol. 10, no. 1, pp. 49–55, 2003.

[35] Gustafsson, E., Jonsson, A., and Perkins, C. E., "Mobile IPv4 regional registration," *Internet Draft, draft-ietf-mip4-reg-tunnel-00.txt, work in progress*, 2004.

[36] Haverinen, H., Asokan, N., and Maattanen, T., "Authentication and key generation for Mobile IP using GSM authentication and roaming," in *Proc. IEEE ICC (ICC'01)*, pp. 2453–2457.

[37] Haverinen, H. and Salowey, J., "EAP SIM authentication," *Internet Draft, draft-haverinen-pppest-eap-sim-16.txt, work in progress*, 2004.

[38] Havinga, P. and *et al*, "The SMART project: exploiting the heterogeneous mobile world," in *Proc. 2nd International Conference on Internet Computing*.

[39] Henderson, T. R., "Host mobility for IP networks: a comparison," *IEEE Network*, vol. 17, no. 6, pp. 18–26, 2003.

[40] Holtzman, J. M. and Sampath, A., "Adaptive Averaging Methodology for Handoffs in Cellular Systems," *IEEE Tran. on Veh. Tech.*, vol. 44, no. 1, pp. 59–66, 1995.

[41] Hsieh, H.-Y., Kim, K.-H., and Sivakumar, R., "An End-to-End Approach for Transparent Mobility across Heterogeneous Wireless Networks," *ACM/Kluwer Mobile Networks and Applications Journal (MONET), Special Issue on Integration of Heterogeneous Wireless Technologies*, vol. 9, no. 4, pp. 363–378, 2004.

[42] Hsieh, R., Zhou, Z. G., and Seneviratne, A., "S-MIP: A seamless handoff architecture for Mobile IP," in *Proc. of IEEE INFOCOM'03*.

[43] Hu, N. and Steenkiste, P., "Evaluation and characterization of available bandwidth probing techniques," *IEEE Journal of Selected Areas in Comm.*, vol. 21, no. 6, pp. 879–894, 2003.

[44] Hui, S. Y. and Yeung, K. H., "Challenges in the migration to 4G mobile systems," *IEEE Communications Magazine*, vol. 41, no. 12, pp. 54–59, 2003.

[45] Inamura, H., Montenegro, G., Ludwig, R., Gurtov, A., and Khafizov, F., "TCP over second (2.5G) and third (3G) generation wireless networks," *Internet Draft, draft-ietf-pilc-2.5g3g-12*, Dec. 2002.

[46] Karn, P., "Qualcomm white paper on mobility and IP addressing." http://people.qualcomm.com/karn/papers/mobility.html.

[47] Kawabata, K., Nakamura, T., and Fukuda, E., "Estimating velocity using diversity reception," in *Proc. IEEE Veh. Tech. Conf.*, pp. 371–374.

[48] Lai, K. and Maker, M., "Measuring link bandwidth using a deterministic model of packet delay," in *Proc. of ACM SIGCOMM*.

[49] Liebsch, M. and *et al.*, "Candidate access router discovery," *Internet Draft, draft-ietf-seamoby-card-protocol-07.txt, work in progress*, 2004.

[50] Malki, K. E. and *et al.*, "Low latency handoff in Mobile IPv4," *Internet Draft, draft-ietf-mobileip-lowlatency-handoffs-v4-01.txt, work in progress*, May 2001.

[51] Maltz, D. and Bhagwat, P., "MSOCKS: an architecture for transport layer mobility," in *Proc. of IEEE INFOCOM'98*, pp. 1037–45.

[52] McNair, J., Akyildiz, I. F., and Bender, M., "An Inter-System Handoff Technique for the IMT-2000 System," in *Proc. of IEEE INFOCOM'00*.

[53] Min, S. and Lee, K. B., "Channel estimation based on pilot and data traffic channels for DS/CDMA systems," in *Proc. IEEE Global Telecommunications Conference*, (Sidney), pp. 1384–1388, 1998.

[54] Min-hua, Y., Yu, L., and Hui-min, Z., "The Mobile IP Handoff Between Hybrid Networks," in *Proc. IEEE PIMRC 2002*.

[55] Misra, A., Das, S., Dutta, A., McAuley, A., and Das, S., "IDMP-based fast handoffs and paging in IP-based 4G mobile networks," *IEEE Communications Magazine*, 2002.

[56] Mohanty, S., "VEPSD: Velocity estimation using the PSD of the received signal envelope in next generation wireless systems," *to appear in IEEE Transactions on Wireless Communications*, 2005.

[57] Mohanty, S. and Akyildiz, I. F., "A Cross-Layer (Layer 2 + 3) Handoff Management Protocol for Next Generation Wireless Systems," *to appear in IEEE Trans. on Mobile Computing*, 2005.

[58] Mohanty, S. and Akyildiz, I. F., "Analytical Modeling for Handoff Performance Evaluation of Mobility Management Protocols Based on Mobile IP, TCP-Migrate, and SIP," *under preparation*, 2005.

[59] Mottier, D. and Castelain, D., "A Doppler estimation for UMTS-FDD based on channel power statistics," in *Proc. IEEE Veh. Tech. Conf.*, pp. 3052–3056.

[60] Narasimhan, R. and Cox, D. C., "Speed estimation in wireless systems using wavelets," *IEEE Tran. Commun.*, vol. 47, no. 9, pp. 1357–1364, 1999.

[61] Neumaier, A., *Introduction to Numerical Analysis*.

[62] Papoulis, A. and Pillai, S. U., *Probability, random variables, and stochastic processes*. Mc Graw Hill, 4 ed.

[63] Perkins, C., "IP mobility support for IPv4," *RFC 3220, IETF*, 2002.

[64] Perkins, C. E. and Johnson, D. B., "Route optimization in Mobile IP," *Internet Draft, draft-ietf-mobileip-optim-11.txt, work in progress*, 2001.

[65] Povey, G. L. R., Grant, P. M., and Pringle, R. D., "A decision-directed spread-spectrum RAKE receiver for fast-fading mobile channels," *IEEE Trans. on Vehicular Technology*, vol. 45, 1996.

[66] Priscoli, F. D., "Interworking of a satellite system for mobile multimedia applications with the terrestrial networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 17, no. 2, pp. 385–394, 1999.

[67] Ramjee, R., Varadhan, K., Salgarelli, L., Thuel, S. R., Wand, S.-Y., and Porta, T. L., "HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396–410, 2002.

[68] Rigney, C. and *et al*, "Remote Authentication Dial In User Service (RADIUS)," *RFC 2865, IETF*, 2000.

[69] Rosenberg, J. and *et al.*, "SIP: Session Initiation Protocol," *RFC 3261, IETF*, 2002.

[70] Salgarelli, L., "EAP SKE authentication and key exchange protocol," *Internet Draft, draft-salgarelli-pppext-eap-ske-03.txt, work in progress*, May 2003.

[71] Salkintzis, A. K., Fors, C., and Pazhyannur, R., "WLAN-GPRS Integration for Next-Generation Mobile Data Networks," *IEEE Wireless Communications*, 2002.

[72] Seol, S., Kim, M., Yu, C., and Lee, J., "Experiments and Analysis of Voice over Mobile IP," in *Proc. of The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2002)*.

[73] Snoeren, A. C. and Balakrishnan, H., "An end-to-end approach to host mobility," in *Proc. of 6th ACM/IEEE Internation Conference on Mobile Computing and Networking (Mobicom'00)*.

[74] Stemm, M. and Katz, R. H., "Vertical handoffs in wireless overlay networks," *ACM/Springer Journal of Mobile Networks and Applications (MONET)*, vol. 3, no. 4, pp. 335–350, 1998.

[75] Stemm, M. and Katz, R. H., "Vertical handoffs in wireless overlay networks," *ACM/Springer Journal of Mobile Networks and Applications (MONET)*, vol. 3, no. 4, pp. 335–350, 1998.

[76] Stewart, R. and *et al.*, "Stream control transmission protocol," *RFC 2960, IETF*, 2000.

[77] Stuber, G. L., *Principles of Mobile Communication*.

[78] Tepedelenlioglu, C. and Giannakis, G. B., "On Velocity Estimation and Correlation Properties of Narrow-Band Mobile Communication Channels," *IEEE Trans. on Vehicular Technology*, vol. 50, no. 4, pp. 1039–1052, 2001.

[79] Tsao, S. and Lin, C., "Design and Evaluation of UMTS-WLAN Interworking Strategies," in *Proc. IEEE VTC 2002*.

[80] Tsao, S. and Lin, C., "VGSN : A Gateway Approach to Interconnect UMTS/WLAN Networks," in *Proc. IEEE PIMRC 2002*.

[81] Turkboylari, M. and Stuber, G. L., "Eigen-matrix pencil method-based velocity estimation for mobile cellular radio systems," in *Proc. IEEE Veh. Tech. Conf.*, pp. 690–694.

[82] Valko, A., "Cellular IP: A new approach to Internet host mobility," *ACM SIGMOBILE Computer Communication Review*, vol. 29, no. 1, pp. 50–65, 1999.

[83] Varma, V. K., Ramesh, S., and *at el*, "Mobility Management in Integrated 3G/WLAN Networks," in *Proc. IEEE ICC '03*.

[84] Wakikawa, R., Uehara, K., Ernst, T., and Nagami, K., "Multiple care-of addresses registration," *Internet Draft, draft-wakikawa-mobileip-multiplecoa-03.txt, work in progress*, 2004.

[85] Wang, L., Silventoinen, M., and Honkasalo, Z., "A new algorithm for estimating mobile speed at the TDMA-based cellular system," in *Proc. IEEE Veh. Tech. Conf.*, pp. 1145–1149.

[86] Wang, W. and Akyildiz, I. F., "On the estimation of user mobility pattern for location tracking in wireless networks," in *Proc. IEEE GLOBECOM*.

[87] Wedlund, E. and Schulzrinne, H., "Mobility support using SIP," in *Proc. of Second ACM/IEEE International Conference on Wireless and Mobile Multimedia (WoWMoM'99)*.

[88] Wisely, D. and Mitjana, E., "Paving the Road to Systems Beyond 3G-The IST BRAIN and MIND Projects," *Journal of Communications and Networks*, vol. 4, no. 4, pp. 292–301, 2002.

[89] Xiao, C., Mann, K. D., and Olivier, J. C., "Mobile speed estimation for TDMA-based hierarchical cellular systems," *IEEE Tran. on Veh. Tech.*, vol. 50, no. 4, pp. 981–991, 2001.

[90] Xie, J., "User-independent paging scheme for Mobile IP," in *Proc. of IEEE Globecom*.

[91] Xie, J. and Akyildiz, I. F., "A Novel Distributed Dynamic Location Management Scheme for Minimizing Signaling Costs in Mobile IP," *IEEE Trans. Mobile Computing*, vol. 1, no. 3, pp. 163–175, 2002.

[92] Xie, J. and Akyildiz, I. F., "A hybrid control resource allocation scheme for policy-enabled handoff in wireless heterogeneous overlay networks," *submitted for publication*, 2005.

[93] Yang, Y. R. and Lam, S. S., "General AIMD congestion control," in *Proc. of the 2000 International Conference on Network Protocols*.

[94] Yokota, H., Idoue, A., Hasegawa, T., and Kato, T., "Link layer assisted Mobile IP fast handoff method over Wireless LAN Networks," in *Proc. of ACM MobiCom'02*, pp. 131–139.

[95] Zhang, N. and Holtzman, J. M., "Analysis of handoff algorithms using both absolute and relative measurements," *IEEE Trans. on Vehicular Technology*, vol. 45, no. 1, 1996.

[96] Zhang, Q. and *et al.*, "Efficient mobility management for vertical handoff between WWAN and WLAN," *IEEE Communications Magazine*, 2003.

[97] Zhang, Q., Guo, C., Guo, Z., and Zhu, W., "Efficient Mobility Management for Vertical Handoff between WWAN and WLAN," *IEEE Communications Magazine*, vol. 41, no. 11, 2003.

[98] Zhao, X., Castelluccia, C., and Baker, M., "Flexible network support for mobility," in *Proc. of ACM Mobicom*.

VITA

Shantidev Mohanty was born in Kushi, Jajpur, Orissa, India in April 1978. He received his B. Tech. (Hons.) degree from the Indian Institute of Technology, Kharagpur, India and the M.S. degree from the Georgia Institute of Technology, Atlanta, Georgia, in 2000 and 2003, respectively, both in electrical engineering. He attended the doctoral program in the School of Electrical and Computer Engineering at the Georgia Insitute of Technology, Atlanta, GA, from August 2001 to December 2005. At the same time, he was a graduate research assistant in the Broadband and Wireless Networking Laboratory (BWN-LAB) at the Georgia Institute of Technology. He received the Doctor of Philosophy degree in December 2005 in Electrical and Computer Engineering from the Georgia Institute of Technology. His current research interests include wireless networks, mobile communications, mobility management, and cross-layer protocol design. From 2000 to 2001 he worked as a mixed signal design engineer for Texas Instruments, Bangalore, India. He worked as a summer intern for Bell Labs, Lucent Technologies, Holmdel, New Jersey, during the summers of 2002 and 2003 and for Applied Research, Telcordia Technologies, Piscataway, New Jersey, uring the summer of 2004.