Performance Improvement of Mobility Management in IP-based Wireless Networks

Ved Prasad Kafle

**DOCTOR OF
PHILOSOPHY**

Department of Informatics,

School of Multidisciplinary Science,

The Graduate University for Advanced Studies (SOKENDAI)

2006 (School Year)

September 2006

Advisory Committee:
Prof. Shigeki Yamada   National Institute of Informatics, SOKENDAI
Assoc. Prof. Shunji Abe   National Institute of Informatics, SOKENDAI
Prof. Yanghee Choi   Seoul National University
Assoc. Prof. Yusheng Ji   National Institute of Informatics, SOKENDAI
Prof. Tomohiro Yoneda   National Institute of Informatics, Tokyo Institute of Technology

(Alphabetical order of last name except chair)

Abstract

The proliferation of wide band wireless networks and the availability of abundant services through the Internet have necessitated a ubiquitous access to the Internet by mobile hand-held devices such as hand phones, laptop computers, and personal digital assistants (PDAs). Researches on mobility management of a single mobile node in the existing cellular networks as well as in the IP-based networks have extensively been done by the respective research communities. However, as new networking systems such as network mobility and overlapped heterogeneous access networks have been emerging, mobility management of mobile nodes to maintain seamless network connectivity in these systems is being a challenging task. In this dissertation, we present our research on designing, modeling, evaluating, and optimizing the mobility management solutions in regard to these networking systems.

Network mobility management: Network mobility (NEMO) management concerns with mobility management of an entire network (called mobile network) that provides an uninterrupted Internet connectivity to many mobile devices moving together in the mobile network. Although the NEMO basic support protocol, developed by the Internet Engineering Task Force (IETF), offers basic operations of NEMO management, it has the side effect of increasing packet delivery overheads due to pinball routing and multi-layer encapsulation of data packets. Moreover, the protocol does not address the large handover latency that causes a large number of packet losses and, consequently, communication service interruption of all the mobile nodes belonging to the mobile network. We address these problems by developing route optimization and handover management schemes.

We propose a mobile router-assisted route optimization (MoRaRo) scheme to reduce the delays and overheads associated with data delivery paths. This scheme is simple to implement because it requires only a slight change in the implementation of the NEMO basic support protocol in the local components of a mobile network, such as mobile routers and mobile nodes. No change is required in correspondent nodes, home agents, or any other network components located in domains beyond the mobile network. The MoRaRo scheme enables a correspondent node to forward packets directly to the mobile network without any tunneling. It thereby reduces packet delays and encapsulation overheads in the core network. We evaluated the performance of the MoRaRo scheme in terms of the packet delivery overheads, efficiency, and delay between correspondent nodes and mobile network nodes. We found that the scheme improved the packet delivery efficiency of the network mobility support protocol by about 100%. In addition, the packet delivery delay was significantly reduced.

We also propose a cooperative mobile router-based handover (CoMoRoHo) scheme for

long, vehicular-multihomed mobile networks. This scheme exploits the advantage of a multihoming environment of a long, vehicular mobile network that can have two or more mobile routers spatially separated by a certain distance. It makes the packet loss independent of the handover latency by establishing a local tunnel between an access router and a mobile router of the mobile network by using only one signaling message. We compared the performance of the CoMoRoHo scheme with the Fast Handover for Mobile IPv6 protocol by formulating analytical models, and found that this scheme outperformed in regard to the number of packet losses, signaling-message overhead and packet-delivery overhead in the access network. CoMoRoHo has good scalability because it works well, even when the access network is overloaded.

Overlapped heterogeneous networks: The other type of access network environment we are considering for mobility management is the overlapped heterogeneous wireless system. When there are different types of wireless networks available at a place, a mobile user with multi-mode network interfaces should be capable of effectively carrying out the following two functions: (a) select an optimal access network for a given application, and (b) gracefully transfer a connection from one access network to another when the previous one becomes suboptimal or unavailable. In this dissertation, we address the issue of an optimal network selection by evaluating network performances from the user's perspective, and consequently proposing an access network selection algorithm. Similarly, to optimize the connection transfer process, we develop a graceful vertical handover mechanism by harmonizing the functions performed by different layers of the networking system.

To select an optimal network that maximizes user satisfaction, we formulated the bandwidth utility functions (BUF) for three types of applications: rigid, elastic, and adaptive. We use this BUF and handover latency to derive the user satisfaction function. Through analysis, we found that selecting a high bandwidth access network does not guarantee higher user satisfaction if the user happens to perform a handover to a lower bandwidth network. It is also observed that after getting an estimate of the user's movement, we should assign the network with most availability to the user's call request so that the user can remain in the same network throughout the call duration. By doing this, we can prevent the user satisfaction from being degraded. In accordance with the evaluation, we propose an algorithm for selecting an access network that maximizes user satisfaction in heterogeneous networks.

When the selected network becomes suboptimal or unavailable due to some reason such as mobility and increasing network congestion, a mobile node has to perform a vertical handover to a new network, in most cases, to an inferior one. In such cases, to smoothen the inevitable performance change caused by the resource limitations of the new network, we have devised a graceful vertical handover scheme. This scheme harmonizes the functions performed by different layers, i.e., from the link layer to the application layer, and consequently smoothens the

change in the throughput and reduces the handover-related data losses in the network. Through simulations, we evaluated the TCP performance with the graceful handover, which is significantly better than the performance without one. The graceful handover avoids packet losses by smoothing TCP sending rates and establishing a route to the previous access router via a relay node.

Student Number: 032A18

Acknowledgements

Contents

List of Figures

iv

v

List of Tables

List of Acronyms

3G Third Generation

3GPP Third Generation Partnership Project

ABC Always Best Connected

ABR Available Bit Rate

Ack Acknowledgement

AIPN All IP Network

AP Access Point

AR Access Router

AR-BU Access Router Binding Update

BACK Binding Acknowledgement

BU Binding Update

BUF Bandwidth Utility Function

CA Certificate Authority

CBR Constant Bit Rate

CBU Correspondent Binding Update

CN Correspondent Node

CoA Care-of Address

CoMoRoHo Cooperative Mobile Router-based Handover

CR Correspondent Router

DNS Domain Name Server

FBack Fast Binding Acknowledgement

FBU Fast Binding Update

FMIPv6 Fast Mobile IP version 6

FNA Fast Neighbor Advertisement

FTP File Transfer Protocol

GDR Graceful Degradation Rate

GDT Graceful Degradation Time

GSM Global System for Mobile communications

HA Home Agent

HA-BU Home Agent Binding Update

HA-BACK Home Agent Binding Acknowledgement

HACK Handover Acknowledgement

HAWAII Handoff Aware Wireless Access Internet Infrastructure

HI Handover Indication

HMIPv6 Hierarchical Mobile IP version 6

HoA Home Address

IETF Internet Engineering Task Force

IP Internet Protocol

IPSec Internet Protocol Security

IPv4 Internet Protocol version 4

IPv6 Internet Protocol version 6

kbps Kilo Bit Per Second

LAN Local Area Network

LFN Local Fixed Node

LMN Local Mobile Node

MAP Mobility Anchor Point

Mbps Mega Bit Per Second

MIPv6 Mobile Internet Protocol version 6

MN Mobile Node

MNN Mobile Network Node

MNP Mobile Network Prefix

MoRaRo Mobile Router-assisted Route optimization

MR Mobile Router

MRBC Mobile Router Binding Cache

MSN Mobility Support Node

NA Neighbor Advertisement

NAACK Neighbor Advertisement Acknowledgement

NAR New Access Router

NCoA New Care-of Address

NEMO Network Mobility

NII National Institute of Informatics

ns-2 Network Simulator – 2

ORC Optimized Route Cache

PAN Personal Area Network

PAR Previous Access Router

PCH Path Control Header

PCoA Previous Care-of Address

PDA Personal Digital Assistant

PHS Personal Handy-phone System

PR Personal Router

PrRtAdv Proxy Router Advertisement

QoS Quality of Service

RFC Request For Comments

RH Routing Header

RN Relay Node

RO Route Optimization

ROBC Route Optimization Binding Cache

RR Return Routability

RRH Reverse Routing Header

RtAdv Router Advertisement

RTO Retransmission Timeout

RtSol Router Solicitation

RtSolPr Router Solicitation for Proxy

RTT Round Trip Time

SA Security Association

TCP Transmission Control Protocol

UMP User Mobility Pattern

USF User Satisfaction Function

VMN Visiting Mobile Node

VBR Variable Bit Rate

WLAN Wireless Local Area Network

WWAN Wireless Wide Area Network

Chapter 1

Introduction

Wireless networks such as the GSM (Global System for Mobile communications) networks, third generation (3G) cellular networks, wireless local area networks (LANs), and PHS (Personal Handy phone System) networks are composed of many small units of coverage called cells. When mobile devices like hand phones, personal digital assistants (PDAs), and laptop or palmtop computers move from one cell to another, they have to carry out a number of tasks to remain connected to the network and continue ongoing communication services. The process of managing these tasks is called mobility management of mobile devices or mobile nodes. The mobility management operation should be carried out transparently to mobile users such that the users are unaware of the process and time of changing cells; otherwise, mobility will adversely affect the network service quality perceived by the users. Moreover, it should be very efficient in terms of both the network resource (e.g., bandwidth) consumption and the mobile nodes' battery and computational power consumption.

1.1   IP Mobility Management

In the existing wireless systems, mobility of mobile nodes that move within the same type of wireless network is managed by the link-specific mechanism, which works well only for the given link-access technology. However, the next generation of mobile communication networks is expected to be a heterogeneous multimedia system widely comprising different radio access networks. Each of these networks may possess some advantages over the others in terms of bandwidth, coverage, cost, reliability, etc. To exploit these advantages, the heterogeneous system may appear in an overlay form [1, 2]; one access network (e.g., wireless LAN) overlapping the service area of the other access networks (e.g., 3G networks). In such networking environments, a mobile node may access any wireless network that best serves the user's requirements.

1

For instance, the wireless LAN may be accessed for a bulk data transfer, whereas the 3G cellular networks may be used for low-bandwidth voice calls. Therefore, having a common mechanism that can manage the mobility of all mobile nodes in all types of wireless networks is being an essential requirement for realizing the future ubiquitous computing systems. For heterogeneous wireless systems, Mobile IP protocols (for IP version 4 as well as version 6) [3, 4] are considered to be universal solutions for mobility management because they can hide the heterogeneity in the link-specific technologies used in different networks. Although Mobile IP protocols deal well with the mobility of a single mobile node on the Internet topology, they are not sufficient to handle emerging new types of networks, such as moving or mobile networks. Moreover, besides maintaining the IP layer connectivity, there are other issues of mobility management, such as optimal network selection in heterogeneous overlay systems and smooth transfer of connections from one network to another when the previous one becomes suboptimal. This dissertation investigates these new type of problem domains and proposes some solutions.

1.2   Network Mobility Management

In vehicles or personal area networks, a number of wireless devices move together and change their Internet connection points simultaneously. For example, when a number of wireless devices are carried by a person, these devices move together when the person moves; or when a number of communicating devices are installed on moving vehicles such as buses, trains, ships, and aircrafts, these devices along with the mobile devices carried by the passengers move together when the vehicles move. In such cases, managing mobility of individual devices via the devices themselves would be inefficient and sometimes unfeasible [19]. Therefore, having a mobile network to serve the group of mobile nodes that move together, and then managing mobility of the mobile network instead of the individual mobile nodes would provide better mobility management solutions. Considering this fact, the Network Mobility (NEMO) working group [20] within the Internet Engineering Task Force (IETF) has proposed a conceptual architecture of mobile networks and developed the NEMO basic support protocol [22].

A mobile network, which is composed of one or more IP subnets, moves as a single unit on the Internet topology [21]. It has at least one mobile router and many mobile network nodes (MNNs). The mobile router uses the NEMO basic support protocol to provide uninterrupted connectivity to MNNs. When the mobile router moves away from its home link, it configures a care-of address in the visited link and registers the address with its home agent by sending a prefix-scope binding-update message. The home agent then creates a binding cache entry associating the mobile network's subnet prefix with the mobile router's care-of address. Following the registration, the home agent intercepts data packets addressed not only to the mobile router

but also to any MNNs that have obtained an address from the mobile network's subnet prefix. The home agent then tunnels the packets to the mobile router's current location.

Although the NEMO basic support protocol offers the basic operation of NEMO management, it has the side effect of increasing packet delivery overheads and delays due to pinball routing and multi-layer encapsulation of data packets. Moreover, the protocol does not address the large handover latency that causes a large number of packet losses and communication service interruption of all the mobile nodes belonging to the mobile network.

To solve these problems, we have extended the NEMO management operation by developing a mobile router-assisted route optimization (MoRaRo) [8] scheme and a cooperative mobile router-based handover (CoMoRoHo) [9] scheme. With MoRaRo, a mobile node performs route optimization with a correspondent node only once at the beginning of a communication session, and on handover, i.e., when the mobile network moves, the mobile router performs route optimization on behalf of all active mobile nodes. The virtue of MoRaRo is that it requires only a slight modification of the implementation of the NEMO basic support protocol at local entities such as the mobile router and mobile nodes, leaving entities in the core or in other administrative domains untouched. The MoRaRo scheme enables a correspondent node to forward packets directly to the mobile network without any tunneling, thus reducing packet delay and encapsulation overheads in the core network. To enable the scheme to be evaluated, we present the results of both mathematical analysis and simulation.

The basic idea behind the CoMoRoHo scheme is to enable different mobile routers of a multihomed mobile network to access different subnets during a handover and cooperatively receive packets destined for each other. That is, when a mobile router is performing handover to a new subnet, it requests the access router of the old subnet to tunnel packets to some other mobile router that is still located in the old subnet. The other mobile router then forwards the packets to the mobile router that has requested the access router for tunneling. Similarly, when the mobile router completes its handover to the new subnet, it helps the other mobile router to perform a lossless handover to the new subnet. To evaluate the CoMoRoHo scheme, we carried out its performance modeling in regard to the handover latency, packet loss, signaling overhead, and packet delivery overhead in access networks.

1.3 Access Network Selection and Graceful Vertical Handover

A mobile node with multi-mode network interfaces can take advantages of the availability of overlapped wireless networks only when the mobile node is equipped with mechanisms that select an optimal network and smoothly transfer a connection from one access network to another when the previous one becomes suboptimal or unavailable. Without such mechanisms, mobile

nodes always try to access higher bandwidth networks because bandwidth-hungry multimedia applications are increasing on the Internet. However, if all the mobile nodes try to use the high bandwidth network, they have to frequently perform a vertical handover to lower bandwidth networks due to various reasons, such as limited coverage and increasing network congestion. Consequently, the handover latency and the change of bandwidth during a communication session may greatly affect the quality of service perceived by the user.

We address the issue of selecting an optimal wireless network by defining a user satisfaction function as a function of bandwidth utility and handover latency [11]. The bandwidth utility function of a network service, in turn, is a measure of the degree of network capability to fulfill bandwidth requirements of multimedia applications. We analyze how a user's movement across heterogeneous networks having varied bandwidths affects user satisfaction. We also investigate the effect that the change in user contexts such as call holding time and user movement probability has on the satisfaction function. We found that selecting a high bandwidth access network does not necessarily guarantee higher user satisfaction if the mobile user happens to perform a vertical handover to a low bandwidth network soon after the beginning of the call. It is also observed that after getting an estimate of the user's movement, we should assign the network with the most availability to user's call request, so that the user will remain in the same network throughout the call duration, thus maximizing user satisfaction. Based on the evaluation, we proposed an optimal network selection algorithm to maximize the user satisfaction in the overlapped heterogeneous network environment.

In case the optimal network chosen by the above algorithm becomes unavailable or suboptimal, the mobile node has to perform a vertical handover to a new network, in most cases, to an inferior one. Because of the disparities in the overlay networks' such characteristics as bandwidths, latencies, and loss rates, the handover process may result in an abrupt change in the throughput, a large number of packet losses, and disruption of the communication services [52, 66]. To solve these problems, we present a new scheme for vertical handover management, which we call the *graceful vertical handover* scheme. This scheme provides a graceful degradation of performance by smoothing the throughput change and reducing the possible packet losses during a handover. To achieve such a gracefulness in performance change, this scheme implements its functionalities in the link layer, IP layer, and TCP or application layer. Through simulations, we evaluated the TCP performance with the graceful handover, which is significantly better than the performance without one.

1.4 Organization

This dissertation is organized as follows. In Chapter 2, we overview related work on IP network mobility and heterogeneous networks. The subsequent first two chapters deal with network mobility and the next two chapters present the schemes for optimal access network selection and graceful vertical handover in heterogeneous overlay wireless networks. We present a mobile router-assisted route optimization (MoRaRo) scheme in Chapter 3, and a cooperative mobile router-based handover (CoMoRoHo) scheme in Chapter 4. In Chapter 5, we define a user satisfaction function, evaluate the impact that different parameters such as user mobility, call holding time, and handover latency have on the user satisfaction function, and propose an optimal network selection algorithm. We describe a graceful vertical handover mechanism in Chapter 6. Finally, we summarize the dissertation and outline future work in Chapter 7.

Chapter 2

Background and Related Works

2.1   Node Mobility – Mobile IP Protocols

The proliferation of wideband wireless networks and the availability of abundant multimedia services through the Internet have necessitated a ubiquitous access to the Internet using mobile handheld devices, such as hand phones, laptop or palmtop computers, and personal digital assistants (PDAs). Mobile IP protocols [3, 4] play a major role in bringing the IP-technology in the mobile world. These protocols aim at maintaining the IP-layer connectivity between mobile devices and the network to provide uninterrupted communication services to mobile devices, irrespective of their mobility in the IP-based wireless access networks. Since Internet Protocol version 6 (IPv6) is superior in many aspects, such as mobility management, security, and efficient routing, to its precursor Internet Protocol version 4 (IPv4), and IPv6 is expected to be an indispensable protocol for the near future Internet, we choose Mobile IPv6 [4] and its derivatives for our this research.

In Mobile IPv6 (MIPv6) protocol, besides mobile devices, called mobile nodes, there are home agents (HAs) and correspondent nodes (CNs). The mobile node and the correspondent node are peer nodes exchanging data packets between them. Whereas, the home agent is a router on the mobile node's home link, which forwards packets to/from the mobile node when the mobile node is away from its home link. The mobile node possesses two addresses: a home address (HoA) to identify the node, and a care-of address (CoA) to locate the node. The HoA is an IP address assigned to the mobile node from its home subnet prefix on its home link. The CoA, on the other hand, is configured from a foreign subnet prefix when the mobile node is attached to the foreign link away from home. Upon having a CoA, the mobile node registers it at the home agent by sending a Binding Update message that creates a binding between the mobile node's HoA and CoA at the home agent.
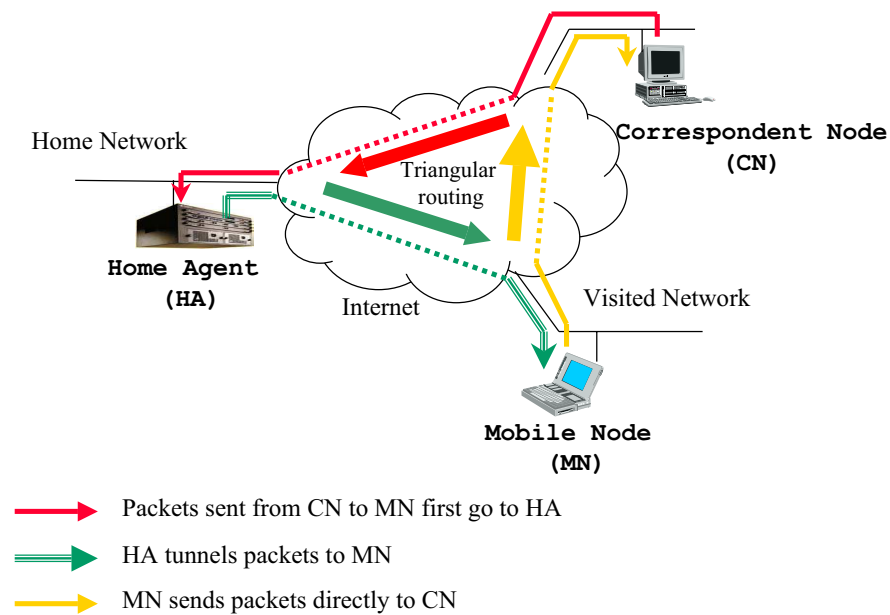
Figure 2.1: Triangular routing in Mobile IPv6.

Figure 2.1 shows the packet forwarding paths in MIPv6 protocol. A correspondent node, which may be a stationary node or a mobile node, initially knows only the HoA of a mobile node. When the correspondent node initiates a connection to the mobile node, it thus sends packets to the HoA of the mobile node. These packets reach the home link of the mobile node, where the home agent intercepts the packets and tunnels them to the current location, i.e., CoA of the mobile node using IP-in-IP encapsulation [7]. The mobile node, on the other hand, can send packets directly to the correspondent node, without going through the home agent. Therefore, a triangular routing occurs among the mobile node, correspondent node, and home agent. To eliminate the triangular routing problem, the mobile node can provide its location information to the correspondent node by performing a correspondent registration. For this purpose, the mobile node first performs a return routability test to get authentication information from the correspondent node, and then sends a binding update message to establish a binding between the mobile node's HoA and CoA at the correspondent node. After the registration, the correspondent node addresses the mobile node by its CoA. As a result, the packets are routed directly to the current location of the mobile node.

To localize signaling messages and reduce the handover latency of frequently moving mobile nodes, the Internet Engineering Task Force (IETF) has extended the MIPv6 operation to access networks and drafted two protocols: Hierarchical MIPv6 (HMIPv6) [5] and Fast Handover for MIPv6 [6]. The first protocol assumes a hierarchical structure of routers in the foreign

domain to localize signaling messages, whereas the second protocol assumes the cooperative routers that provide packet-tunneling services between them during a handover of mobile nodes. However, both of these schemes require a special configuration of the foreign networks, which may not be feasible in heterogeneous mobile network systems where the component networks may have no relationship or cooperation amongst them.

2.2   Network Mobility – NEMO Basic Support Protocol

Network mobility (NEMO) concerns itself with the mobility management of a moving network or mobile network. A mobile network, which is composed of one or more IP subnets, moves as a single unit on the Internet topology [21]. As shown in Fig. 2.2, a mobile network uses a mobile router (MR) as a gateway to provide Internet connectivity via an access router (AR) to the mobile network nodes (MNNs). The MNNs are categorized into three groups: local fixed nodes (LFNs), local mobile nodes (LMNs), and visiting mobile nodes (VMNs). The home HoAs of the LFNs and LMNs are associated with the mobile network's subnet prefix, whereas the HoAs of the VMNs are associated with other networks. Therefore, a MIPv6-enabled VMN arriving at a mobile network first configures a CoA from the mobile network's IP prefix and then registers the CoA with its home agent (HA). A VMN may represent a single host or a network itself, such as a personal area network (PAN), resulting in nested mobility. In the case of nesting, MRs belonging to each mobile networks form a hierarchy, with the (upper) parent-MR providing connectivity to (lower) sub-MRs. Only the top-level or root-MR performs mobility management when the whole mobile network moves as a single unit, with other sub-MRs and MNNs (local as well as visiting) remaining unaware of network mobility.

The NEMO basic support protocol [22], which is developed by the IETF, provides uninterrupted connectivity to MNNs by enhancing the operation of MIPv6 in mobile routers and their home agents. The basic operation of the protocol is illustrated in Fig. 2.2. When a mobile router moves away from its home link, it configures a CoA in the visited link and registers the CoA with its home agent by sending a binding update message. The binding update message in NEMO is slightly different from that in MIPv6. In MIPv6 a binding update message contains the CoA and HoA of a mobile node, whereas in NEMO the message contains additional information – the IP subnet prefix(es) of the mobile network, which is called the mobile network prefix (MNP). The presence of an MNP in the message directs the home agent to create a binding cache entry associating the MNP with the mobile router's CoA. Following the registration, the mobile router's home agent intercepts data packets addressed not only to the mobile router but also to any MMNs that have obtained an address from the MNP, and tunnels the packets to the mobile router's current location.
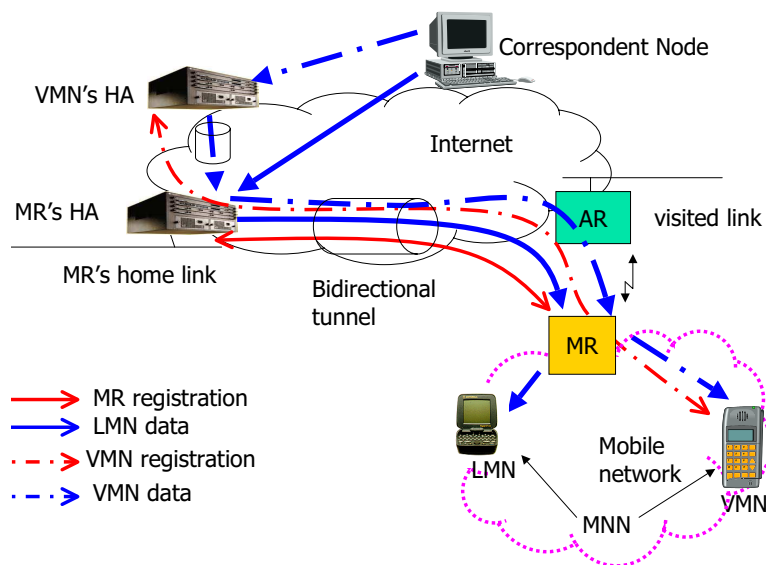
Figure 2.2: Conceptual architecture of mobile network and NEMO basic support protocol operation.

With the NEMO basic support protocol, only the mobile router configures and performs prefix scope binding updates with its home agent when the mobile network moves. As the MNNs are not required to configure and register their CoAs as long as they are inside the mobile network, the signaling volumes in wireless as well as wired networks are significantly reduced [23].

A correspondent node (CN) normally sends data packets to the HoA of a mobile node. As the HoAs of the local nodes are associated with the MNP, the mobile router's home agent intercepts the packets destined for local nodes and tunnels them to the mobile router's CoA. Similarly, packets addressed to a visiting mobile node first arrive at the node's home agent, which tunnels the packets to the nodes's CoA. Since the visiting mobile node's CoA belongs to the MNP, the packets go to the mobile router's home agent, which again tunnels them to the mobile router's CoA. For both local and visiting MNNs, the mobile router decapsulates packets and forwards them on to the interfaces to which the MNNs are attached. Similarly, outbound data packets originated from the MNNs follow the same route in the reverse direction.

The process of packet delivery between a CN and an MNN is full of overheads because of multi-angular sub-optimal routing and multi-layer encapsulation. This problem is worse if there are multiple levels of nested mobile networks where packets need to be tunneled from every mobile router's home agent, as shown by "Pinball path" in Fig. 2.3. This multi-angular routing problem is referred to as pinball routing [24]. Pinball routing consumes excessive bandwidth not only in the wired part but also in the wireless part of the network. When the access router forwards a multi-level encapsulated packet to an MNN, the encapsulation headers unnecessar-
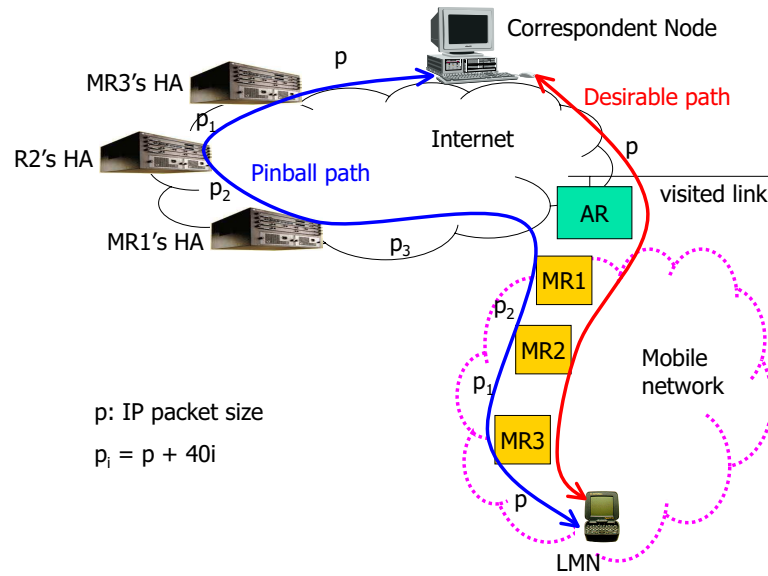
Figure 2.3: Packet routing paths in mobile network.

ily consume the bandwidth of wireless links between the access router and MNN. Moreover, pinball routing affects the network performance in several ways; it reduces network throughput, increases delays in packet delivery, and imposes excessive computational overheads on home agents. To overcome these problems, we propose a mobile router-assisted route optimization (MoRaRo) scheme. The objective of MoRaRo is to enable packets to travel along an optimal path from source to destination as indicated by "Desirable path" in Fig. 2.3. Previous work related to the route optimization in NEMO will be discussed in Chapter 3.

The mobile network concept described above finds an appropriate application in vehicles, especially in fast-moving, long vehicles such as trains. In such vehicles, a mobile router can provide Internet connectivity to a large number of MNNs. However, as the vehicle moves fast, the mobile router has to perform handovers frequently. In IP networks, a handover process generally consists of three sequential operations: link switching, CoA configuration, and binding update [4]. The handover latency, namely, the sum of times taken by these operations, can be sufficiently large to lead to the loss of a large number of packets belonging to many MNNs. This may result in impairing the quality of multimedia communications. Addressing this issue, we propose a cooperative mobile router-based handover (CoMoRoHo) scheme that minimizes packet losses of a long vehicular mobile network during a handover. We will explain more about previous work on network mobility handover management in Chapter 4.

2.3   Moving Networks in the 3GPP System

The 3rd Generation Partnership Project (3GPP) includes Moving Networks in its All IP Networks (AIPN) specifications [16]. The 3GPP's Moving Network concept is similar to the mobile network of the IETF, except that in the former the consolidated traffic is routed through a gateway that can be a moving base station, wireless access router, or mobile router. We describe about these gateways in the subsequent paragraphs.

A moving base station, which is part of an access system and is owned by the AIPN operator, is responsible to provide radio access to all the mobile devices of the Mobile Network. The mobile devices therefore have to use the wireless technology supported by the AIPN access system. In this case, mobility of the Moving Network is fully managed by the AIPN access system.

A wireless access router (e.g., a wireless LAN access point) can be owned by a third party network operator to connect to the AIPN access system. The wireless access router has freedom to implement its own radio technology to connect the user mobile devices. Moreover, the wireless access router is free to select any AIPN access system that best suits its requirements. In this case, mobility management functions are divided between the AIPN and the access systems. The access system manages mobility when the wireless access router moves between the cells of the same access system, whereas the AIPN manages mobility when the wireless router moves across the access systems.

The purpose and functions of a mobile router in Moving Networks are similar to that in the mobile networks defined by the IETF. The mobile router is equipped with some wireless technology to connect to the AIPN directly, i.e., without going through an access system. In this case, mobility of the mobile router is handled completely by the AIPN.

As explained in the previous section, the IETF mobile networks use only the mobile router to provide Internet connectivity to the mobile devices, and mobility of the mobile router is managed by the IPv6 network. In addition to the mobile router, the 3GPP AIPN's Moving Network may include moving base stations or wireless access routers. However, in our research we follow the IETF model of mobile networks and consider only the mobile router.

2.4   Heterogeneous Wireless Networks

Up until the last decade, voice and data communication networks were being independently developed. However, for the last few years the demarcation between voice networks and data networks has been vanishing as both networks are attempting to provide multimedia services over IP-based wireless mobile networks. For instance, the 3G networks provide both voice

and limited-bandwidth data services to mobile users, and the wireless LAN based on the IEEE 802.11 standards, which was initially designed for larger-bandwidth data services, is being developed to support voice services.

Nonetheless, the fact is that none of the networks are perfect in all aspects; the 3G networks provide wider coverage but limited bandwidths (i.e., less than 2 Mbps), whereas the wireless LANs provide larger bandwidths but limited coverage. Therefore, to utilize all the good points, we have to integrate these networks into overlay forms [1, 2]; that is, the wireless LAN covers some areas that have already been covered by the 3G networks.

In such heterogeneous network environments, a mobile node with multi-mode network interfaces, such as the one recently launched by NTT DoCoMo in Japan, can access a network that best serves the user's requirement. However, because bandwidth-hungry multimedia applications supported by wireless networks have been increasing, a mobile user located in the overlay area will always try to access a network that provides the maximum bandwidth. However, depending on the location and mobility characteristics, the user has to transfer its network connection to a lower bandwidth network when the higher bandwidth network becomes unavailable due to various reasons, such as limited network coverage, increasing congestion, and call admission control policies. The process of transferring a connection from one overlay to another is referred to as a vertical handover [64]. In particular, a vertical handover that occurs from a high bandwidth (low coverage) network to a low bandwidth (high coverage) network is called an upward vertical handover, and one that occurs in the reverse direction is called a downward vertical handover. On the other hand, the process of connection transfer from one cell to another of the same overlay is called a horizontal handover. Compared to horizontal handovers, there are more challenges posed for properly managing vertical handovers.

In a single homogeneous network like the 3G cellular network or wireless LAN, the horizontal handover is generally seamless. The term seamless indicates that the handover process is so fast and lossless that it goes transparent to user applications [65]. However, it would be impracticable to say that a vertical handover, especially an upward vertical handover, can always be seamless if there are large disparities in the overlay network characteristics, such as bandwidths, latencies, and loss rates [66]. These disparities may result in abrupt changes in the throughput, a number of packet losses, or disruption of services for some time during the handover [52].

In this thesis, we analyze the impact that mobility has on the network service quality from a user's prospective. We measure the user-centric performance of a network service by a user satisfaction function [53, 54, 55]. There has been a lot of research on the evaluation of network-centric performance or network-level quality of service (QoS), which is concerned with optimizing the network characteristics. The network-centric evaluation indicates, for instance, that

the larger the bandwidth, the better the network performs. However, it cannot answer the following question: how much large bandwidth and smaller latency or loss rate are appropriate for a user's application. To answer this question, we need to evaluate the user-centric performance. The user-centric performance, which is also a measure of the user-perceived QoS, relates user application requirements with the network service characteristics or quality. Note that a user requires network resources to be just sufficient enough to satisfy its application's requirements. Any extra resource beyond the requirement, may not give any additional benefit to the user. In such cases, users may not opt for a network that has the highest resources; they may rather select a network that provides the optimal performance at the lowest user-centric cost. The user-centric cost includes the price of network service as well as the resource consumption, such as the battery power of the mobile terminal. It is indisputable that a mechanism should exist that enables mobile users to carry out intelligent decisions for optimal network selection. Without such a mechanism, mobile users cannot get benefits from the availability of different types of networks; instead they would be overloaded with choices.

There are only a few published research papers dealing with this issue, however, from different aspects. Lee et al. [17] applied a software agent based approach in a personal router (PR) that lies between the user and network. The PR selects a suitable service for the user based on network information and user preferences. The PR continuously gets feedback of the user's subjective evaluation in terms of quality and cost of service, and accordingly adjusts the user preference parameters. However, their work does not include any quantitative analysis of the relationship between the user's requirements and the network's information. Moreover, the subjective evaluation of a particular user cannot be universally valid to all users. Altmann et al. [18] explained a mechanism for enabling users to select a service from a fixed number of priority-level/price pairs provided by a single network. Their work is more concerned with improving the efficiency of switching service levels when the observed network performance changes than evaluating the user-centric performances.

Complementing above work, we develop and evaluate an analytical framework for selecting an optimal network in a heterogeneous wireless network environment. In addition, we also propose a *graceful vertical handover* scheme, which provides graceful degradation of performance by smoothing the throughput change and reducing the possible packet losses during a handover.

Chapter 3

NEMO Route Optimization

3.1 Introduction

We know from experience with MIPv6 that route optimization (RO) can resolve the problems of sub-optimal routing. However, the NEMO basic support protocol does not yet include any mechanism for RO. The use of MIPv6-type RO without any enhancement is inefficient because a different care-of address (CoA) process is used for mobile nodes in NEMO. In MIPv6, a mobile node's CoA gives the actual location of the mobile node and packets are forwarded directly to that CoA following RO. In contrast to MIPv6, the CoA of a mobile network node (MNN) in NEMO does not provide the actual location of the MNN as the CoA belongs to the mobile network prefix (MNP) of the mobile network, which is also moving.

Various approaches to RO in NEMO have been proposed in the literature [24]-[30]. However, most of them require additional network entities or functionalities to be implemented outside the mobile network domain, which may hinder their deployment in the existing Internet. Moreover, these approaches provide only partial solutions by optimizing only some sectors of the route between correspondent nodes and MNNs.

To overcome these problems, we propose a mobile router-assisted route optimization (MoRaRo) scheme. With MoRaRo, an MNN performs RO with the correspondent node only once, at the beginning of a communication session. After that, when the network moves, the root-MR performs RO on behalf of all active MNNs, so that network movement remains transparent to the sub-MRs and MNNs. MoRaRo is simple to implement as it requires only slight modification of the implementation of the NEMO basic support protocol at the local components, in particular the MRs and MNNs of a mobile network, leaving network entities in the core or in other administrative domains of the MIPv6 or NEMO basic support protocols, such as CNs and home agents, untouched. This scheme uses the well-established functionality of MIPv6 to enable a

CN to route data packets directly to a mobile network without any encapsulation (or angular routing). It thus reduces delays in packet delivery and encapsulation overheads in the core networks.

To evaluate the proposed MoRaRo scheme, we developed analytical models for estimating packet delivery overheads, efficiency, and delay. We also performed an analysis and simulation to assess performance during handover in terms of the correspondent binding update latency and overheads.

This chapter is organized as follows. In Section 3.2, we summarize previous work relating to route optimization in NEMO. The MoRaRo scheme is described in detail in Section 3.3. We present an evaluation of the scheme's performance in Section 3.4, analysis and simulation results in Section 3.5, and a conclusion in Section 3.6.

3.2   Related Work

RO is a mechanism that not only shortens the data delivery path between a mobile node and a CN, but also reduces the potential level of encapsulation. Nevertheless, RO requires some route update signaling and/or additional information in the IP headers of data packets to enable packets to follow the optimal path and reach their destination intact. The generic consideration in designing an RO scheme is to use a minimum of signaling and/or additional information in the packet header.

Since we are proposing an extended MIPv6 type RO for NEMO, we first briefly describe how MIPv6 RO operates, and then review other solutions available in the literature for NEMO.

3.2.1   Route Optimization in Mobile IPv6

MIPv6 [4] has defined mobile node (MN)-initiated RO mechanism, which is triggered when the MN receives the first data packet or Binding Refresh Request from a CN tunneling via a home agent. The RO mechanism consists of two tasks: a Return Routability (RR) address test and a Binding Update (BU). The MN performs an RR address test to provide the CN with reasonable assurance that the MN is in fact addressable at both its claimed CoA and HoA. To achieve this, the MN sends a Home Test Init message from its HoA (via the home agent) and Care-of Test Init messages from its CoA to the CN. Then, the CN responds with Home Test and Care-of Test messages, containing authorization data (such as home/care-of init cookies, keygen, nonce indices), via the MNN's HoA and CoA, respectively. The MN uses these authorization data, while sending a binding update to the CN. Successful reception of the binding update by the CN completes the route optimization process and the CN can then send packets directly to the

CoA of the mobile node, shunting the home agent. Completion of route optimization thus takes about 1.5 round-trip times between the MN and CN.

Using MIPv6 route optimization without any enhancement in NEMO creates some problems. Firstly, as the CoA of an MNN in NEMO is associated with the home link of the MR, it does not provide the actual geographical location of the MNN. Therefore, the MNN's CoA cannot be used as a destination address in an IP data packet for the route optimization purpose. Secondly, when the MR's care-of address is used by an MNN to perform RO, a binding update implosion could occur as all MNNs need to update their bindings at the home agents and CNs simultaneously when an MR changes its CoA. To resolve these problems, we propose a mobile router-assisted route optimization scheme for NEMO support.

3.2.2 Other Solutions

Some solutions to the problems of RO in NEMO have been published [24]-[30]. Most of them essentially require the introduction of new components or mechanisms in the existing networking system. They include generic overviews of various types of route optimization possibilities in NEMO [19, 24]. Thubert et al. proposed the use of a new routing header, Routing Header (RH) type 4, also called a reverse routing header (RRH), for MNN-originated outbound packets, and a modified RH type 2 for inbound packets destined for MNNs [25]. The RH type 4 collects the CoAs of all nested MRs, which are later included in the modified RH type 2 to reduce the number of nested encapsulations for inbound packets. This scheme, however, optimizes the path between the home agent and the MR serving the MNN, not between the CN and MR. Moreover, it requires MRs to modify packet headers, which would increase computational overheads.

Na et al. [26] proposed that the MRs' HAs should assist RO between a correspondent router (CR), a router serving the CN, and the MR. However, this scheme has some disadvantages. It requires preservation of a soft-state at the CR, in addition to that at the MR, and it does not describe how to inform the CR promptly of changes in the MR's CoA. Route updating could be done faster if an MR instead of a home agent initiated the RO process when the network moved. Our scheme uses an MR-initiated RO approach.

Ohnishi et al. [27] described RO problems relating to nested mobility. This scheme uses hierarchical configuration of MRs and enables the MNN's HA to send packets directly to the root-MR's CoA, skipping other MRs' HAs. By optimizing the route between the MNN's HA and the root-MR, but not between the CN and root-MR, this scheme again provides only a partial solution to route optimization problems.

Perera et al. [28] proposed assigning the MR a prefix pertaining to the visited network and
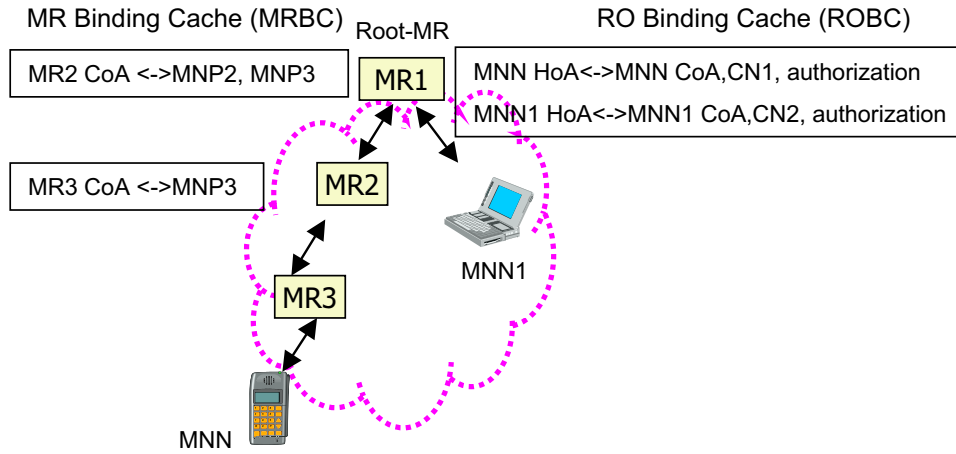
Figure 3.1: MR-Binding Cache (MRBC) and RO-Binding Cache (ROBC) used for route optimization in NEMO support network. MNP2 and MNP3 are mobile network prefixes of network containing MR2 and MR3, respectively. MNN is assumed to be a visiting mobile node.

advertising the prefix in a router advertisement to the MNNs. MIPv6-enabled MNNs first configure their CoAs from the prefix using a stateless address auto-configuration, and then perform RO by sending binding updates to their HAs and CNs. As this scheme requires MNNs to configure their CoAs every time the network moves, it could possibly cause a binding implosion problem [19]. In contrast to this scheme, our MoRaRo scheme requires the MNN to configure and register a CoA with the HA only once when it first enters the mobile network.

Ernst et al. [29] proposed a technique in which the mobile network registers a multicast address on the domain name server (DNS), and uses the address to notify the prefix scope binding update when the network moves. CNs are required to subscribe to the multicast address to learn the up-to-date location of the mobile network. Similarly, Wakikawa et al. [30] proposed the use of optimized route cache (ORC) management protocols in some interior gateway routers.

However, since these proposals advocate the use of additional components or protocols distributed over the Internet, they would increase the complexity of NEMO deployment in the existing networks and standards.

3.3   Proposed Route Optimization Scheme: MoRaRo

The proposed MoRaRo scheme requires implementation of RO related mechanisms only at a mobile network's local entities, i.e., the MR and MNNs. We describe the operations of the MoRaRo by considering a nested mobile network that has multiple levels of MRs as shown in Fig. 3.1. In this scheme, all MRs keep a binding cache, which we call the *MR-binding cache*

(MRBC) for all the nested mobile routers behind them. Additionally, the root-MR (i.e. the top-level MR) keeps another binding, which we call the *RO-binding cache (ROBC)* for all active MNNs that have ongoing communication sessions with CNs.

The MRBC is used to store bindings between the CoAs of nested MRs and their MNPs. As illustrated in Fig. 3.1, when the sub-mobile router MR3 is attached to a parent mobile router MR2, the former registers its CoA and mobile network prefix (MNP3) with the latter. MR2 uses this information to forward a packet addressed to the MNN's CoA, which is associated with MNP3. Similarly, MR2 registers its CoA and all the mobile network prefixes (MNP2 and MNP3) that are accessible through it with MR1 (the root-MR) so that MR1 can forward data packets to MNNs that have CoAs associated with MNP2 and MNP3. Every entry in the MRBC has a lifetime, which must be renewed by re-registering if the sub-MR is going to stay in the mobile network longer than the time allocated in the previous registration. Moreover, the root-MR keeps updating the sub-MRs and active MNNs of the root-MR's CoA and HoA. As NEMO is an extension of the MIPv6 protocol, it is logical to set the value of the MRBC entry's lifetime to MAX_RR_BINDING_LIFETIME (i.e., 420 seconds) of MIPv6. Both the very long and very short values of the lifetime have disadvantages. If the lifetime is very long, the cache may contain stale and useless binding information. On the other hand, a short lifetime will require the sub-MR to refresh binding very frequently, thus increasing signaling overheads.

The ROBC is maintained by the root-MR for two purposes: (1) to obtain the CoA of active MNNs to enable data packets to be tunneled to them, and (2) to perform a BU with the CNs on behalf of active MNNs when the network moves. The ROBC includes bindings between the MNN's HoA and other information such as the CoA, CN's address, and authorization data. The information required for creating the binding is supplied by the MNNs in *RO-inform message* (explained below) when they first perform RO with the CNs. To keep the ROBC up-to-date, every entry has a lifetime, which is renewed by packets traveling between the MNN and its communication peer, i.e., a CN. If there is no packet traversal between the peers for some predefined lifetime, the entry is deleted. This lifetime can be less than or equal to the lifetime of the MRBC entries.

Caches like MRBC and ROBC are very frequently used for packet routing in mobile networking systems. Two well established protocols, Cellular IP [37] and HAWAII [38], use caches to store routing and paging information of mobile nodes.

3.3.1   Route Optimization for Mobile IPv6 Nodes

First, we describe the RO operation of a MIPv6-enabled visiting MNN, which can easily be extended to non-MIPv6 supporting local MNNs (described in section 3.4). The RO process is
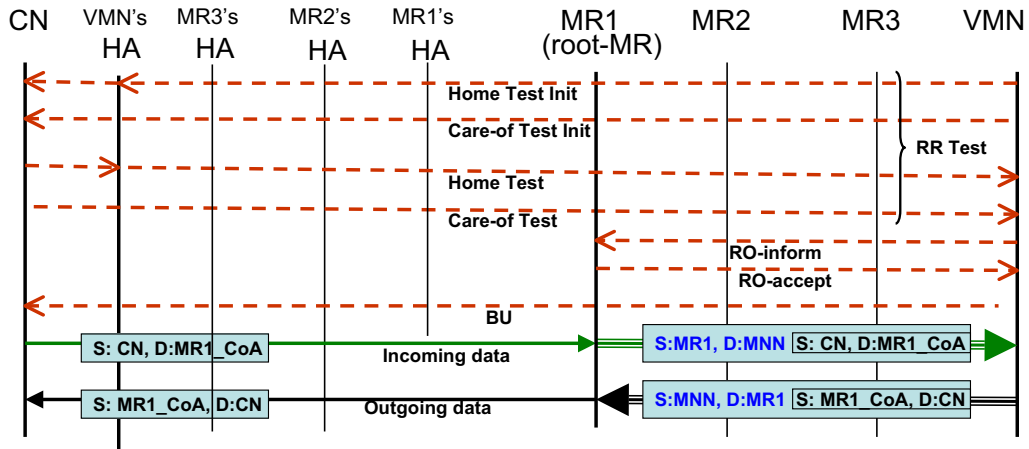
Figure 3.2: Packet flow sequence in proposed MoRaRo scheme. Dotted arrows show signaling paths and solid arrows show data paths. 'S' and 'D' in boxes represent packet header source and destination address fields, respectively. Both data and signaling messages originated from or destined for the VMN traverse all nested MRs.

triggered when an encapsulated data packet destined for an MNN arrives at the mobile router MR3 via the MR3's home agent. MR3 decapsulates the packet and forwards it to the MNN. Then, as in MIPv6 RO, the MNN performs a return routability (RR) test with the CN by exchanging messages as shown in Fig. 3.2. After the test, the MNN forms a new control message, RO-inform, containing the MNN's HoA, CoA, CN's address, authorization data, etc., and sends the message to the root-MR. The root-MR uses RO-inform message to create an entry for the MNN in the ROBC and then replies with an RO-accept to the MNN.

The MNN then formulates a binding update (BU) message using the root-MR's CoA in the alternate care-of address option, and sends the BU to the CN. On arrival of the BU, the CN creates a binding between the MNN's HoA and the CoA mentioned in the alternate care-of address option field. The CN then sends packets directly to the root-MR's CoA using Routing Header (RH) type 2. RH type 2 is an extension header defined in Mobile IPv6 [4]. When the MNN leaves the mobile network managed by the MR, it acquires a new CoA and performs the RO in accordance with the MIPv6 protocol.

3.3.2   Packet Processing in Mobile Network

The root-MR checks the home-address option field of the RH type 2 header of an inbound packet to get the HoA of the MNN that the packet is addressed to. The MNN's HoA is used to search for the corresponding CoA in the ROBC, which in turn is used to tunnel the packet to the MNN.

Similarly, MNN-originated outbound packets have the root-MR's CoA and the address of the CN in the source and destination address fields, respectively. These packets are tunneled to the root-MR using the root-MR's HoA as the destination and the MNN's CoA as the source address in the outer IP header. The root-MR decapsulates and forwards the packet normally to the CN, as depicted in Fig. 3.2.

3.3.3   Mobile Network Handover

When the root-MR performs a handover to a new link and configures a new CoA, it sends a BU to all CNs that have active sessions with the MNNs. As the root-MR has already stored the addresses of the CNs and binding authorization data for all active MNNs in the ROBC, it uses this information to send BU messages on behalf of the MNNs. That is, no RR address test is performed on handover. The root-MR also notifies all MRs and active MNNs of the CoA change so that they can use the new CoA as the source address in outbound data packets. Note that in our scheme, the network handover does not trigger any update in home registration except in the HA of the root-MR. That means our scheme avoids increasing signaling volume due to handover management.

3.3.4   Route Optimization for non-Mobile IPv6 Nodes

The MR operates as a proxy for performing RO for non-MIPv6 nodes. The non-MIPv6 nodes are those nodes that are local to the mobile network; that is to say, they have their IP addresses from the MNP and can access the Internet only through the MR. In such cases, since the MR's home agent tunnels data packets destined for any node that has an IP address associated with the MNP, the MR receives the tunneled packets addressed to the non-MIPv6 node. The MR then decapsulates and forwards the packets to the node, and performs RO tasks on behalf of the node using the MR's own CoA as the node's CoA. The operation of the MR in the RO is similar to the operation of the MIPv6-enabled MNN, as explained above.

3.3.5   Security and Privacy Considerations

We believe the security issues related to the proposed MoRaRo can be resolved with the security mechanisms available for the MIPv6 protocol and its derivatives such as the NEMO basic support protocol and Hierarchical MIPv6 (HMIPv6) protocol [5]. As both the MR and its HA belong to the same administrative domain, they use their pre-established security association (SA). The MNN and MR exchange their mutually trusted identities to establish the SA. A trusted identity can be an IP address or a certificate signed by a Certificate Authority (CA) that

both the MR and MNN trust. Similarly, the MNN and CN are mutually trusted through the
return routability test. Furthermore, as in MIPv6, all signaling messages exchanged among the
MNN, MR, HA and CN are authenticated by IPSec [31].

Performing route optimization reveals the location of the MNN to the CN. The location
information is considered as personal information, the revealing of which may create a privacy
exposure problem. However, the MNN can use privacy rules to protect its location information
against possible misuse. The privacy rules regulate the CN's activities regarding the collection,
use, disclosure, and retention of location information of the MNN. In our MoRaRo scheme, the
MNN can perform privacy negotiation with the CN by using the privacy protection frameworks
being developed by the Geographic Location/Privacy (geopriv) working group of the IETF [32].

Although the use of privacy rules protects the location information against misuse, it cannot
conceal the location information from the CN. To conceal the location information of the MNN,
the following two techniques could be helpful.

(a) We can arrange the ARs in hierarchical order and use HMIPv6 to get the address of the
   top level AR (i.e, MAP: mobility anchor point) as the CoA of the root-MR. In such case,
   the use of the root-MR's CoA for RO would not reveal the precise location of the MNN,
   as the ARs in hierarchical order might be geographically spread over a large area.

(b) Alternatively, we can modify the operation of the proposed MoRaRo scheme to use the
   home address (instead of CoA) of the root-MR for the route optimization of those MNNs
   that demand a high degree of location privacy. This technique can prevent the location of
   the MNN from being revealed at the cost of sacrificing the most optimal data path. It can
   optimize the data delivery path partly by allowing packets to travel directly from the CN
   to the root-MR's home agent, without going through the home agents of nested MRs or
   MNNs.

For a more general overview of security and privacy in ubiquitous computing environments
we refer the reader to [33].

3.3.6 Advantages and Disadvantages

The MoRaRo scheme has a number of advantages over other schemes.

a. Shortest path selected: In MoRaRo, data packets are routed directly from the CN to the
   root-MR. Unlike other schemes [25, 27], our scheme does not require HAs to intercept
   packets after RO as no packets go to HAs.

b. Little modification of existing system required: Implementing MoRaRo requires modifying the NEMO basic support implementation only at the MR and MNNs. The NEMO-enabled home agents and MIPv6-enabled CNs are left untouched.

c. End-to-end principle preserved: There is no modification of IP packets, even headers, between the CN and MNN. There is also no tunneling in the core network. Moreover, to preserve the end-to-end semantics of IP packet transportation it requires only one tunneling between the root MR and the MNN irrespective of the nested mobility levels.

d. No additional entity implemented in core network: MoRaRo does not require any additional components or mechanisms to be implemented in foreign network domains or core networks. As the RO functions are completely confined to the mobile network, our scheme is relatively easy to deploy and, in addition, it does not require any coordination with other parties or network owners.

e. No need for VMN to perform location update to HA when network moves: As the VMN retains the same CoA address as long as it stays in the mobile network, there is no need to perform a location update with its home agent when the mobile network moves.

f. Faster reaction to handover: The root-MR can inform CNs of CoA changes faster than any other network entities can. Accordingly, as the root-MR performs a BU with the CNs, the optimal route can be re-established shortly after a handover, avoiding or at least minimizing disruption of the network service.

The MoRaRo scheme does have some potential disadvantages.

a. Location of MNN exposed: When the MNN performs a BU using the root-MR's CoA, its location is exposed to the CN. This problem can be solved as explained in section 3.5.

b. Root-MR is a single point of failure: The root-MR executes two different functions: it acts as a redirection point for all packets flowing in or out of the mobile network, and it performs BUs with CNs on behalf of active MNNs when its CoA is changed. The amount of BU-related workload on the root-MR depends on the following parameters: (1) ratio of active MNNs to total MNNs associated with mobile network, (2) number of CNs for each active MNN, and (3) the rate of mobility of the mobile network. We found, through our evaluation, that given reasonable ranges for these parameters, the improvement in system performance due to lessening packet delivery overheads exceeded the burden imposed by BU signaling. In addition, the performance of a root-MR can be improved by using multi-homing, which enables the root-MR to access the Internet through different interfaces

[21]. Availability of multi-homing reduces the BU signaling volume over an interface of the MR.

3.4   Performance Evaluation

In this section, we describe the development of analytical models to evaluate the performance of the MoRaRo scheme in terms of packet delivery (from the CN to MNN) overheads, efficiency, and delay. We chose these metrics because they are directly related to the objectives of route optimization, i.e., to shorten the data delivery path and hence improve the performance. Although we evaluated the overhead and delay for inbound packets destined for an MNN, it could easily be extended to incorporate MNN-originated outbound packets as outbound packets follow the same paths (and undergo equal levels of encapsulation) as inbound packets.

We also analyzed the performance during handover, especially in terms of delays and overheads for correspondent binding updates. While performing the evaluation, we compared the MoRaRo scheme with the path control header (PCH) [26] and reverse routing header (RRH) [25] schemes. First, we give a brief overview of these schemes.

In the PCH scheme, when an MR's home agent receives a packet reversely tunneled from the MR, it piggybacks a hop-by-hop option header, called a path control header (PCH), on to the packet, and forwards the piggybacked packet to the CN. When the packet reaches a CR (correspondent router), the PCH contains the care-of addresses of all the MRs along the reversely tunneled path. Then, the CR establishes an RO-tunnel between itself and the MR using a three-way handshake process. Afterwards, any packet destined for the MNN is routed through the CR-MR tunnel using routing header (RH) type 0 which contains care-of addresses of all the MRs along the way to the MNN.

In the RRH scheme, when an MR receives a reversely tunneled outbound packet, it replaces the source address of the outer header with its CoA and appends the previous source address to RH type 4 slots. After passing through all nested MRs, the slots contain the CoAs of all the MRs. When it receives the packet, the home agent of the MR serving the MNN creates a mapping between the MNN's HoA and the CoAs of all the MRs. These CoAs are used in a modified RH type 2 inserted in an encapsulation header for redirecting packets to the MNN.

We chose these two schemes because we wanted to compare the MoRaRo with schemes that use a similar and a different approach. The PCH uses both signaling and header information for RO, which is similar to our approach in MoRaRo. In contrast, the RRH uses only header information for RO, which differs completely from our approach.

For the purposes of performance analysis, we define a *mobility support node* (MSN) as a network entity, such as an MNN, CN, MR, or HA, which participates in processing data packets
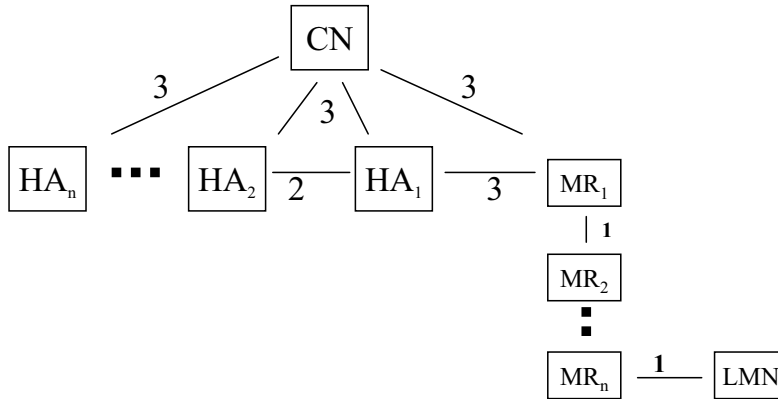
Figure 3.3: Network configuration for performance evaluation. Numbers represent distances in terms of IP hops.

Table 3.1: Number of MSNs (M_n) involved in n level of nested mobility and packet delivery overhead types in different schemes.

Scheme	M_n	Overhead type
Basic	$2n + 2$	Encapsulation
PCH	$n + 3$	Encapsulation + RH type 0 + BU
RRH	$n + 3$	Encapsulation + RH type 2(modified)
MoRaRo	3	RH type 2 + Encapsulation + BU

to a greater extent than just normal routing. Processing may include packet encapsulation, decapsulation, or inserting, deleting or changing routing headers using information maintained as soft-state at an MSN or contained in the packet header. For a nested mobile network with n level of MRs, as depicted in Fig. 3.3, the numbers of MSNs (M_n) for different schemes are shown in the second column of Table 3.1.

3.4.1   Analysis of Packet Delivery Overhead

Packet delivery overhead is the additional cumulative traffic load on the network induced by mobility management functions [34]. It includes the following two components.

1) Header overhead – caused by additional information in packet headers, such as encapsulation, RH type 0, and RH type 2, required for routing packets from source to destination. That is, the header overhead does not include the basic IPv6 header, but includes extension or optional headers.

2) Signaling overhead – caused by BU signaling required for performing route optimization with a CN (or CR).

Of the four schemes shown in Table 3.1, the basic support scheme and RRH have only header overhead, whereas the others have both header and signaling overheads.

We define header overhead as the product of the additional header size and the distance it travels in the network. Header overhead is measured in units of 'octet×hop'. The per packet header overhead of an RO scheme s in a mobile network with n levels of nested mobility is given as

$$H_o(s, n) = \sum_R P_i^+ \times D_i, \tag{3.1}$$

where R is the ordered set of MSNs, excluding the MNN, along the data delivery path from a CN to MNN. That is, $R = \{MSN_i, i = 0 \text{ to } m\}$, where MSN_0 is the CN and MSN_m is the MSN located just before the MNN. The value of m, which depends on both s and n, is equal to $M_n - 1$, (M_n is shown in Table 3.1). Similarly, D_i is the distance in terms of IP hops (routers) between MSN_i and MSN_{i+1}, and P_i^+ is the additional header size (excluding the basic IPv6 header) that is transported over the distance D_i. Based on the description of the schemes, P_i^+ for different schemes can be expressed as shown in Table 3.2. In this table, h and a are the IPv6 header size (i.e., 40 bytes) and IPv6 address size (i.e., 16 bytes), respectively. Content inside pairs of square brackets represents an extension header size. n represents the number of nested mobility levels.

Table 3.2: Additional header size (P_i^+) in different schemes.

Schemes	P_i^+	Conditions
Basic	$h \times i$	$0 < i \leq n$
	$h \times (2n - i)$	$n < i < 2n$
PCH	$h + [8 + a \times (n - 1)]$	$i = 1$
	$8 + a \times (n - 1)$	$2 \leq i \leq n$
RRH	$h + [8 + (a \times n)]$	$i = 1$
	$[8 + a \times n]$	$2 \leq i \leq n$
MoRaRo	$[8 + a]$	$i = 0$
	$h + [8 + a]$	$1 \leq i < n$

Next we define the signaling overhead. The signaling messages required for RO are carried in the mobility header (MH), which is an extension header defined in Mobile IPv6 [4]. Based on the specification, the sizes of mobility headers containing different messages are listed in Table 3.3. With these messages, we use a similar definition as used for header overhead to

Table 3.3: Route optimization messages and their size in octets (IPv6 header is not included).

Type	Used in	Size
Binding Update	MoRaRo, PCH	80
Binding Ack.	MoRaRo, PCH	40
Binding Request	PCH	16
Home Test Init	MoRaRo	16
Care-of Test Init	MoRaRo	16
Home Test	MoRaRo	24
Care-of Test	MoRaRo	24
RO-inform	MoRaRo	80
RO-accept	MoRaRo	40

compute the signaling overhead $S_o(s, n)$, which is shown below.

$$S_o(s, n) = \sum_Q \sum_R SP_i \times D_i, \tag{3.2}$$

where SP_i is the size of a signaling packet including one or more IPv6 headers that travel a distance D_i, and Q is the set of messages required to complete the RO process. Like the header overhead, the signaling overhead is also measured in units of 'octet×hop'. Note that in the case of only header overhead, the additional header in an IP data packet (not the basic IPv6 header and payload) is an overhead-yielding element. However, in signaling overhead, the whole signaling packet (message plus one or more IP headers) is an overhead-yielding element.

Header overhead is present in every data packet, but signaling overhead is incurred only once when RO is performed. Therefore, the average signaling overhead per packet can be computed by dividing the signaling overhead by the number of data packets transmitted over the optimized path that was established through signaling. Let $E(D)$ be the average number of packets of a data session forwarded over the optimized path. Then, the average (total) overhead per packet, $T_o(s, n)$, is as given below.

$$T_o(s, n) = H_o(s, n) + \frac{S_o(s, n)}{E(D)}. \tag{3.3}$$

As mentioned earlier, the objective of deploying an RO scheme is to reduce packet delivery overheads. In other words, the NEMO basic support protocol has large overheads that have to be reduced by an RO scheme. To check how effective an RO scheme is in terms of reducing packet delivery overheads, we divide the average overheads of the schemes under consideration

by that of the basic support protocol and get an overhead ratio $R_o(s, n)$, as given by Eq. (3.4).

$$R_o(s, n) = \frac{T_o(s, n)}{T_o(basic, n)} \qquad (3.4)$$

Note that as we expect an RO scheme to produce a lower packet delivery overhead than the basic support protocol, the overhead ratio should be less than 1. This ratio can be used to compute the percent reduction of the packet delivery overhead due to deploying an RO scheme as: $(1 - R_o(s, n)) \times 100$. Therefore, the smaller the ratio, the better the performance of the RO scheme.

Next we derive the packet delivery efficiency, which measures how effectively the network resources are utilized for delivering the IP payload. The efficiency is computed by dividing the total traffic load when an IP packet is routed normally (without using mobility support) from a source to a destination by the traffic load when the same packet is routed with mobility support from the same source to the same destination. That is, when a packet is routed normally, there is no associated overhead as long as the packet is routed through the shortest path as determined by routing algorithms.

Let P be the size of an IP packet (that includes a basic IPv6 header and payload) and $D_{TL}(n)$ be the traffic load due to normal routing of the packet from a source to a destination. Then, $D_{TL}(n)$ is simply,

$$D_{TL}(n) = P \times d_s, \qquad (3.5)$$

where d_s is the shortest distance between the source and destination. However, when the same packet is transported using mobility support, the traffic load is larger because of possible sup-optimal routing and overheads. Let $M_{TL}(s, n)$ be the traffic load in a mobility support network. Then, $M_{TL}(s, n)$ has three components: (a) traffic load due to an IP packet of size P routed through MSNs, (b) header overhead, and (c) signaling overhead. After combining the header and signaling overheads together, we get $M_{TL}(s, n)$ as given below:

$$M_{TL}(s, n) = \sum_R (P \times D_i) + T_o(s, n). \qquad (3.6)$$

Note that the dimensions of both $D_{TL}(n)$ and $M_{TL}(s, n)$ are denoted by 'octet×hop'.

Finally, we compute the packet delivery efficiency $E(s, n)$ as follows:

$$E(s, n) = \frac{D_{TL}(n)}{M_{TL}(s, n)}. \qquad (3.7)$$

The packet delivery efficiency of a scheme is less than or equal to 1. The higher the efficiency, the better the scheme's performance.

3.4.2   Analysis of Packet Delivery Delays

Packet delivery delay (also known as routing delay) refers to the length of time required to move a packet from a source (CN) to a destination (MNN) through the Internet. The delay depends on several factors, such as the bandwidth of intermediate links, network congestion in links, and the physical distance to be traveled. Furthermore, mobility support incurs more delay due to introducing additional functionality at the MSNs. In this evaluation, for simplicity, we assume that all wired and wireless links have the same bandwidth of B_w and B_{wl}, respectively. The network congestion level is considered in terms of the load on the MSNs (or MSN utilization), which refers to the degree to which the MSN is busy, and is calculated as the ratio of the rate of packet arrival to packet processing. The physical distance between the source and destination is considered in terms of the latency of wired and wireless links, say L_w and L_{wl}, respectively. Let t_r be the routing table look up and processing delay in every hop (router), and t_p be the additional processing (binding table lookup and encapsulation and/or header modification) delay at an MSN. Let $t(P_i)$ denote the delivery delay of a packet of size P_i ($P_i = P + P_i^+$) from MSN_i to MSN_{i+1}. Then $t(P_i)$ can be expressed as follows.

$$t(P_i) = t_p + [t_r + \frac{P_i}{B_w} + L_w] \times D_i, \tag{3.8}$$

where D_i, the distance between MSN_i and MSN_{i+1}, is assumed to have wired links. We substitute L_w (B_w) in Eq. (8) with L_{wl} (B_{wl}) for wireless links.

Since we want to evaluate the effect of the functionality specific to mobility management on packet delivery delay, we assume that t_p is a variable that is dependent on MSN utilization, whereas, t_r is fixed and is the same for all routers. We also assume that the packet inter-arrival time and processing time in the MSN have exponential distributions with parameters λ and μ, respectively [35]. Then, processing in the MSN becomes an M/M/1 queuing system [36] and t_p is the mean waiting time in the system. We therefore have

$$t_p = \frac{\rho}{\lambda(1 - \rho)}, \tag{3.9}$$

where ρ ($=\lambda/\mu$) is the load or utilization on the MSN. Then, the overall packet delivery delay from a CN to an MNN in scheme s with n levels of nesting is given as:

$$T(s, n) = \sum_{wired} t(P_i) + \sum_{wireless} t(P_i). \tag{3.10}$$

3.4.3   Analysis of Handover Delays and Overheads

In the NEMO basic support protocol, the root-MR sends a BU containing a new CoA to its home agent via the new AR during handover. After that, the home agent sends packets to the new CoA. With route optimization schemes, however, the root-MR may receive data packets not only via its home agent but also via other tunnel entry-points such as the correspondent router in the PCH, the HA of a bottom-level MR serving the MNN in the RRH, and the CN in the proposed MoRaRo scheme. Therefore, these points also need to be notified by performing a correspondent binding update (CBU).

In addition, the Mobile IP handover is hard, i.e., the connection to the old AR is broken, and a connection to the new AR is made (also known as a break-before-make type handover). Therefore, the old AR has to buffer and redirect in-flight packets that are tunneled by the tunnel entry-point from the moment when the connection of the root-MR to the old AR is broken to the moment when the tunnel entry-point receives the CBU from the root-MR via the new AR. The duration (or latency) of the handover can be broken down into three parts: detection, configuration, and completion time. The detection time refers to the time that elapses while the MR detects that it has moved to a new IP subnet by listening to IP-layer beacons. The configuration time refers to the time taken by the MR to configure a CoA in the new subnet. Finally, the completion time refers to the time required by the MR to inform the tunnel entry-point of its CoA change by performing a CBU. Considering that the former two times are the same for all schemes (and depend on specifications in the Mobile IPv6 protocol), we focus only on the third part of handover latency, i.e., CBU latency. If this is large, the old-AR will be required to buffer a huge volume of data packets, which may result in packet loss. Therefore, a good RO solution should keep this latency as small as possible.

The CBU latency is equal to the one-way delay of a BU packet traveling from the root-MR to the corresponding tunnel entry-point. Let $T_{X\_Y\_Z}$ denote the one-way delay of a BU packet from MSN X to MSN Y via MSN Z. Then the CBU latencies of different schemes are as follows:

$$T(basic) = T_{rootMR\_HA} \text{ (via no intermediate MSN)}$$

$$T(PCH) = T_{rootMR\_CR\_allHAs}$$

$$T(RRH) = T_{rootMR\_HA_1} \text{ (via no intermediate MSN; position of } HA_1 \text{ is as shown in Fig. 3.3)}$$

$$T(MoRaRo) = T_{rootMR\_CN} \text{ (via no intermediate MSN)}$$

It is logical to assume that the PCH and RRH schemes use TCP acknowledgement packets to carry new CoA information for CBU purposes (since we are evaluating a case of CN-originated communication). In this case, the sizes of the CBU packets in all the above schemes are roughly

the same. Therefore, we can infer that apart from the PCH, all other schemes have similar CBU latency assuming that the distances between any two MSNs in the Internet are approximately the same.

However, the low CBU latency in our scheme is obtained by requiring the root-MR to perform a CBU on behalf of all the active mobile nodes. This results in an additional signaling overhead, which is analyzed next. We introduce the following new parameters:

t_{sr} – MR's average subnet/cell resident time;

t_{pia} – packet inter-arrival time of MNN-originated outbound packets;.

t_{sd} – average session duration;

N_p – average number of outbound packets from an MNN during a session (i.e., $N_p = t_{sd}/t_{pia}$).

Let N_h denote the average number of times the MR performs handovers during a session, which is expressed as $N_h = \lceil t_{sd}/t_{sr} \rceil$. Then the percentage of CBU packets with respect to the total data packets departing from the MNN via the MR, P_{cbu}, is given as:

$$P_{cbu} = \frac{N_h}{N_p} \times 100. \tag{3.11}$$

We also carried out a performance simulation to evaluate P_{cbu}, which is described in the next section.

3.5   Analysis and Simulation Results

3.5.1   Packet Delivery Overhead and Efficiency

For the purpose of performance comparison, we used the distances between any two MSNs as specified in Fig. 3.3. Using that configuration we calculated the header overheads in 'octet×hop'

Table 3.4: Header overheads H(s,n) in 'octet×hop'.

Nesting level (n)	*Basic*	*PCH*	*RRH*	*MoRaRo*
1	120	144	192	72
2	360	216	280	136
3	720	320	400	200
4	1200	456	552	264

units for different nesting levels (*n*) and listed them in Table 3.4. The results show that the header overhead of the MoRaRo scheme is the lowest of all the schemes. They also indicate that the header overhead increases with an increasing level of nested mobility. However, in this situation, our MoRaRo scheme also has the lowest rate of increase. The header overhead in all the RO schemes is lower than that in the basic support protocol, except when $n = 1$. In this case, as the PCH and RRH schemes use RH type 0 and RH type 2 extension headers, respectively, in the encapsulation header, their overheads are larger than that of the basic support protocol, which uses only one encapsulating IPv6 header.

Figures 3.4(a) and (b) show the results of the analysis of overhead ratios for $n = 2$ and $n = 4$, with respect to the average number of packets transported, $E(D)$, over an optimal path. It shows that the MoRaRo scheme performed best when there were a large number of packets. MoRaRo has relatively larger overheads when only a few packets are transported via the RO path. This is because it uses a MIPv6-type RO process consisting of a number of signaling messages. However, as the number of packets increases, the overhead decreases abruptly, becoming the lowest of all the schemes. It follows that for a smaller data transfer session, use of RO is undesirable.

If we compare Figs. 3.4(a) and (b), we can see that the packet delivery overhead is largely reduced by the use of an RO scheme when there are more levels of nested mobility. For instance, MoRaRo reduces the per packet overhead by 60 and 78% for $n = 2$ and $n = 4$, respectively, when the number of transported packets is 500.

To see how dependent the results were on the distance between two MSNs, we evaluated the overhead ratios for different distances, as shown in Fig. 3.5. For simplicity we assumed that the distances between any two adjacent MSNs in the wired network section were the same. Figure 3.5(a) shows that MoRaRo outperformed all other schemes with 2 levels of nested mobility at a distance of more than one hop. Moreover, for a nested mobility level of 4, as shown in Fig. 3.5(b), our scheme always performed best in terms of the packet delivery overhead. That is, the advantages of our scheme are greater with longer distances. Although we have shown the curves of overhead ratios for $E(D) = 100$, we found that the tendency of MoRaRo performance was similar for other values, except when $E(D)$ was very low (fewer than 50 packets). For very low values of E(D), as the overheads were larger, MoRaRo required longer distances between MSNs to outperform the other schemes.

Figure 3.6 compares the packet delivery efficiency of different schemes against packet size for $n = 2$. In terms of efficiency, the MoRaRo scheme performed best. For smaller packets such as TCP acknowledgments, packet delivery is less efficient because mobility-specific headers consume a comparatively larger share of network resources. The comparison shows that the efficiency of the basic support protocol is very low (about 40%) for any packet size. On the
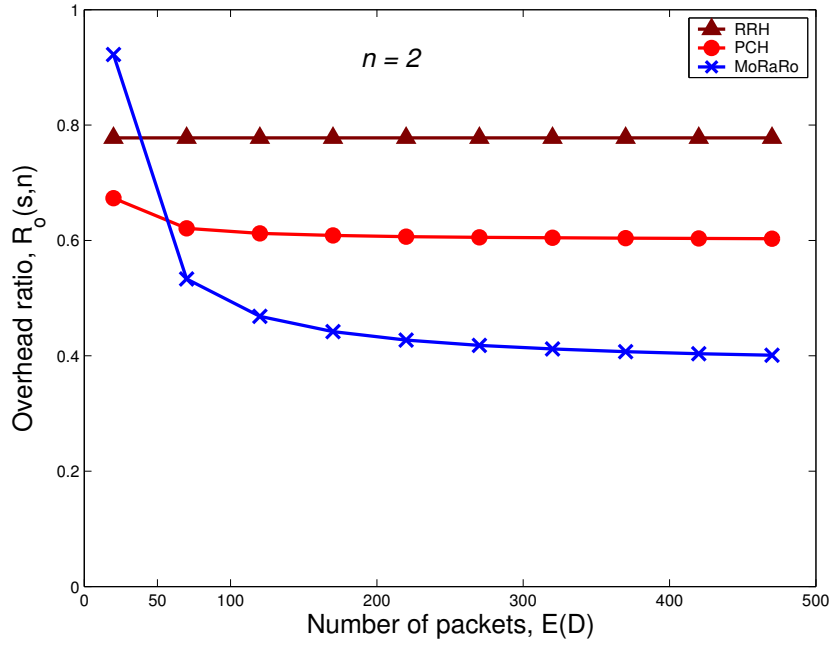
(a)
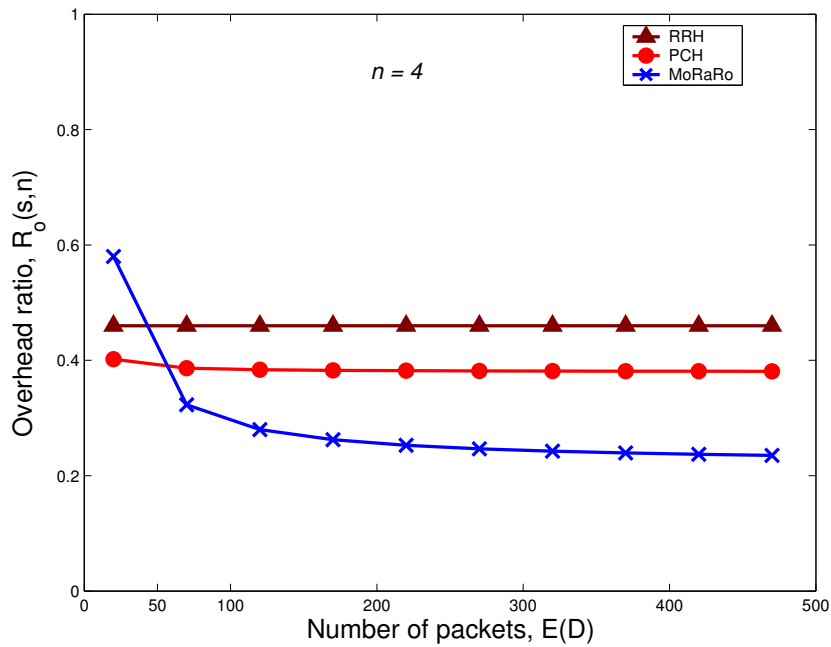
(b)

Figure 3.4: Overhead ratio versus number of transported packets through the optimized path, for nested mobility levels of 2 (top) and 4 (bottom).
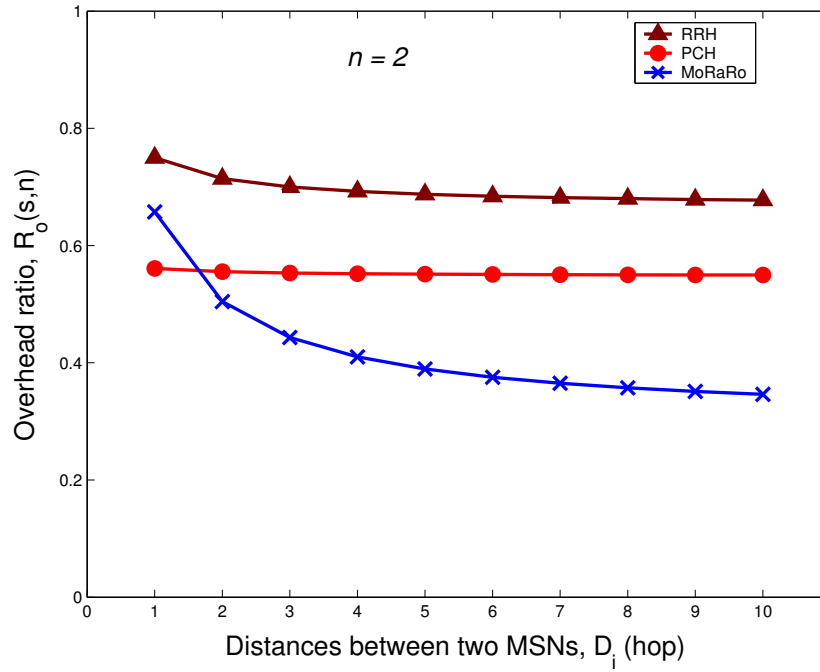
(a)
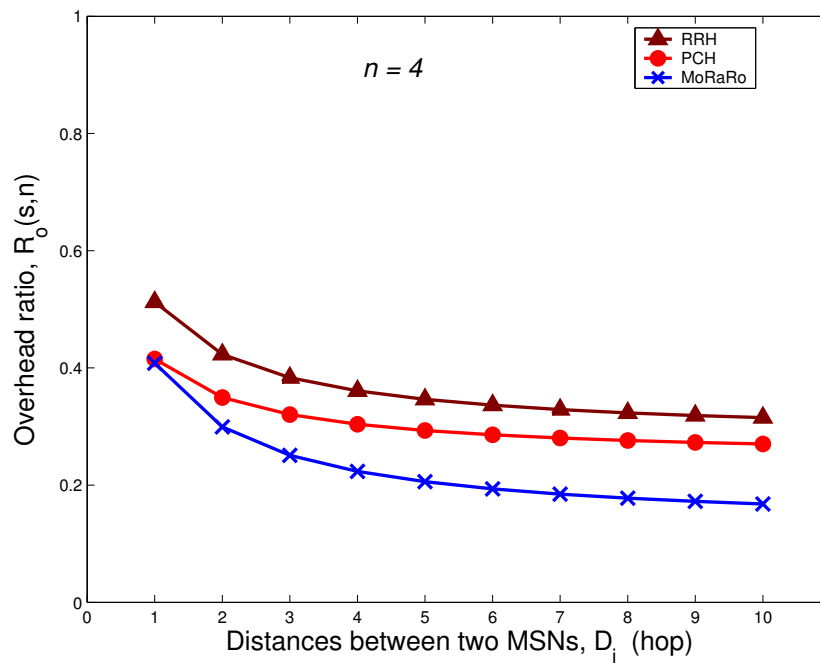
(b)

Figure 3.5: Overhead ratio versus distance between two MSNs for nested mobility levels of 2 (top) and 4 (bottom), with $E(D) = 100$ packets.
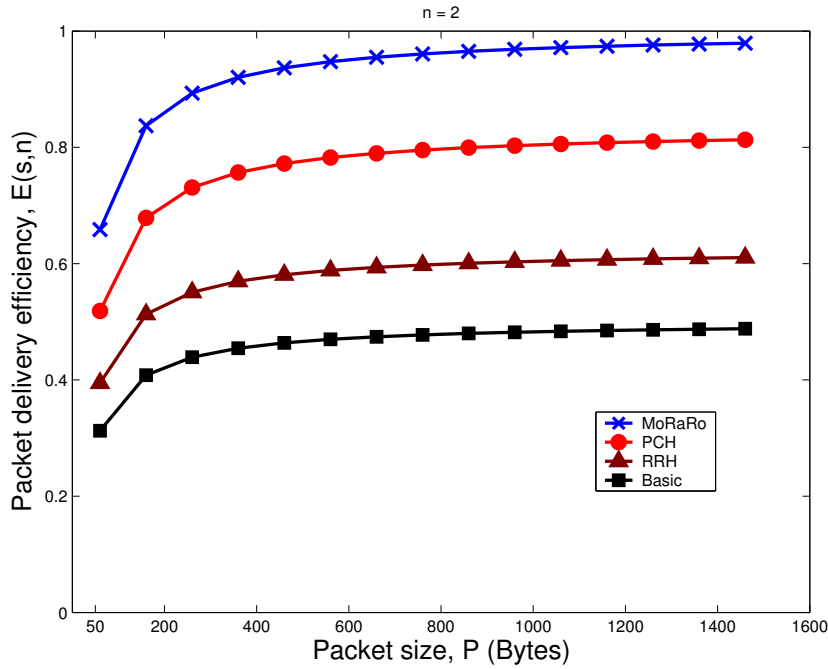
Figure 3.6: Packet delivery efficiency.

Table 3.5: Parameters for packet delivery delay analysis.

B_w	B_{wl}	L_w	L_{wl}	t_r	P	$1/\mu$
100 Mbps	11 Mbps	0.5 ms	2 ms	0.001 ms	1000 Bytes	4 ms

other hand, MoRaRo has an efficiency of more than 80% when the packet size is around (or larger than) 200 bytes. This indicates that MoRaRo increases the packet delivery efficiency of the NEMO basic support protocol by more than 100%. Furthermore, it is obvious that increasing the nested mobility level further reduces the efficiency of the basic support protocol, whereas our scheme is the least affected one.

3.5.2 Packet Delivery Delays

Table 3.5 lists the parameters used for the packet delivery delay analysis, which were taken from a paper by Lo et al. [34]. Figure 3.7 compares the results of the analysis of packet delivery delay for different loads on the MSNs. As expected, the maximum delay occurred in the basic support protocol and the minimum in the MoRaRo scheme. The PCH has slightly less delay than the RRH because it allows packets to be routed from a CR to the MR without going through a home agent. Figures 3.7(a) and (b) are drawn on the same scale for $n = 2$ and $n = 4$, respectively, to

show how the delays were affected by changes in the level of nested mobility. They show that the delays increased rapidly with increasing levels of nested mobility in all schemes except the proposed one. Clearly, our scheme performed optimally for any level of nesting and MSN utilization.

3.5.3 CBU Overhead for Handover

In addition to the analysis, we carried out a performance simulation to evaluate the CBU overhead. We considered a mobile network with 1000 visiting mobile nodes of which 20% were active at any one time. The whole network was served by an MR whose subnet (or cell) residence time was exponentially distributed with mean t_{sr}. Similarly, the inter-arrival time of outbound packets from an MNN via an MR was also exponentially distributed with mean t_{pia}. We assumed that each MNN sent 1000 packets during a session. The simulation was run for 2 hours of simulation time while counting the total number of CBU packets and data packets forwarded by the MR to the CNs.

Figure 3.8 shows the results of the analysis and simulation of CBU overheads. The vertical axis represents the percentage of CBU packets out of the total number of data packets forwarded by the MR to the correspondent nodes, and the horizontal axis represents the mean subnet residence time. It shows that the percentage of CBU packets was very small compared to the total number of data packets, and decreased as the subnet residence time increased. Therefore, the overhead induced by the CBU was very low. Furthermore, when compared to the gain in packet delivery efficiency (more than 100%) of the MoRaRo scheme, the increment in overheads (less than 2%) by the CBU signaling was insignificant.

In summary, the proposed MoRaRo scheme improved the overall performance of the network more than any other scheme. It enhanced the operation of a mobile network by significantly reducing packet delivery overheads and delays without requiring any modification of the protocol implementation at the CN, home agent or other Internet entities.

A possible drawback of MoRaRo is the addition of the correspondent binding update function to a mobile router, which could result in the mobile router becoming a bottleneck when serving a large number of mobile nodes. Our evaluation, however, indicates that the volume of additional signaling packets triggered by a correspondent binding update is very low (less than 2%) and its impact will be insignificant if the MR is capable of handing a large number of data packets.
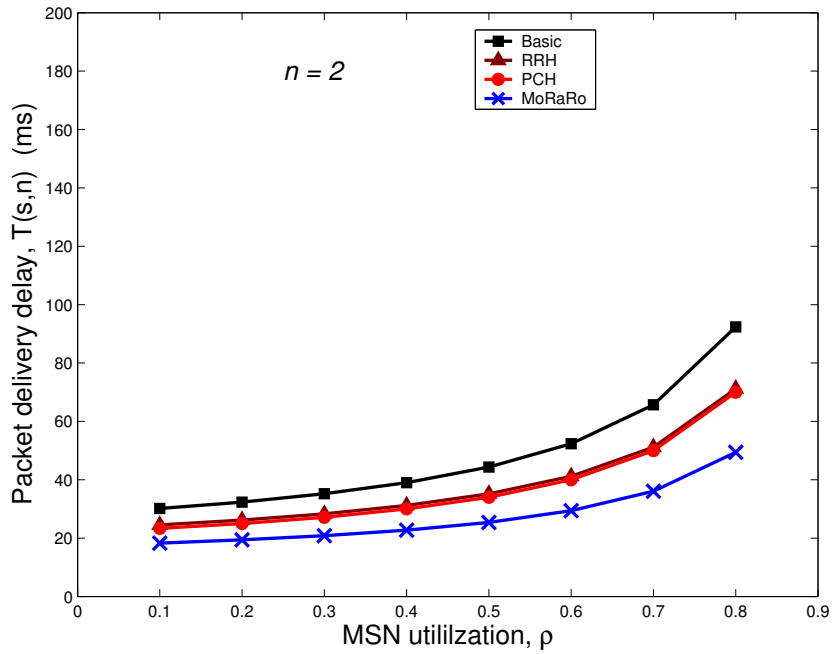
(a)
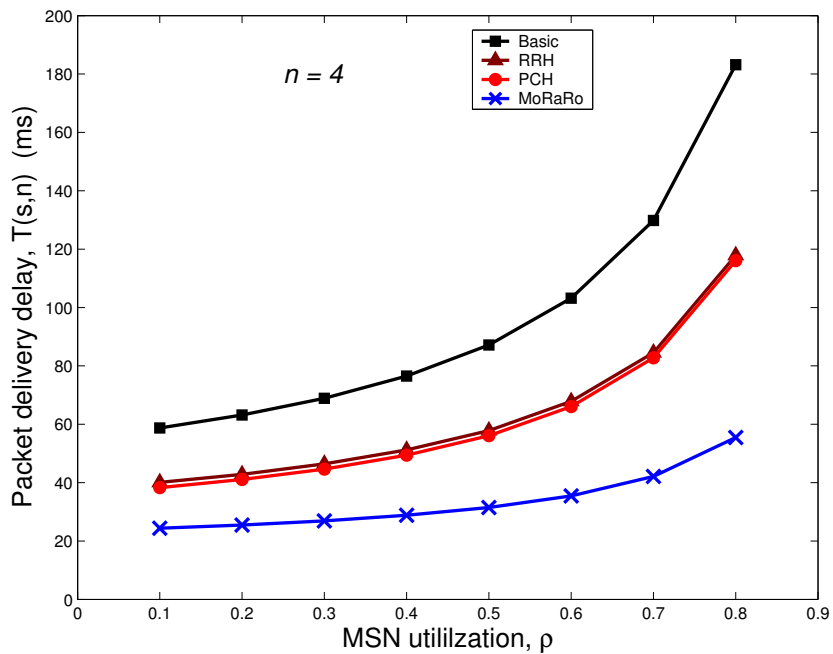
(b)

Figure 3.7: Packet delivery delay for nested mobility levels of 2 (top) and 4 (bottom).
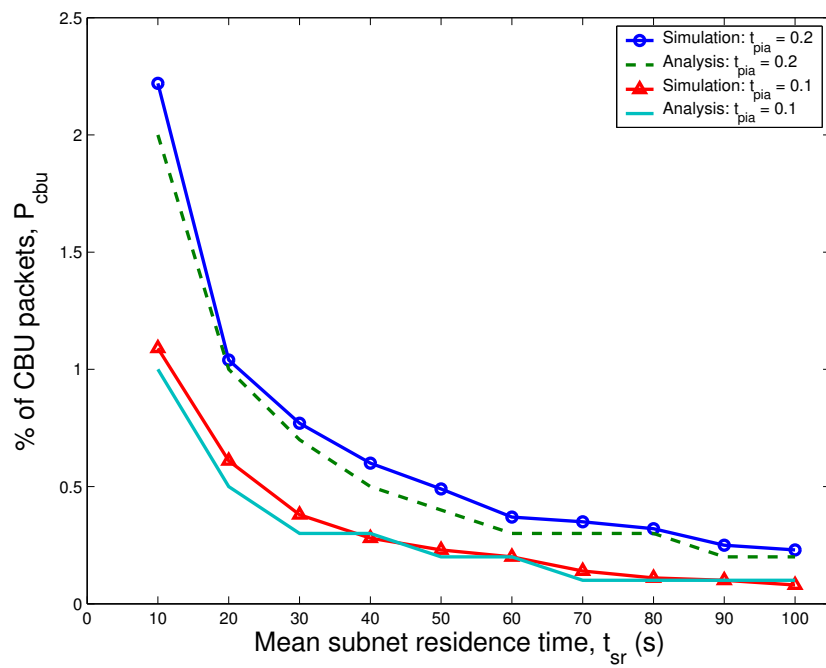
Figure 3.8: Percentage of correspondent binding update packets in MoRaRo.

3.6   Conclusion

The NEMO basic support protocol provides advantages by reducing location update overheads. However, it has the side effect of increasing packet delivery overheads due to pinball routing and multi-layer encapsulation of data packets. To solve this problem, we proposed a mobile router-assisted route optimization (MoRaRo) scheme for NEMO support. Our scheme is simple to implement as it requires only a slight change in the implementation of the NEMO basic support protocol in the local components of a mobile network such as mobile routers and mobile nodes. No change is required in correspondent nodes, home agents, or any other network components located in domains beyond the mobile network. The MoRaRo scheme enables a correspondent node to forward packets directly to the mobile network without any tunneling. Therefore, it reduces packet delays and encapsulation overheads in the core network. We evaluated the performance of the MoRaRo scheme in terms of packet delivery overheads, efficiency, and delay between correspondent nodes and mobile network nodes. We found that the scheme improved the packet delivery efficiency of the network mobility support protocol by about 100%. In addition, the packet delivery delay was significantly reduced.

Chapter 4

NEMO Handover Management

4.1 Introduction

We develop a cooperative mobile router-based handover (CoMoRoHo) scheme to minimize packet losses of a long-vehicular mobile network during a handover. For this purpose, we consider a multihomed mobile network with multiple MRs. A mobile network can be multihomed under different scenarios, such as by having an MR with multiple egress interfaces or by having a number of MRs or MNPs [21]. However, in this work we do not explicitly consider any other requirements for multihoming except having multiple MRs. Although issues relating to multihoming, such as prefix delegation and HA synchronization, have been discussed in the Internet-drafts of the NEMO working group [20], none of the Internet-drafts focuses on handover optimization of long-vehicular mobile networks.

The basic idea behind the CoMoRoHo scheme is to enable different MRs to access different subnets during a handover and cooperatively receive packets destined for each other. That is, when an MR is performing handover to a new subnet, it requests the AR of the old subnet to tunnel packets to some other MR that is still located in the old subnet. The other MR then forwards the packets to the MR that has requested the AR for tunneling. Similarly, when the MR completes its handover to a new network, it helps the other MR to perform a lossless handover to the new subnet. In general, the number of packet losses during a handover is directly proportional to the handover latency; however, the overlapped reception of packets from different subnets during the handover makes possible to minimize packet losses even without reducing the handover latency. To evaluate the CoMoRoHo scheme, we carried out performance modeling of the proposed scheme in comparison with the Fast Handover for Mobile IPv6 (FMIPv6) protocol [6] in terms of the handover latency, packet loss, signaling overhead, and packet delivery overhead in access networks.

This chapter is organized as follows. Section 4.2 summarizes the related literature. The proposed scheme is described in detail in Section 4.3. The analytical models for evaluating the scheme's performance are described in Section 4.4. Analysis results are presented in Section 4.5, and a conclusion is given in Section 4.6.

4.2 Related Work

4.2.1 Multihoming

Regarding multihoming in mobile networks, different types of multihoming scenarios, such as having an MR with a number of egress interfaces or having multiple MRs or MNPs in a mobile network, are outlined in [21] and [40]. Multihoming increases robustness of a mobile network system in terms of fault tolerance and load balancing. However, to the best of the authors' knowledge, there is no work on how multihoming could be utilized to enhance the performance of mobile networks during a handover. In this study, we therefore address the issues of improving the performance of multihomed mobile networks during a handover in terms of packet losses and signaling and packet-delivery overheads.

4.2.2 Handover Optimization

Although not directly related to NEMO, handover optimization schemes for a single mobile node moving on the Internet topology have been widely researched in the MIPv6-related research community. Hierarchical Mobile IPv6 (HMIPv6) [5] and FMIPv6 [6] are some of the widely referenced schemes for preventing the performance of a mobile node from being degraded during a handover. The approaches used in these schemes are more concerned with reducing handover latency and packet losses by employing the following two techniques: localizing signaling domain [5] or predicting mobility in advance to enable a mobile node to complete some part of the handover operation before moving to a new network [6]. These schemes can also be applied to NEMO; however, the NEMO problem domain is different from that of a single mobile node in regard to the aspects described in the following paragraph.

Firstly, in the case of a single mobile node, large handover latency causes the loss of a few packets belonging to the mobile node only. In the case of NEMO, however, large handover latency causes the loss of a large number of packets belonging to all MNNs of the mobile network. Moreover, if the MNNs are running TCP applications, the synchronized loss of their packets takes a longer time to be recovered [41]. This is because all TCP connections assume the packet losses are due to network congestion; consequently, they go through the slow-start phase

roughly at the same time, and may experience congestion simultaneously. Secondly, a mobile network such as a long-vehicular mobile network may require two or more MRs to provide Internet connectivity to a large number of MNNs. Mobile networks thus are more likely to be multihomed than mobile nodes. A multihomed mobile network with widely separated MRs can have simultaneous connections to two or more non-overlapped subnets during a handover. Thirdly, an individual mobile node may change its direction of movement more frequently, whereas a vehicular mobile network moves in a fixed direction for a considerably longer time.

These three aspects were in fact the factors that motivated us to develop CoMoRoHo in this work. The first aspect indicates that the reduction of handover latency, and the consequent effect on packet losses, is a very important requirement in enabling mobile networks to support multimedia communications. The second and third aspects indicate that mobile networks have special characteristics that can be utilized in designing a new type of handover scheme for mobile networks. In this work, both the multihomed mobile network's capability of receiving packets from different subnets and the beforehand knowledge of network movement direction are used in designing the CoMoRoHo scheme.

Note that the CoMoRoHo scheme does not assume any specific structure of the access network. However, this scheme can utilize such optimization as hierarchical configuration of access routers, cooperative access routers, or smart packet forwarding within the access network to further improve its performance.

4.3   Proposed Handover Scheme: CoMoRoHo

4.3.1   System Model and Assumptions

The system model and assumptions used in explaining the operation of the proposed Co-MoRoHo are described next. We consider a long-vehicular multihomed mobile network that has many MRs, from which an MNN selects one according to an algorithm such as [42]. We assume that all MRs are interconnected to each other and have good security associations among themselves.

To explain the operation of our scheme, we use terminology similar to that used in FMIPv6 [6]. When an MR moves out of a subnet of an AR and enters another subnet of another AR, in such a case, we refer to the former AR as a previous AR (PAR) and the latter AR as a new AR (NAR) of the MR. Similarly, we use the terms previous CoA (PCoA) and new CoA (NCoA) to refer to the MR's CoAs associated with the PAR and NAR, respectively.

Although the proposed handover scheme can support any number of MRs, for convenience, in this paper, we take into consideration two MRs: MR1 and MR2, with MR1 in the front and
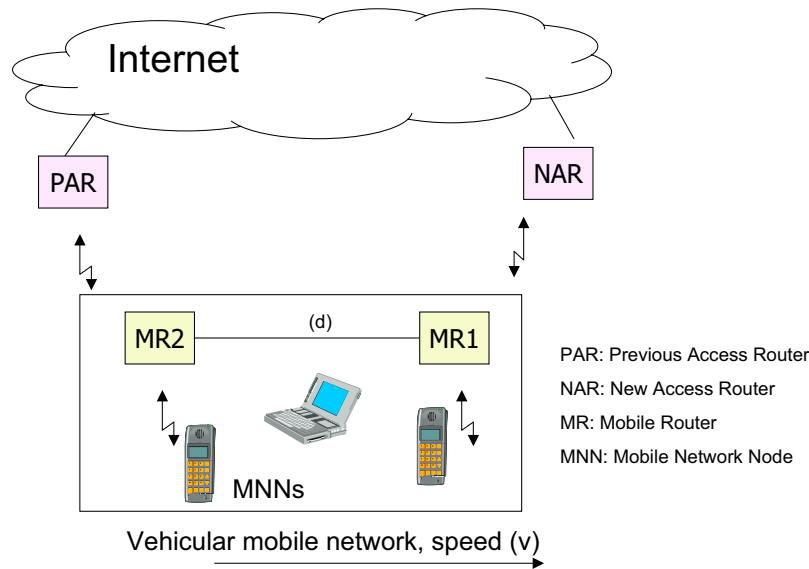
Figure 4.1: System model for CoMoRoHo scheme.

MR2 in the rear part of a vehicular mobile network, as shown in Fig. 4.1. During handover, MR1 leaves the subnet of PAR earlier than MR2, resulting in the mobile network having connectivity with Internet through only a single mobile router MR2 until MR1 completes its handover process. After that the mobile network gets the Internet connection through two different subnets via MR1 and MR2 for some time until MR2 leaves the PAR. Note that our scheme can be applied easily to cases with more than two MRs without any modification. In such cases, each MR should know its counterpart MR that can be requested to receive data packets during the handover of the mobile network.

The time duration (T_d) during which MR2 remains connected to the PAR after MR1 leaves the PAR depends on the distance (d) between MR1 and MR2, and the speed (v) of the mobile network, i.e., $T_d = d/v$. For a very-low-loss handover, MR2 should remain connected to the PAR for a duration greater than or equal to the handover latency (T_h) of MR1, i.e., $T_d \geq T_h$. For example, suppose a mobile network on a train where $d = 100$ meters and $v = 100$ km/hr. The value of T_d will thus be 3.6 seconds, which is sufficient time to allow MR1 to complete binding updates with its home agent before MR2 leaves the PAR if the mobile network and its home network are located in the same country. Similarly, if the distance between the two MRs is 100 meters and the MR can perform handover in 1 second, our scheme can support a train speed of up to 360 km/hr. Conversely, if the MR can perform handover in 50 ms and the mobile network moves at a maximum speed of 200 km/hr, the distance between the two MRs can be as small as 2.8 meters. This analysis indicates that for smaller handover latency, our scheme does not necessarily require the MRs be separated by a long distance. However, if the MRs are separated
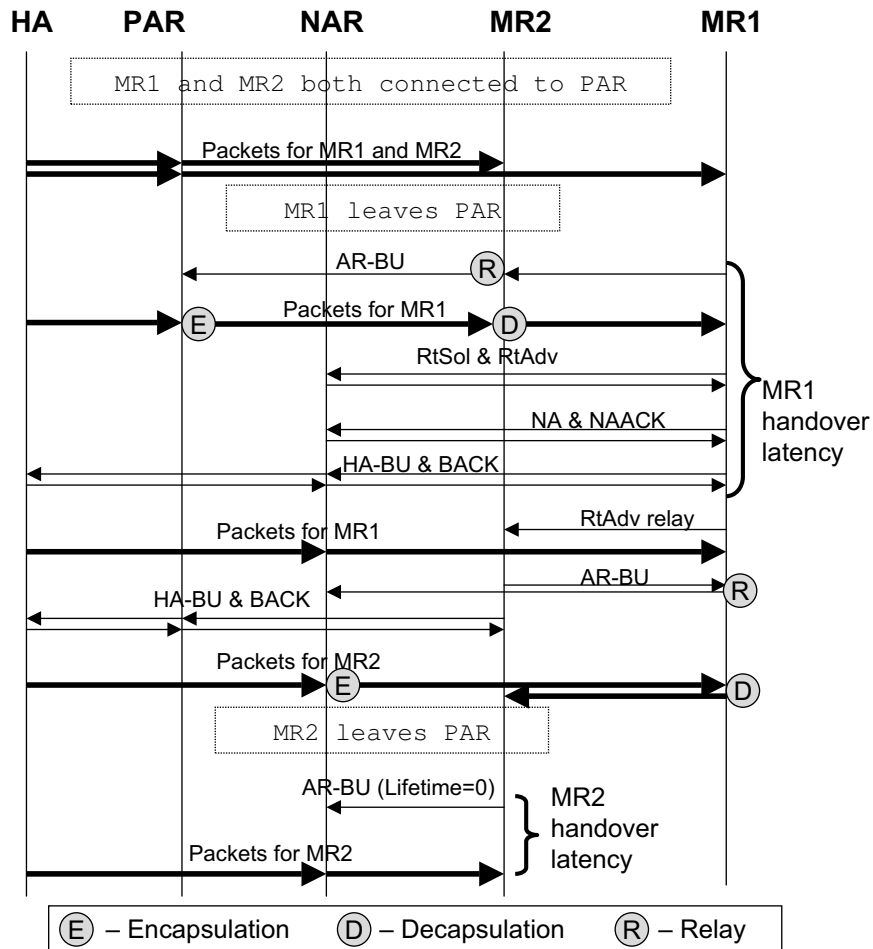
Figure 4.2: Handover signaling message interactions and data-packet flow paths during handover with CoMoRoHo scheme.

by a long distance, the scheme can support fairly high velocities and handover latencies. Our scheme is therefore more applicable to longer vehicles, such as trains, than smaller vehicles, such as cars.

4.3.2 Handover Operation

To achieve a low-loss handover, the objective of CoMoRoHo is to enable MRs to receive packets on behalf of one another via different subnets during the handover period. For this purpose, as soon as an MR detects that it has moved away from the subnet of the PAR, it informs the PAR to tunnel its packets to the other MR, which is still located in the subnet of the PAR. Similarly, after completing its handover to the NAR, the MR helps the other MR to perform lossless handover to the NAR.

The operation of the proposed handover scheme is depicted in Fig. 4.2. Before the handover, both MR1 and MR2 access the Internet via the PAR. When MR1 moves out of the subnet of the PAR, it executes the following functions.

- AR-binding update: MR1 sends a binding update message, which we call an AR-binding update (AR-BU), to the PAR via MR2, which can easily forward the AR-BU to the PAR because it is still in the PAR's subnet. The AR-BU message is similar to a home-agent binding update message of MR1, except the contents of the home-address destination option and alternate care-of address fields of the message. In the AR-BU message, the home-address destination option and alternate care-of address option fields contain MR1's PCoA and MR2's PCoA, respectively. On receiving the AR-BU message, the PAR creates a mapping between MR1's PCoA and MR2's PCoA, which is used to tunnel the packets addressed to MR1's PCoA. Namely, when the packets destined for MR1 arrive, the PAR encapsulates and forwards these packets to MR2's PCoA. MR2 then decapsulates and forwards the packets to MR1 through a wired link.

- NCoA configuration and HA-binding update: To configure an NCoA, MR1 requires the network prefix of the subnet of the NAR. It can obtain the network prefix in two ways: either by sending a router solicitation (RtSol) to request the NAR to send a router advertisement (RtAdv) or by waiting for a periodically broadcasted RtAdv from the NAR. In this paper, we consider the first case, i.e., the network prefix is obtained by requesting with an RtSol. After configuring an NCoA from the network prefix of the NAR, MR1 verifies the uniqueness of the NCoA by sending a neighbor advertisement (NA) to the NAR and consequently receiving a neighbor advertisement acknowledgement (NAACK). MR1 next performs the HA registration by sending an HA binding update (HA-BU) message, and then receiving a binding acknowledgement (HA-BACK) message from the HA. Upon completion of the registration, the HA tunnels packets to the NCoA through the NAR.

- RtAdv relay: After sending the HA-BU message to the HA, MR1 relays the RtAdv obtained from the NAR to MR2 in a 'RtAdv relay' message, so that MR2 can also configure a prospective NCoA for itself, while still remaining connected to the PAR.

Upon receiving the RtAdv relay message from MR1, MR2 infers that it has to perform handover to the NAR in the near future. MR2 then executes a sequence of functions as performed by MR1. It configures an NCoA from the network prefix of the NAR obtained in the RtAdv relay message. It then sends an AR-BU message to the NAR via MR1 to create a mapping between MR2's NCoA and MR1's NCoA. The creation of this mapping enables MR2 to perform the HA registration using its NCoA before moving to the NAR. Accordingly, MR2 completes

the HA registration by sending an HA-BU message to the HA and consequently receiving an HA-BACK message.

Following the HA registration, the HA forwards the packets destined for MR2 to the NCoA. These packets arrive at the NAR, which then tunnels them to MR1's NCoA using the address mapping created by the AR-BU performed by MR2. MR1 decapsulates and forwards the packets to MR2 through a wired link. In this way, having the AR-BU performed prior to the HA registration avoids the possibility of packet losses at the NAR. Similarly, by performing the home registration and using the NCoA before moving to the NAR, MR2 reduces the IP layer handover latency.

After receiving packets via MR1, MR2 decides whether to stay in the PAR or move to the NAR, depending on the availability and strength of the radio signal from the NAR in the current location of MR2. When MR2 enters the NAR's subnet, it can immediately use its already configured and registered NCoA. For this purpose, it has to deactivate its binding at the NAR created by the AR-BU so that the NAR will send packets destined for MR2 to MR2's NCoA, not to MR1's NCoA. MR2 deactivates the binding by sending an AR-BU message requesting the NAR to create a binding whose lifetime is zero. In other words, to create a binding with a zero lifetime means to delete the binding.

From the above explanation and Fig. 4.2, it is clear that the handover operation of MR1 consists of link switching, AR-BU, CoA configuration, and HA registration, whereas the handover of MR2 includes only link switching and AR-BU. Although MR1 takes a longer time than MR2 to perform handover, only few packets are lost because of performing the AR-BU with the PAR.

4.3.3   Security Consideration

We claim that our proposal does not create a new type of security vulnerability. In our approach, the MRs are fixed to the vehicular network, that is, they are not changing their positions with respect to one another. In such a networking environment, it is possible to have wired connections between the MRs, which are obviously more secure than wireless connections. Furthermore, as there are fixed numbers of MRs within vehicular networks, the static security association existing in any layer (i.e., IP or lower layers) can work properly amongst the MRs. In this case, an MR will relay the binding update messages or data packets only if the packets are from a known MR. This will prevent any malicious user from pretending to be a mobile router and harming the network functionality.

Moreover, as an MR creates an AR-BU message for itself and the other MR just relays the message, the AR can verify the origin of the AR-BU message. This also allows the MR that creates the AR-BU message to include an authentication header or to even encrypt the message

using a security association, such as the IPSec [31], existing between the MR and the AR. As outlined by Yamada et al. [33], a security association can be created when the MR presents its credentials to the AR to get authorization to use the access network service. In this way, the AR-binding update message is protected from any possible attacks by using the security associations existing between the MRs themselves, and between the MRs and the ARs.

4.3.4   Comparison with Fast Handover for Mobile IPv6

The operation of CoMoRoHo is compared with that of FMIPv6 protocol [6] in this subsection. We chose FMIPv6 for this comparison because it is a widely researched and a well-established protocol developed by the IETF for handover optimization of a mobile node in MIPv6-based networks. Since NEMO is also an extension to the MIPv6 mobility management solution, FMIPv6 is an appropriate candidate protocol to be adopted in NEMO as a handover optimization solution.

In the following, we discuss the operation of FMIPv6 when implemented in mobile routers, and how it differs from our scheme. FMIPv6 for a mobile router can operate in two different modes: predictive or reactive [6]. In the predictive mode of FMIPv6, the MR is required to complete a number of IP layer tasks, such as requesting the PAR for information about the NAR through a transaction of a Router Solicitation for Proxy (RtSolPr) and a Proxy Router Advertisement (PrRtAdv) message, configuring a CoA, sending a Fast Binding Update (FBU) message to the PAR, getting exchanged a Handover Indication (HI) and a Handover Acknowledgement (HACK) message between the PAR and the NAR, and receiving a Fast Binding Acknowledgement (FBack) message before switching the link from the old network to the new one. All these tasks take a considerable amount of time, during which the MR should be in the overlapped region covered by both the PAR and the NAR. Hence, to execute the predictive handover in high-speed mobile networks, the overlapped region should be extremely large, which may not be economically realistic. Therefore, the more practical mode of operation for FMIPv6 is the reactive mode, which is shown in Fig. 4.3. Here, an MR (such as MR1) initiates the procedure to set up a bi-directional tunnel between the PAR and the NCoA after the MR gets connected to the NAR. For this purpose, the MR encapsulates an FBU message in Fast Neighbor Advertisement (FNA) and sends it to the NAR. On receiving the FNA[FBU] message, the NAR negotiates with the PAR by exchanging FBU, HI, HACK, and FBack messages. This procedure takes some time, during which the packets destined for MNNs via the MR maybe dropped from the PAR.

There is a stringent requirement for the FMIPv6 operations that the neighboring access routers, i.e., the PAR and NAR, have knowledge of each other's information, like the network
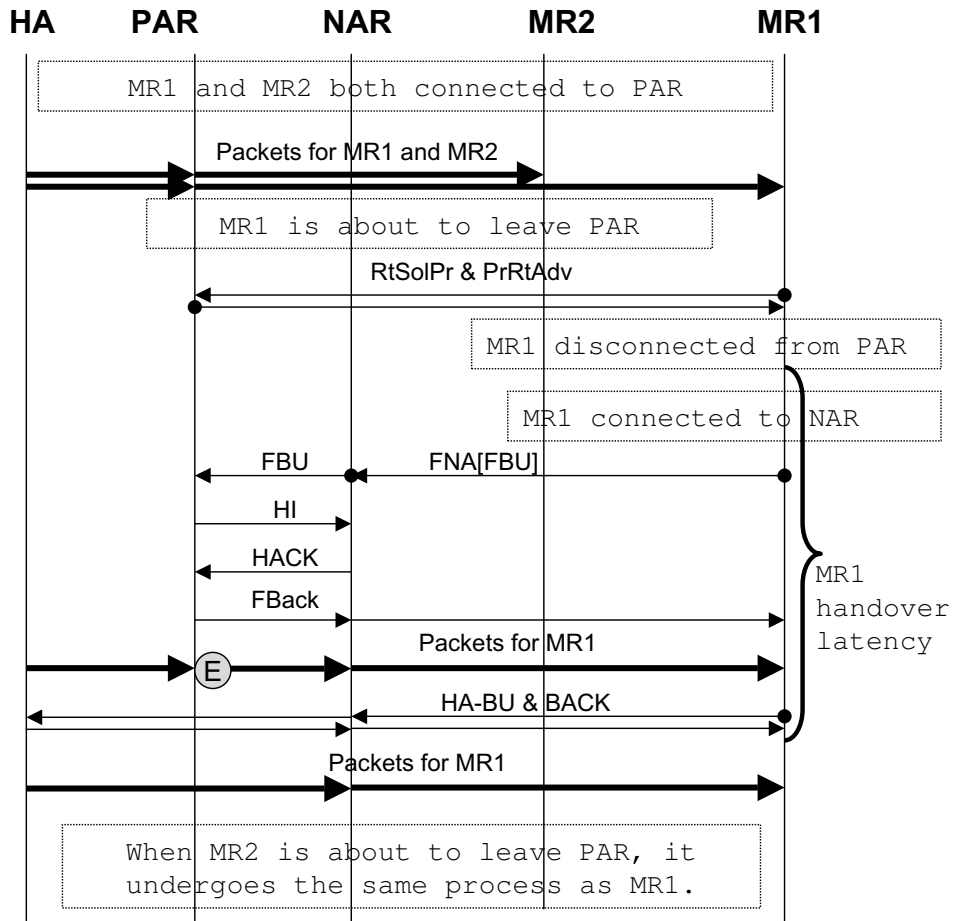
Figure 4.3: Handover signaling message interactions in reactive mode of FMIPv6. Circled E indicates that encapsulation of packets occurs at PAR.

prefixes, link-layer addresses, and security associations. This is only possible if all access routers belong to the same administrative domain. In contrast, our scheme assumes that MRs are cooperative with each other and forward packets for each other, which is a very logical and easy to meet requirement.

4.3.5 Advantages and Disadvantages

The advantages of CoMoRoHo are as follows.

(a) Little modification of existing system: Implementing CoMoRoHo requires modifying the NEMO basic support protocol implementation only at MRs. Except for implementing a tunneling mechanism, no other modifications are required at the MIPv6-enbaled ARs. Moreover, a specific configuration of access network is not required. The NEMO-enabled home agents and MIPv6-enabled correspondent nodes are left untouched.

(b) End-to-end principle preserved: There is no modification of IP data packets between the correspondent nodes and the MNNs.

(c) Signaling overhead minimized: To establish a local tunnel, it requires only one signaling, i.e., an AR-BU message from the MR to the PAR. It thus improves system scalability.

(d) Packet losses minimized: It enables a mobile network to receive packets from two or more subnets during the handover period. It therefore reduces the possibilities of packet losses and network service interruption.

The disadvantage of CoMoRoHo is that it imposes an overhead of an additional encapsulation header attached to data packets over the wireless link between an AR and an MR. However, the encapsulation is mandatory to protect the end-to-end packet integrity. We argue that enduring an overhead is far better than interrupting the network service during the handover period. To check the impact of this overhead on system performance, in Sections 4.4 and 4.5, we evaluate the packet delivery overhead of CoMoRoHo and compare the overhead with that of FMIPv6.

4.4   Performance Modeling

A quantitative method for evaluating and comparing the performance of CoMoRoHo with FMIPv6 is developed in this section. One of the important metrics in evaluating handover mechanisms is the handover latency and its impact on packet loss; the other is the signaling overhead generated in the network by the handover management operation. We first derive the expressions of handover latency, and then relate them to packet loss. For this evaluation, we will consider only MR1, because in our scheme MR2's handover latency is very low, composed of only the link-layer switching time and AR-BU time.

4.4.1   Handover Latencies

The handover latency (T_h) of mobile router MR1 is equal to the sum of the times required for MR1 to carry out the following functions: a. movement detection and link switching (T_{md}), b. AR-BU via MR2 (T_{ar_bu}), c. CoA configuration (T_{coa}), and d. HA-BU (T_{ha_bu}). Namely, the handover latency of the proposed scheme is given as

$$T_h^{CoMoRoHo} = T_{md} + T_{ar\_bu} + T_{coa} + T_{ha\_bu}. \qquad (4.1)$$

Although the link-switching latency varies with the link- and physical-layer technologies used in the MR, we set T_{md} equal to 10 ms. This value corresponds to an achievable link-layer handover latency of wireless media by employing some optimizations [46]. However, the selection of this value has no significant influence on the IP-layer performance comparisons because it increases the IP-layer handover latencies of all the schemes under consideration by an equal amount.

To derive the expressions of the remaining component times of the Eq. (4.1), we use the following parameters with some typical values referenced from [45].

B_w – bandwidth of wired links (100 Mbps);

B_{wl} – bandwidth of wireless links (11 Mbps);

L_w – latency of a wired link, i.e., propagation delay + link-layer delay (0.5 ms);

L_{wl} – latency of a wireless link (2 ms);

D_{mr_ha} – distance between the MRs and the HA in hops;

P_x – IP packet length of a signaling message x;

t_s – time to configure/process a signaling message, which normally takes less than 1 ms [44]. Unless specified, we assign to it a small constant value for all types of signaling messages (0.8 ms);

t_r – routing-table look-up and processing time for a packet in every hop (0.001 ms) [34];

λ_p – mean packet arrival rate of an MNN.

The component times are then expressed as follows.

$$T_{ar\_bu} = t_s + \frac{P_{ar\_bu}}{B_w} + L_w + \frac{P_{ar\_bu}}{B_{wl}} + L_{wl} \tag{4.2}$$

$$T_{coa} = (2t_s + \frac{P_{rtsol} + P_{rtadv}}{B_{wl}} + 2L_{wl}) + t_s + (2t_s + \frac{P_{na} + P_{naack}}{B_{wl}} + 2L_{wl}) \tag{4.3}$$

$$T_{ha\_bu} = (2t_s + \frac{P_{ha\_bu} + P_{ha\_back}}{B_{wl}} + 2L_{wl})$$
$$+ (2t_r + \frac{P_{ha\_bu} + P_{ha\_back}}{B_w} + 2L_w) \times (D_{mr\_ha} - 1) \tag{4.4}$$

For deriving Eqs. (4.2)–(4.4), we used the following logics. When MR1 leaves the PAR, it formulates an AR-BU message and sends it to the PAR via MR2. After that, MR1 starts a process to formulate a CoA for itself. For this purpose, MR1 sends a router solicitation (RtSol) message to the NAR to trigger the NAR to send a router advertisement (RtAdv). MR1 uses the network prefix advertised in the RtAdv to configure its new CoA. As in FMIPv6, MR1 verifies the uniqueness of the NCoA by sending a neighbor advertisement (NA) to the NAR and consequently receiving a neighbor advertisement acknowledgement (NAACK). Note that the uniqueness of a CoA in FMIPv6 is verified through the exchange of FNA and NAACK messages between the MR and the NAR, or HI and HACK messages between the PAR and NAR, depending on the mode of operation. MR1 then sends a binding update to its home agent, which the home agent uses to update the binding between the mobile network prefix and the new CoA of MR1. The home agent then returns a binding acknowledgement (HA-BACK) to MR1. On receiving the HA-BACK via the NAR, the MR1 handover process completes. Henceforth, MR1 will receive packets tunneled from its home agent to its new care-of address. The handover latency of MR1 is therefore the duration of time from the instance when MR1 gets disconnected from the PAR to the instance when it receives the HA-BACK message from the home agent.

Similarly, we derive the expressions of handover latencies of the basic support protocol and FMIPv6 as given by Eqs. (4.5) and (4.6), respectively.

$$T_h^{Basic} = T_{md} + T_{coa} + T_{ha\_bu} \tag{4.5}$$

$$T_h^{FMIPv6} = T_{md} + T_{fcoa} + T_{fbu} + T_{ha\_bu} \tag{4.6}$$

where T_{fco} is the time taken to configure a CoA using RtSolPr and PrRtAdv messages, and T_{fbu} is the time taken to complete a fast binding update between the PAR and the NAR. Referring to Fig. 4.3, we can derive T_{fcoa} and T_{fbu} as given below.

$$T_{fcoa} = (2t_s + \frac{P_{rtsolpr} + P_{prrtadv}}{B_{wl}} + 2L_{wl}) + t_s \tag{4.7}$$

$$T_{fbu} = (2t_s + \frac{P_{fna[fbu]} + P_{fback}}{B_{wl}} + 2L_{wl}) + \sum_Y (2t_s + \frac{P_x}{B_w} + L_w) \tag{4.8}$$

where $x \in Y$ and $Y = \{fbu, hi, hack, fback\}$, i.e., a set of the fast binding update, handover indication, handover acknowledgement, and fast binding acknowledgement messages exchanged between the PAR and the NAR. We would like to note that because the same numbers of signaling messages are involved in executing the predictive and reactive modes of FMIPv6, the handover delays and the signaling overheads in both modes are the same.

4.4.2 Packet Losses during Handover

An analytical model for evaluating the possible number of packet losses during a handover is developed in this subsection. Referring again to Fig. 4.2, during the handover latency given by Eq. (4.1), MR1 can not receive packets from the access network only for the first two parts of the latency period, i.e., starting from the instance that MR1 is disconnected from the PAR to the instance that the PAR receives the AR-BU message sent by MR1 via MR2. Data packets arriving during this time at the PAR and destined for the MNNs connected through MR1 may be dropped at the PAR. The number of packet losses in the case of CoMoRoHo is thus

$$L^{CoMoRoHo} = N_1 \lambda_p (T_{md} + T_{ar\_bu}) \tag{4.9}$$

where N_1 is the number of MNNs having active communication sessions through MR1.

We next derive the possible number of lost packets in the NEMO basic support protocol and FMIPv6. In the case of the basic support protocol, MR1 cannot receive packets from its home agent during handover until it completes the home-agent binding update. The number of lost packets is therefore given by the following expression.

$$L^{Basic} = N_1 \lambda_p T_h^{Basic} \tag{4.10}$$

Since the predictive mode of FMIPv6 performs almost all of the handover-related tasks while being connected to the old network, the packet loss may be negligible if the handover prediction is made at an appropriate time [49]. Therefore, for our comparison we have to take into consideration the packet loss of the reactive mode of FMIPv6. Referring to Fig. 4.3, in the case of FMIPv6, the mobile router cannot receive its packets until it completes the fast binding update between the PAR and the NAR. The number of lost packets in FMIPv6 is thus

$$L^{FMIPv6} = N_1 \lambda_p (T_{md} + T_{fbu}) \tag{4.11}$$

where it is assumed that the mobile router configures an NCoA while remaining connected to the PAR. If the mobile router moves before configuring the NCoA, the time taken to configure the NCoA will also be included in the packet loss period.

4.4.3 Handover Signaling Overheads

To manage a handover effectively, mobile routers exchange signaling messages with other entities in the access and core networks. In this subsection, we derive the expression of the signaling overhead of the proposed CoMoRoHo scheme and that of FMIPv6 with respect to the NEMO

basic support protocol.

Signaling overheads can be classified into two categories, local and global, based on the scope of signaling message traversal in networks [23]. Local signaling messages, such as the messages required to set up a local tunnel between ARs and MRs, are confined to a small region in the access network or mobile network. On the other hand, global signaling messages, such as the home-agent binding updates, traverse a longer distance on the Internet. The global signaling overhead is the same in all the schemes, namely, the proposed scheme, FMIPv6, and the NEMO basic support protocol, because in these schemes, on every handover, each MR is required to send a BU message to the HA and consequently receive a BACK message from the HA. To compare their performances, we therefore take into consideration only the local signaling overheads.

We define the signaling overhead as the cumulative traffic-load imposed by the signaling messages in the network. In other words, the signaling overhead is a function of three parameters: the number of messages, the size of each message, and the distance traversed by each message, and measured as the cumulative product of the message size and the distance it travels in the network [8, 34]. Note that this definition captures the notion of both the bandwidth consumption and the number of signaling messages. In an IP network, the distance between two network entities is measured in terms of the number of hops. Furthermore, there has been a well-established belief that the transmission cost of a wireless hop is higher than that of a wired hop due to frame retransmissions and media contentions at the data-link layer of a wireless link, which usually has a lower bandwidth than a wired link [43, 47, 48]. We denote this additional cost by a factor β ($\beta > 1$), representing that the overhead of transmitting a message over a unit length of a wireless link is β times the overhead of transmitting the same message over a unit length of a wired link.

Referring to Figs. 4.2 and 4.3, we express the local signaling overheads of the proposed scheme and FMIPv6 by Eqs. (4.12) and (4.13), respectively. Eq. (4.12) uses a denominator of 2 to estimate the average overhead associated with a single mobile router of a mobile network with two mobile routers.

$$
\begin{aligned}
S_o^{CoMoRoHo} = [P_{ar\_bu}(2 + 3\beta) + (P_{rtsol} + P_{rtadv})\beta \\
+ (P_{na} + P_{naack})\beta + P_{rtadv\_relay}]/2
\end{aligned}
\tag{4.12}
$$

$$
\begin{aligned}
S_o^{FMIPv6} = (P_{rtsolpr} + P_{prrtadv})\beta + P_{fna[fbu]}\beta \\
+ P_{fbu} + P_{hi} + P_{hack} + P_{fback}(1 + \beta)
\end{aligned}
\tag{4.13}
$$

 Note that the local signal overhead of CoMoRoHo is independent of the distance between the PAR and the NAR, whereas that of FMIPv6 increases with increasing distance. In deriving above expressions of signaling overheads, we assumed that the distance is one hop. Similarly, the signaling overhead of the basic support protocol is given below.

$$S_o^{Basic} = (P_{rtsol} + P_{rtadv})\beta + (P_{na} + P_{naack})\beta \qquad (4.14)$$

where we assumed that the basic support protocol also uses NA and NAACK messages to check the validity of the NCoA.

To get the signaling overheads of CoMoRoHo and FMIPv6 with respect to the NEMO basic support protocol, we express the signaling overheads in ratios (S_R) by dividing the overheads of CoMoRoHo and FMIPv6 by that of the basic support protocol. That is,

$$S_R^{CoMoRoHo/FMIPv6} = \frac{S_o^{CoMoRoHo/FMIPv6}}{S_o^{Basic}} \qquad (4.15)$$

where the '/' sign in the subscripts stands for 'or', not a division sign. In fact, by using this sign, the single equation represents two equations: one for CoMoRoHo and the other for FMIPv6. In other words, the signaling overhead ratios ($S_R^{CoMoRoHo/FMIPv6}$) represent the factor by which the signaling overheads of CoMoRoHo and FMIPv6 increase as compared with the basic support protocol.

4.4.4 Packet Delivery Overheads

Handover management also involves additional packet delivery overheads as data packets from access routers to mobile routers are forwarded via tunneling, which requires an additional header to be attached to each packet. In fact, the signaling overheads discussed in the previous subsections, excluding the overhead of a new CoA configuration, are also used to establish the tunnel through which data packets are routed during a handover. We can therefore combine the tunneling header overhead with the signaling overhead to estimate the average per packet overhead required to prevent a packet from being dropped from access routers during a handover.

To evaluate the packet delivery overhead introduced by the handover management schemes, we first derive the average value of per-packet traffic load of each scheme, and then derive the packet delivery overheads of CoMoRoHo and FMIPv6 by comparing their per-packet traffic load with the traffic load of the NEMO basic support protocol. The average per-packet traffic loads, from the access network to the mobile router, of CoMoRoHo and FMIPv6 are given by

the following equations.

$$T_o^{CoMoRoHo} = \frac{S_T^C + n_p^C[(H + P_{data})\beta + P_{data}]}{n_p^C} \tag{4.16}$$

$$T_o^{FMIPv6} = \frac{S_T^F + n_p^F(H + P_{data})(1 + \beta)}{n_p^F} \tag{4.17}$$

In Eqs. (4.16) and (4.17), S_T^C and S_T^F are the signaling overheads required to establish a local tunnel from the PAR to the mobile router, and n_p^C and n_p^F are the numbers of packets that are forwarded via tunneling in the proposed scheme and FMIPv6, respectively. H and P_{data} are the IPv6 header size and the IP data-packet size arriving at the PAR, respectively. Similarly, since there is no additional tunneling from the access router to the mobile router in the case of the NEMO basic support protocol, the per-packet traffic load is given as follows.

$$T_o^{Basic} = P_{data}\beta \tag{4.18}$$

We divide Eqs. (4.16) and (4.17) by (4.18) to get per-packet delivery overhead ratios (P_R) of CoMoRoHo and FMIPv6, respectively. That is,

$$P_R^{CoMoRoHo/FMIPv6} = \frac{T_o^{CoMoRoHo/FMIPv6}}{T_o^{Basic}} \tag{4.19}$$

where, like in Eq. (4.15), the '/' sign in the subscripts stands for 'or'. Eq. (4.19) represents the factors by which the traffic loads imposed by a data packet during handover under CoMoRoHo and FMIPv6 increase in comparison with the NEMO basic support protocol.

Table 4.1: Signaling messages size in octets (IPv6 header is not included).

Message	Used in	Size
Router Solicitation (RtSol)	CoMoRoHo	16
Router Advertisement (RtAdv)	CoMoRoHo	56
Access Router Binding Update (AR-BU)	CoMoRoHo	60
Neighbor Advertisement (NA)	CoMoRoHo	32
Neighbor Advertisement Acknowledgement (NAACK)	CoMoRoHo	48
Home Agent Binding Update (HA-BU)	CoMoRoHo, FMIPv6	60
Home Agent Binding Acknowledgement (HA-BACK)	CoMoRoHo, FMIPv6	20
Router Solicitation for Proxy (RtSolPr)	FMIPv6	40
Proxy Router Advertisement (PrRtAdv)	FMIPv6	104
Handover Indication (HI)	FMIPv6	72
Handover Acknowledgement (HACK)	FMIPv6	32
Fast Binding Update (FBU)	FMIPv6	60
Fast Neighbor Advertisement (FNA)	FMIPv6	24
Fast Binding Acknowledgement (FBack)	FMIPv6	12
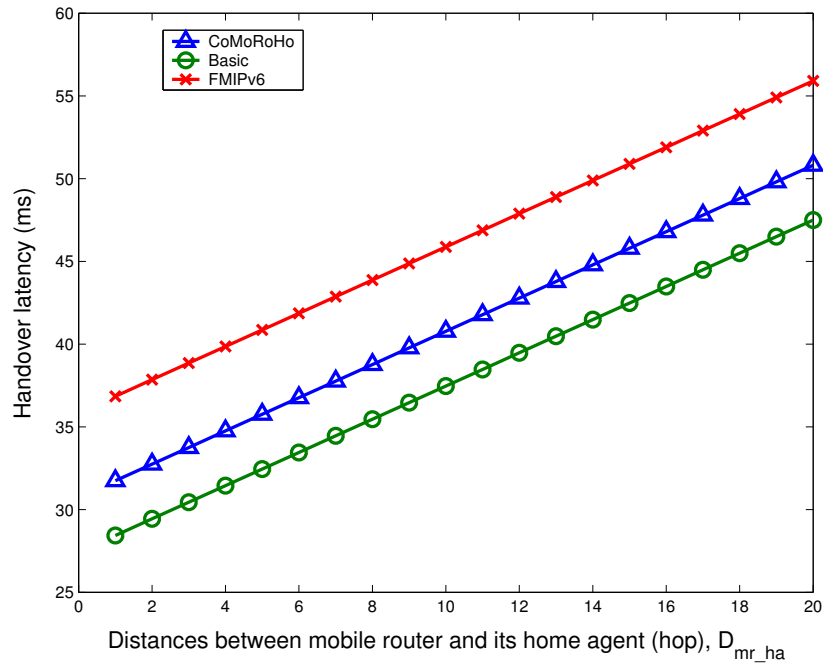
Figure 4.4: Handover latencies of different schemes.

4.5   Results and Discussions

We now present the analysis results of the performance of CoMoRoHo by comparing it with FMIPv6 in terms of the handover latencies, packet losses, signaling overheads, and packet delivery overheads. To get these numerical results, we use the parameter values as specified in Section 4.4.1 and Table 4.1. The sizes of the signaling messages in Table 4.1 are referenced from the RFC 3963 [22], RFC 4069 [6], and RFC 2461 [50].

Figure 4.4 shows the handover latencies of the NEMO basic support protocol, FMIPv6, and the proposed CoMoRoHo. It shows that handover latencies of all protocols increase linearly with increasing the distances between the mobile router and its home agent. The NEMO basic support protocol has the lowest handover latency, as it does not take any additional time to establish a local forwarding path before performing the HA binding update. On the other hand, FMIPv6 has the highest handover latency, as it requires a transaction of a number of messages to establish a local tunneling between the PAR and the NCoA before performing the HA binding update. The handover latency of our CoMoRoHo scheme is less than FMIPv6 because, to establish a local tunneling, our scheme needs only one signaling message from mobile router MR1 to the PAR via MR2.

Although the handover latencies of CoMoRoHo and FMIPv6 are larger than that of the NEMO basic support protocol, in the former two schemes mobile router MR1 can receive data packets after setting up a local forwarding path from the PAR to MR1 (via MR2 in CoMoRoHo and via the NAR in FMIPv6). Figure 4.5 shows the possible numbers of packets destined for MNNs via MR1 but lost at the PAR during the handover of MR1. This figure shows that the lowest number of packets are lost in the proposed CoMoRoHo scheme, which is independent of the distance between MR1 and its HA, because of setting a local forwarding path between the PAR and MR1 using a single signaling message. FMIPv6 also establishes a tunneling path between the PAR and MR1, but using a number of signaling messages. The number of packets lost in FMIPv6, although independent of the distance between MR1 and the HA and less than that of the basic support protocol, is therefore more than that of the proposed scheme. In contrast, the basic support protocol has to perform home-agent binding update, which requires time proportional to the distance between the MR and its HA, to enable the HA to forward packets to the MR via the NAR. Until receiving the HA-binding update message, the HA keeps forwarding the packets to the PAR, which are latter dropped from the PAR. The basic support protocol therefore has the highest packet loss that increases with increasing the distance between the mobile router and its home agent.

Figure 4.6 shows that the advantage of having a lower number of packet losses in FMIPv6, compared to the basic support protocol, vanishes as the computation time required by a mobility
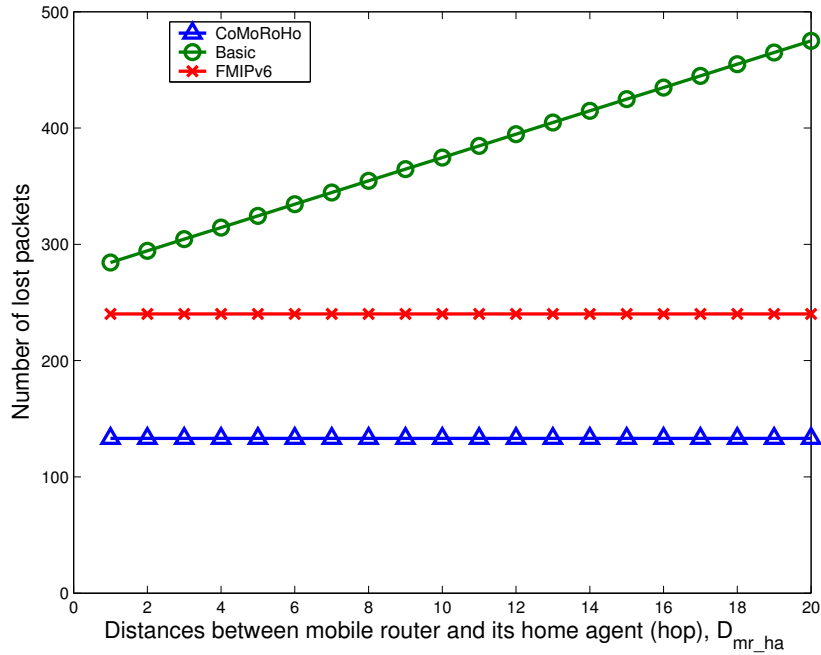
Figure 4.5: Number of packet losses with different schemes versus distance between mobile router and its home agent ($N_1 = 100$ nodes, $\lambda_p = 100$ pkt/s).
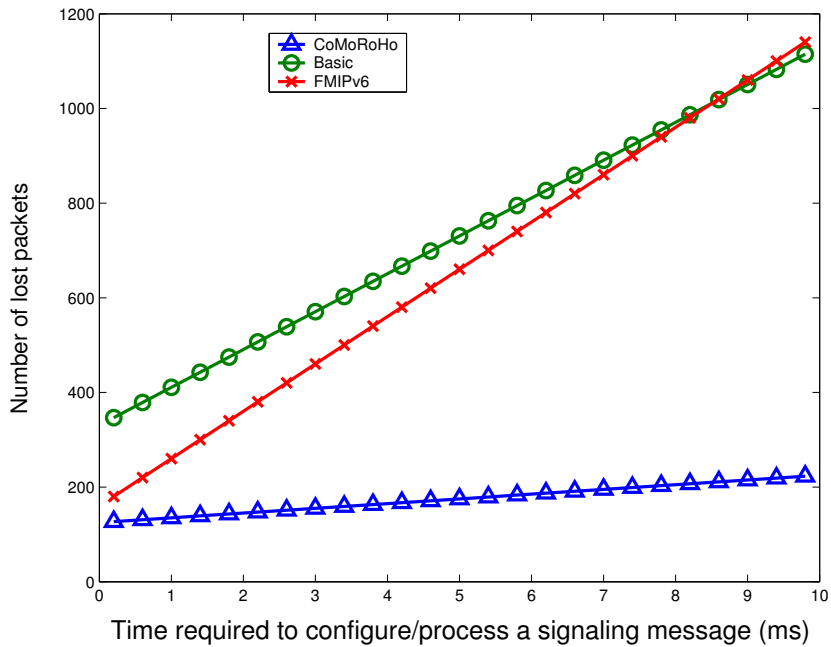
Figure 4.6: Number of packet losses with different schemes versus time required to configure/process a signaling message in a mobility agent ($N_1 = 100$ nodes, $\lambda_p = 100$ pkt/s, $D_{mr\_ha} = 12$ hops).
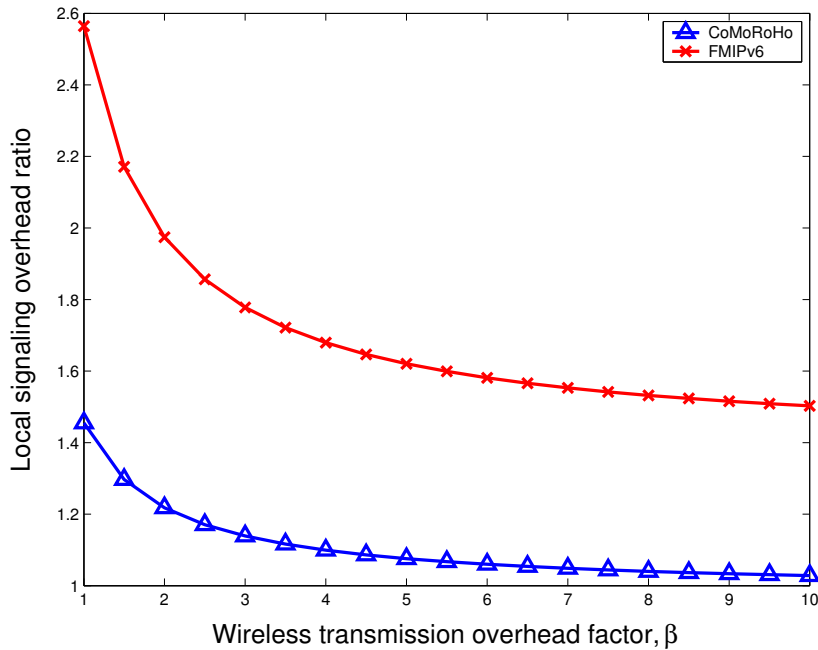
Figure 4.7: Local signaling overhead ratios of CoMoRoHo and FMIPv6 schemes.

agent (e.g., MR, AR, or HA) to process or configure a signaling message increases beyond 5 ms. A message may take longer time to be processed when a mobility agent is overloaded with signaling messages, data packets, or retransmitting functions due to the bad condition of wireless links. The figure shows that although the number of packet losses in our CoMoRoHo scheme increases slowly with time, it remains lowest for all the values of the processing time. The results clearly indicate that our scheme's performance remains intact even when the access network is overloaded.

Figure 4.7 compares the results of the analysis of signaling overhead ratios for different values of wireless transmission overhead factor, β. It is clear that CoMoRoHo induces less signaling overheads in the access network than FMIPv6, irrespective of the values of β. In comparison with the basic support protocol, CoMoRoHo increases the signaling overhead in the network by a factor of less than 1.5, whereas FMIPv6 increases the signaling overhead by a factor of more than 1.5. The figure shows that the signaling overhead ratios of both the proposed CoMoRoHo and FMIPv6 decrease with increasing β for the following reason. The local signaling overhead of the basic support protocol is composed of two messages: router solicitation and router advertisement, both of which are transmitted over the wireless link. However, the local signaling overheads of CoMoRoHo and FMIPv6 are composed of additional messages, some of which are transmitted over wired links. The increase in β therefore has a relatively higher impact on the signaling overhead of the basic support protocol; consequently, the overhead ratios
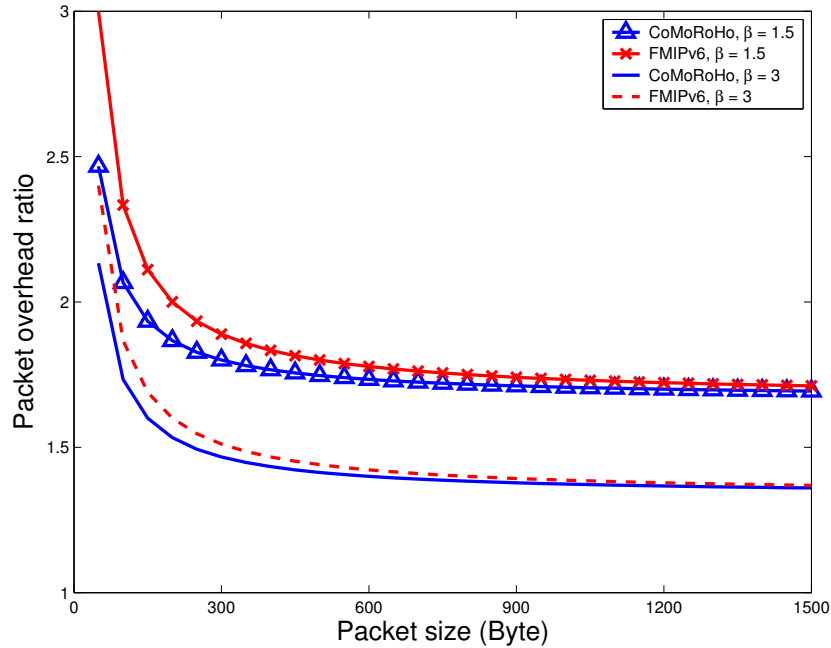
Figure 4.8: Packet delivery overhead ratios of CoMoRoHo and FMIPv6 schemes.

of CoMoRoHo and FMIPv6 decrease with increasing β.

Figure 4.8 shows the packet-delivery overhead ratios of CoMoRoHo and FMIPv6 for different packet sizes. The figure shows that, in terms of packet-delivery overhead, our CoMoRoHo scheme performs better than FMIPv6. For smaller packet sizes such as TCP acknowledgements, packet-delivery overhead is high for both the schemes, because tunneling headers consume a comparatively larger share of network resources. Similarly, for both the schemes the packet-delivery overhead ratio decreases as the value of β increases. This is because with increasing β per packet traffic load of the basic support protocol increases faster than the traffic load of CoMoRoHo or FMIPv6 as the traffic load of the basic support protocol consists of only the wireless transmission from the AR to the MR. The results indicate that although the performance difference, in terms of packet-delivery overhead, between CoMoRoHo and FMIPv6 diminishes as the packet size increases, our scheme performs better for smaller packet sizes, irrespective of the values of β.

In summary, the proposed CoMoRoHo scheme improves the performance of a vehicular mobile network during handover by reducing the number of packet losses, at the cost of introducing relatively lower amounts of signaling-message and packet-delivery overheads in the network. A possible limitation of this study is that it does not consider the dynamics of the link-layer and wireless transmission mechanisms in evaluating the handover latency and packet losses. Nonetheless, modeling the IP-layer behavior of the proposed scheme and comparing

it with other established protocols such as FMIPv6 allow us to draw the conclusion that the scheme is suitable for handover management of long-vehicular mobile networks.

4.6   Conclusion

We developed a cooperative mobile router-based handover (CoMoRoHo) scheme for long-vehicular-multihomed mobile networks. This scheme exploits the advantage of a multihoming environment of a long vehicular mobile network, which can have two or more mobile routers spatially separated by a certain distance. It thereby makes the packet loss independent of the handover latency by establishing a local tunnel between an access router and a mobile router by using only one signaling message. We compared the performance of the CoMoRoHo scheme with the Fast Handover for Mobile IPv6 protocol by formulating analytical models, and found that our scheme outperformed this protocol in regard to the number of packet losses, signaling-message overhead, and packet-delivery overhead in the access network. Our scheme worked well, even when the access network was overloaded.

Chapter 5

Optimal Wireless Network Selection

5.1 Introduction

In this chapter, we evaluate the impact of mobility on a user satisfaction function (USF), a user centric measurement of network service quality [53, 54, 55], and present an optimal network selection algorithm. We quantify user satisfaction as a function of the bandwidth utility and handover latency [15]. The bandwidth utility function of a network service is a measure of the degree of network capability to fulfill the bandwidth requirement of applications.

We analyze how user satisfaction changes when a user moves from a high bandwidth network to a low bandwidth network, under two types of movements: uniform random and directional movements. In the case of a random movement, we have no idea about the direction in which the user will move in the near future. To model such a random system, we use the well-known and widely used random-walk mobility model. Our evaluation shows that in a random movement, the change in user satisfaction due to a handover to a lower bandwidth network is not significant. However, when we consider a directional movement, where the movement probability to a low bandwidth network is much higher and, consequently, the user uses a high bandwidth network for a short time and hands over to a low bandwidth network to complete the call, the satisfaction degradation is significant. Therefore, to avoid such degradation, we propose an access network selection algorithm that uses a prediction of the user's mobility pattern.

The remainder of this chapter is organized as follows. We define the bandwidth utility and user satisfaction functions in Section 5.2. In Section 5.3, we derive mobile user's movement probabilities using the random-walk mobility model. Section 5.4 presents the analytical results and an access network selection algorithm to maximize user satisfaction. Section 5.5 concludes this chapter.
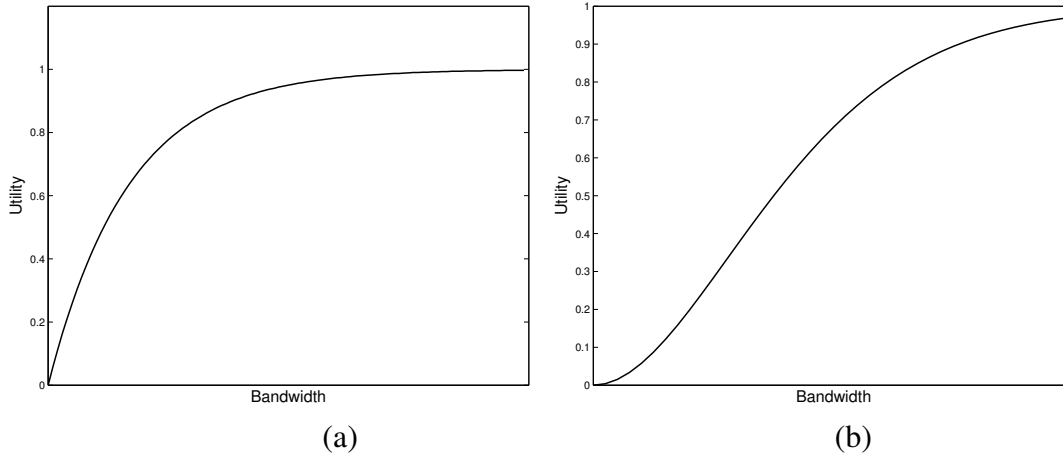
Figure 5.1: Utility functions of (a) elastic and (b) adaptive applications.

5.2   Bandwidth Utility and User Satisfaction Functions

The user satisfaction obtained from a network service depends on the performance of the network, which in turn depends on the capability of the network to satisfy resource (bandwidth) requirements of the user's request. The degree of fulfillment of the bandwidth requirement of a user application from the available bandwidth of a network is measured in terms of the bandwidth utility function (BUF).

5.2.1   Bandwidth Utility Function

The bandwidth requirement of a user's request depends on the application type. Based on the stringency of bandwidth requirements, applications can be divided into three categories: rigid, elastic, and adaptive. We follow the concept of rigid, elastic, and adaptive applications and the natures of their bandwidth utility functions from Shenker et al. [53, 54].

Rigid Applications: The real-time applications such as traditional telephony or voice calls that need a fixed amount of bandwidth to support them are classified as rigid applications. In other words, rigid applications are constant bit rate (CBR) applications whose data need to reach receivers within a given delay bound. If b is the network's available bandwidth, the utility function $U(b)$ of a rigid application is given by

$$U(b) = \begin{cases} 0 & \text{for } b < B_{min} \\ 1 & \text{otherwise} \end{cases} \tag{5.1}$$

where B_{min} is the minimum bandwidth required to support the application.

Elastic Applications: The traditional Internet applications like FTP, Web, e-mail, and Tel-

net that do not need a fixed amount of bandwidth to support them are classified as elastic applications. In other words, elastic applications are available bit rate (ABR) applications that can tolerate delays and follow the diminishing marginal rate of performance improvement as bandwidth increases [54]. This means that when the bandwidth is low, an increment of the bandwidth can increase the application quality by a greater amount than the same increment does at a high bandwidth. This type of utility function is strictly concave in nature. The utility function of an elastic application is shown in Fig. 5.1(a) and is expressed below.

$$U(b) = 1 - e^{-\frac{\alpha_e b}{B_{max}}} \tag{5.2}$$

where α_e is a scaling constant and B_{max} is the maximum bandwidth the application can utilize to improve the utility.

Adaptive Applications: Between the rigid and elastic applications there are adaptive applications such as delay-adaptive audio/video streaming and rate-adaptive scalable multimedia services. Delay-adaptive applications are insensitive to occasional delay bound violations and packet losses, whereas rate-adaptive applications are variable bit rate (VBR) applications that generate variable amounts of traffic depending on the network condition and can tolerate occasional bandwidth fluctuations. However, adaptive applications have intrinsic bandwidth requirements because they must maintain their data flow at or above some minimum level, below which the application quality suffers badly. For an available bandwidth of b, the utility function of an adaptive application is shown below.

$$U(b) = 1 - e^{-\frac{b^2}{\kappa+b}} \tag{5.3}$$

where κ is a constant. Figure 5.1(b) shows that the utility function of an adaptive application is convex but not concave in a neighborhood around zero. The point where the curve is convex depends on κ. The larger the value of κ, the larger the value of the bandwidth where the curve is convex. This means that the marginal utility of an additional bandwidth is slight when the bandwidth is very small or very large.

5.2.2   Effect of Mobility and Handover Latency on User Satisfaction

When a mobile user stays in a single network throughout a call duration, the user satisfaction from the communication service is equal to the bandwidth utility function of that network. However, if the user moves from one network to another while holding the call, the overall satisfaction is a function of the bandwidth utility of each network and the handover latency.

It is shown by Hidaka et al. that a user's satisfactory level decreases with response time t
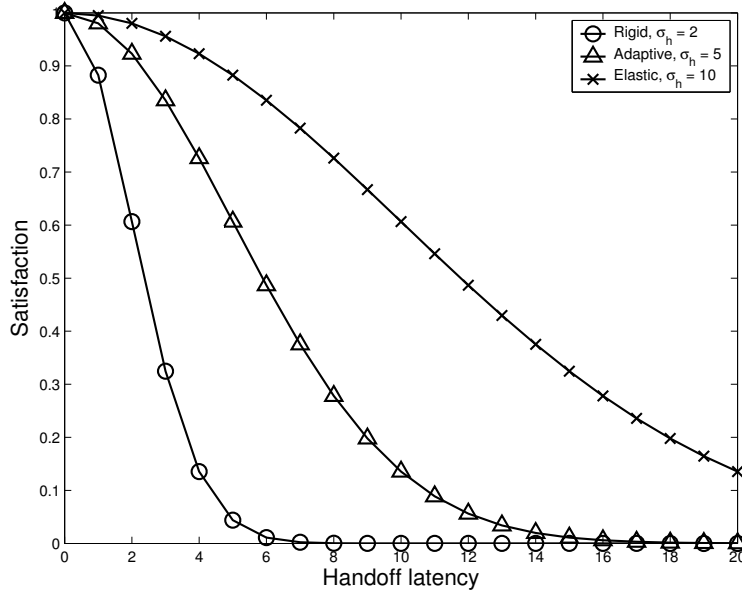
Figure 5.2: Satisfaction degradation of different applications as handover latency increases.

as $e^{-\frac{t^2}{2\sigma^2}}$, where σ is a constant that depends on application characteristics [55]. This can be interpreted, in other words, to mean that user satisfaction degrades with increasing handover latency. Therefore, $s_d(t_h)$, the factor by which satisfaction degrades due to the handover latency, t_h, can be expressed as below.

$$s_d(t_h) = e^{-\frac{t_h^2}{2\sigma_h^2}}$$
(5.4)

where σ_h is a constant that depends on application type; it is larger for non-real time (adaptive and elastic) and smaller for real time (rigid) applications. Figure 5.2 shows how user satisfaction degrades as the handover latency increases for different types of applications. It is clear that when the handover latency, t_h, is 0, the satisfaction value is 1, indicating that there is no satisfaction degradation at all when there is no handover latency. Any handover latency larger than zero results in satisfaction of less than 1. The figure also shows that satisfaction from a rigid application degrades more quickly than that from an adaptive or elastic application as the handover latency increases. For any value of σ_h satisfaction degrades by 60% when the handover latency is equal to σ_h.

Figure 5.3 shows the cellular configuration of the heterogeneous networks used for performance evaluation in this study. We considered two cellular networks: high bandwidth network (A) and low bandwidth network (B). When a mobile user is in the overlay area, it can connect to either of the networks. When the user moves out of the high-bandwidth overlay network, however, it has access only to the low bandwidth network.
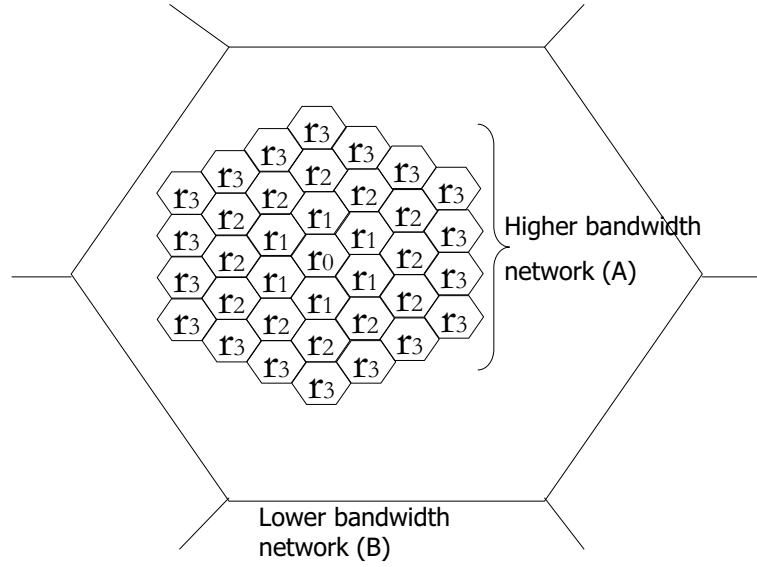
Figure 5.3: Cellular configuration of heterogeneous networks. Small hexagonal cells belong to high bandwidth network (A) and the big cell belongs to low bandwidth network (B).

Let $P_{A,B}$ be the probability that the user moves from network A to network B during a call, and $U(b_A)$ and $U(b_B)$ be the utilities from network A and B, respectively. Then, the effective value of user satisfaction, USF, is as expressed below.

$$
\begin{aligned}
USF &= f(U(b_A), U(b_B), s_d(t_h)) \\[2mm]
&= (1 - P_{A,B})U(b_A) + P_{A,B}[\frac{T_A U(b_A) + (T_c - T_A)U(b_B)}{T_c}]s_d(t_h) \quad (5.5)
\end{aligned}
$$

where T_A is the portion of call duration T_c spent in the high bandwidth overlay. We derive $P_{A,B}$ and T_A in the next section. The first term of Eq. (5.5) represents the user satisfaction when a call is completed in the high bandwidth overlay, whereas the second term represents the case when the call is handed over to the low bandwidth network and is completed there. More precisely, the second term is the time average of the bandwidth utilities, multiplied by the satisfaction degradation of handover latency.

The logic used to derive this equation is as follows. We know that the average of bandwidth utilities of both networks for a call duration (i.e., the content of the square brackets in Eq. (5.5)) ranges from 0 to 1, and so does the satisfaction degradation of handover latency. We also know that the smaller bandwidth and larger handover latency together affect the network performance more severely than does only one of these factors. Suppose the average bandwidth utility $U(b)$ is 0.5 and $s_d(t_h)$ is 0. In this case, the overall satisfaction must be zero because the zero value

of satisfaction of handover latency means that the latency is so long for an application that the application must be terminated. We assume that any application terminated unsuccessfully has the user satisfaction value of zero. Similarly, suppose that there are two cases, one with U(b) = 0.8 and $s_d(t_h)$ = 0.5, and the other with $U(b)$ = 0.5 and $s_d(t_h)$ = 0.8. Both the cases would yield equal overall satisfactions, considering that $U(b)$ and $s_d(t_h)$ are equally important and independent of each other. This explanation shows that the general tendency of the overall satisfaction is quite similar to a joint probability function, where the joint probability of two independent events happening together is equal to the product of the probabilities of the two individual events. Therefore, we have multiplied $U(b)$ and $s_d(t_h)$ to get the total satisfaction of the user who performs a handover from the high bandwidth to low bandwidth network.

We would like to note that our approach to defining the user satisfaction function as a function of the bandwidth utility and handover latency may not be the sole possible solution to assess user satisfaction in heterogeneous networks. Other solutions may be feasible as well. However, we believe that our evaluation method, with its moderately reasonable logic, is a novel start to defining and evaluating user satisfaction in heterogeneous overlay network systems.

The extreme values of satisfaction are 1 and 0, representing respectively that the user is fully satisfied and not satisfied at all. Any value between 1 and 0 represents the level of satisfaction the user gets from the network service for the given bandwidth and handover latency. The user satisfaction value from a rigid application, for example, will be 1 if the service is supported by both the high and low bandwidth networks and the handover latency is zero; in this case, there is no satisfaction degradation. However, if the handover latency is greater than zero, the overall satisfaction may be less than that would have been obtained from the use of only the low bandwidth network from the beginning of the call.

As the characteristics of adaptive applications lie between those of rigid and elastic applications, we take into consideration only the adaptive application for illustrating the effect of various conditions on the user satisfaction function.

5.3 Mobility Modeling

Different mobility models are used in literature to describe the behavior of mobile users in wireless/mobile networks, depending on the mobility characteristics of users. When mobile users are likely to move in a limited geographical area with frequent changes in direction and speed, the widely used random walk mobility model is suitable for modeling users' mobility in the network [56, 47].
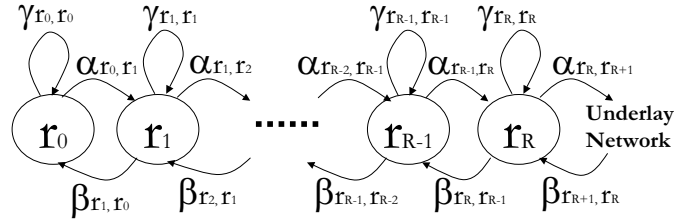
Figure 5.4: States diagram of random walk model.

5.3.1 Random Walk Mobility Model

In the case of overlay heterogeneous networks, a mobile user is more likely to be a pedestrian, moving mostly in a limited coverage area such as a campus, crowded market place, airport, or railway station. Based on this assumption, we consider the random walk model to be appropriate for this study.

As shown in Fig. 5.3, the hexagonal cells of the network are arranged in concentric rings. We arbitrarily represent the central ring, comprising a single cell, as r_0. The other rings surrounding the central ring are denoted by r_i with $i = 1, 2, 3...R$, where i is the distance of a ring r_i from the center in terms of the number of cells. The number of cells in a ring r_i is $6i$, and the total number of cells in a cellular network having R rings of hexagonal cells is $(1 + \sum_{i=1}^{R} 6i)$.

In the random walk mobility model, time is divided into discrete slots, and in each slot the mobile user moves out of the current cell with a given probability p (or may stay in the current cell with the probability $1 - p$) [56]. When the mobile user moves from the current cell, it has equal probability of entering any one of the adjacent cells. This means that in the hexagonal cellular network, during one time slot the mobile user can move to any of its six adjacent cells with a probability of $\frac{1}{6}p$. In what follows we derive the probabilities by which the mobile user moves from one ring to another.

5.3.1.1 Transition Probability

It is noted that when the user is in the innermost ring (r_0), a movement out of it always results in moving to the outer ring (r_1). However, when the user is in a cell of ring r_i, $1 \leq i \leq R$, a movement away from the cell can result in entering any adjacent cells belonging to the same ring, inner ring, or outer ring with different probabilities. It is shown by Ho et al. that the probabilities by which a movement of the user located in a ring r_i results in moving to a cell of the same ring, outer ring (r_{i+1}), and inner ring (r_{i-1}) are $\frac{1}{3}$, $\frac{1}{3} + \frac{1}{6i}$, and $\frac{1}{3} - \frac{1}{6i}$, respectively [56].

We use a Markov chain, shown in Fig. 5.4, to represent the location of the mobile user in the network. A state of the chain is equivalent to a ring of the network. The chain is said to be in a

state r_i if the user is currently located in ring r_i. We define three transition probabilities, $\alpha_{r_i,r_{i+1}}$, $\beta_{r_i,r_{i-1}}$, and γ_{r_i,r_i} to represent the probabilities that the user located in ring r_i moves to outer ring r_{i+1}, inner ring r_{i-1}, and remains in the current ring, respectively.

$$\alpha_{r_i,r_{i+1}} = \begin{cases} p & \text{for } i = 0 \\ p(\frac{1}{3} + \frac{1}{6i}) & \text{for } 1 \leq i \leq R \end{cases} \tag{5.6}$$

$$\beta_{r_i,r_{i-1}} = p(\frac{1}{3} - \frac{1}{6i}) \qquad \text{for } 1 \leq i \leq R \tag{5.7}$$

$$\gamma_{r_i,r_i} = \begin{cases} (1 - p) & i = 0 \\ (1 - \frac{2}{3}p) & \text{for } 1 \leq i \leq R. \end{cases} \tag{5.8}$$

5.3.1.2   Steady State Probability

Given the Markov chain in Fig. 5.4, the steady state probability π_{r_i} for state r_i ($1 \leq i \leq R$) can be derived in terms of steady state probability π_{r_0} of state r_0 as [36]:

$$\pi_{r_i} = \pi_{r_0} \prod_{j=1}^{i} \frac{\alpha_{r_{j-1},r_j}}{\beta_{r_j,r_{j-1}}} \qquad \text{for } 1 \leq i \leq R. \tag{5.9}$$

With the requirement $\sum_{i=0}^{R} \pi_{r_i} = 1$, we get

$$\pi_{r_0} = \frac{1}{1 + \sum_{i=1}^{R} \prod_{j=1}^{i} \frac{\alpha_{r_{j-1},r_j}}{\beta_{r_j,r_{j-1}}}}. \tag{5.10}$$

We use Eq. (5.9) to calculate the steady state probabilities of the mobile user residing in any ring r_i ($1 \leq i \leq R$). The larger the size of a network, the smaller the probability that the mobile user will be located in the outermost ring.

5.3.2   Call Transfer Probability

Call arrival is widely acknowledged to be a Poisson process. Although we use the term call arrival, we use it to refer both to a call originating from the mobile user as well as to a call addressed to the mobile user from any other correspondent user. It is assumed that the call holding time (i.e., duration of a call) and cell residence time follow exponential distributions with means of $1/\lambda_H$ and $1/\lambda_R$, respectively [56, 57].

Suppose that a call arrives when the mobile user is in the outermost ring of the high bandwidth network. The call has to be transferred to the low bandwidth network if the mobile user moves beyond the overlaid coverage while continuing the call. Therefore, the probability that

a call that arrives at the outermost ring of the high bandwidth network (i.e., network A) is transferred to the low bandwidth network (i.e., network B) is

$$P_{A,B} \quad = \quad P(\text{call holding time} > \text{cell residence time}) \times \ P(\text{user moves outward})$$

$$= \quad (\frac{\lambda_R}{\lambda_R + \lambda_H}) \times \alpha_{r_R, r_{R+1}} \tag{5.11}$$

Let X be the duration of time spent in the high bandwidth overlay before a handover occurs to the low bandwidth network. We suppose that a handover occurs randomly and X follows an exponential distribution as:

$$P(X \leq T_A) = 1 - e^{-\mu_h T_A} \tag{5.12}$$

where $1/\mu_h$ is the average duration of time until the handover occurs. Therefore, $P(X \leq T_A)$ is equivalent to the probability that a handover occurs at time T_A. Note that Eq. (5.11) also gives the probability of handover from the high to low bandwidth networks. Therefore, $P_{A,B} = P(X \leq T_A)$, and solving this we get T_A as:

$$T_A = -\frac{1}{\mu_h} log(1 - P_{A,B}) \tag{5.13}$$

5.3.3   Justification for Selecting the Model

As explained in Section 5.2, when a user moves within the same type of network, there will be no change in its satisfaction. Therefore, we are not concerned with the user movement behavior within a single network; rather we focus on the movement across the border of a high bandwidth network. To accomplish this, we have derived the user movement probability from a high bandwidth to low bandwidth network in Eq. (5.6) as $\alpha_{r_R, r_{R+1}}$.

Our analytical model can be extended to the cases other than the handover from a high bandwidth network to a low bandwidth one. For example, when we wish to extend the analytical method to the case of handover from the low bandwidth network to high bandwidth one in the overlay area, the boundary crossing probability in the reverse direction (from low to high bandwidth network) can be obtained by setting $i = R + 1$ in Eq. (5.7). As a result, the overall satisfaction function, as given by Eq. (5.5) with the subscripts 'A' and 'B' interchanged, will have some gains depending on the bandwidth utilities and handover latency.

However, because we are dealing with the problem of selecting the *best* access network by using currently available information on the mobile user's location, mobility, and call type, we think that it is more important (and also sufficient in our case) to consider only the movement

from the high bandwidth network to low bandwidth network.

5.4   Analysis and Consideration of Access Network Selection

In this section, we numerically investigate the effect that movement across heterogeneous overlay networks has on user satisfaction. Based on the analysis, we propose an access network selection algorithm that maximizes user satisfaction.

5.4.1   Numerical Analysis

We carry out numerical analysis by using the following parameters. The bandwidths available for a mobile user in the high and low bandwidth networks are assumed to be 2 units and 1 unit, respectively. The unit of bandwidth can be 100 kbps or Mbps. The number of rings in the high bandwidth overlay is taken as 5, so the total number of cells in the overlay is 61. Similarly, σ of the utility degradation factor due to handover latency is taken as 5, and the average cell residence time $(1/\lambda_R)$ is 10 minutes. We discuss the impact of various factors on the user satisfaction function by assuming two types of movements: uniform random movement (as described by the random walk model in the previous section) and directional movement. The use of the uniform random model provides the notion of aggregate satisfaction degradation of all the users, who use the high bandwidth network in the overlay area and move to the low bandwidth network while holding a call. On the other hand, in the directional movement we use a prediction that mobile users move in a given direction with a higher probability. This movement is used to show the performance degradation of the users who stay in the high bandwidth network for a very short time and move out of it while continuing calls.

Figure 5.5 shows the satisfaction of a mobile user located in the outermost ring of the overlay network using the random walk model. The figure shows that user satisfaction decreases as the call holding time or handover latency increases. This is because as the call holding time increases, the probability that the user moves out of the overlay network while holding a call increases, and the overall satisfaction decreases due to the handover latency and the lower bandwidth of the new network. Under the given parameters the user satisfaction values are 0.78 and 0.46 if the user uses only the high bandwidth network and only the low bandwidth one, respectively. In other words, these numbers represent the bandwidth utility values of the high and low bandwidth networks, respectively. When the call holding time is 1 minute, the probability that a call is transferred to the low bandwidth network is about 1.7% (as given by Eq. (5.11)) which is negligible, and even when the call holding time is 20 minutes, the probability of call transfer rises only to 12.5%. Because the probability of a call transfer does not increase much when

Figure 5.5: User satisfaction from adaptive application of a mobile user located at outermost ring of high bandwidth network, with random-movement.

the call holding time increases, its impact on user satisfaction is not significant. However, the satisfaction values shown by this figure are the average of satisfactions of all the users, of which some migrate to the low bandwidth network and some do not. For those users who are near the border of the overlay network and who migrate and spend most of their call holding time in the low bandwidth network, their satisfaction degrades significantly, which cannot be evaluated by this uniform movement-based analytical model.

To capture the impact on satisfaction of mobile users that leave the high bandwidth overlay soon, we need to gather information about the users' mobility patterns. Suppose that by any means (such as [58, 59]) we could know that the mobile user is most probable (say by 90%) to move out of the overlay network. In this case, depending on the values of cell residence time and call holding time, the mobile user will spend a variable amount of time in the high bandwidth and low bandwidth networks while holding the call. With the mean value of cell residence time of 10 minutes, the probability that the call is transferred to the low bandwidth network is about 8% and 60% when the call holding times are 1 and 20 minutes, respectively (as given by Eq. (5.11)). In such cases, Fig. 5.6 shows that user satisfaction decreases significantly as the call holding time increases. As before, user satisfaction obtained from the sole use of the high bandwidth and low bandwidth networks are 0.78 and 0.46, respectively. However, when using the high bandwidth network first and then migrating the call to the low bandwidth network, the user satisfaction value for longer call holding time is less than 0.46, and the deterioration is
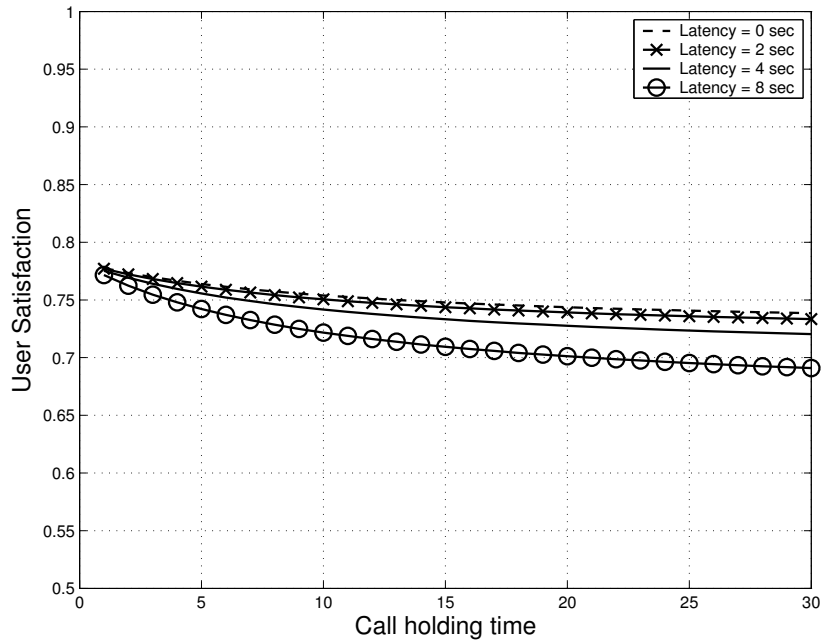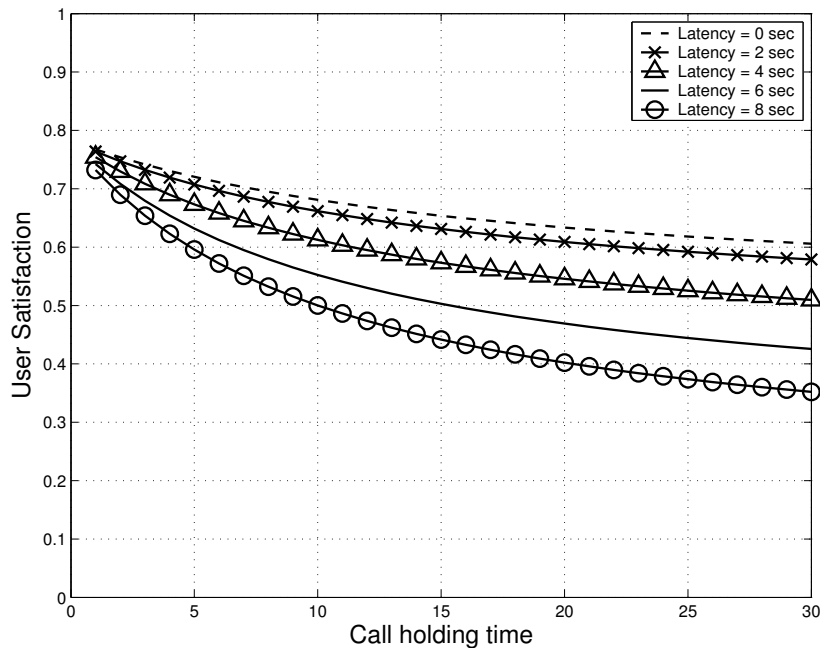
Figure 5.6: User satisfaction from adaptive application of a mobile user located at outermost ring of high bandwidth network, with 90% movement probability to low bandwidth network.

more prominent when the handover latency is high (higher than 4 seconds). The user would have gotten better satisfaction (0.46) if the low bandwidth network was assigned to the user from the beginning of the call.

Table 5.1 presents the values of the bandwidth utility and effective user satisfaction for different available bandwidths in the high and low bandwidth networks. USF(rand) and USF(90%) represent the effective user satisfactions for the random movement and directional movement with a 90% moving out probability, respectively. Here, the call holding time is 10 minutes, the cell residence time is 20 minutes, and the handover latency is 4 seconds. The first row shows that when the available bandwidths of the low and high bandwidth networks are 4 and 1 unit, respectively, there is a vast contrast in the bandwidth utilities of these two networks. The user satisfaction remains larger than the utility of the low bandwidth network in both the random and directional movements. The second row shows the case when the available bandwidths are 4 and 3 units. In this case, although the bandwidth utilities of both the networks are comparable, the effective user satisfaction is equal to or less than the bandwidth utility of the low bandwidth network due to the satisfaction degradation by handover latency.

We thus conclude that if the bandwidth of the low bandwidth network is slightly lower than that of the high bandwidth network in an overlay area, selection of the low bandwidth network that is available for a longer time gives better user satisfaction.

Table 5.1: User satisfactions.

b_A	b_B	$U(b_A)$	$U(b_B)$	USF(rand)	USF(90%)
4	1	0.96	0.46	0.88	0.59
4	3	0.96	0.91	0.91	0.72

5.4.2 Access Network Selection Algorithm

It is clear from the analyses of the results in the previous subsection that merely assigning an access network with higher bandwidth does not necessarily lead to the improved user satisfaction when the users are likely to perform a vertical handover to a lower bandwidth network soon after they begin a communication session. In such cases, the users would get better satisfaction using a lower bandwidth access network for a longer time. Based on this observation, we present the following algorithm for selecting an optimal access network in heterogeneous overlay network environments.

```
When a call request arrives:
1  predict user's mobility pattern
2  predict call holding time
3  list available access networks
4  sort networks in the descending
   order of their bandwidths,
   N(i), i =1,2,...,M}
5  consider network N(i), i = 1
6  estimate user's movement
   probability away from N(i)
7  estimate handover latency from
   N(i) to N(i+1)
8  estimate user satisfaction, USF(i)
9  IF USF(i) > BUF(i+1)
    select network N(i)
   ELSE
    i = i + 1
    consider network N(i)
    go to 6
   END IF
```

As shown above, when a call request from/to a mobile user arrives, the network selection mechanism first predicts the user's mobility pattern and call holding time (how it predicts is described in the following subsection). Then, it lists all the access networks that are available from the user's current location and can serve the call request. This list is then sorted in the descending order of bandwidths. From the list, the network with highest bandwidth is considered first. Based on the network layout and user's mobility pattern, the probability that the user moves from the network under consideration is estimated. With this probability we estimate the effective user satisfaction and compare it with the bandwidth utility function of the next network in the list. If the user satisfaction from the use of the current network is larger than the bandwidth utility function (BUF) of the next network, we select the current network. Otherwise, we choose the next network from the list and repeat the above steps to find an access network that provides optimal user satisfaction.

The above algorithm assumes that a number of user-related parameters (such as mobility pattern and call holding time) as well as network-related parameters (such as available bandwidth and handover latency) are available to evaluate the user satisfaction function. We are going to explain a few techniques for estimating these parameters in the following subsection.

5.4.3   Techniques for Parameter Estimation

In a real situation, user mobility has a high degree of predictability due to temporal and spatial localities [58]. The temporal locality indicates that many mobile users stay in or pass through an almost predictable set of cells during certain times. For example, many users remain at home early in the morning, then go to work and come back home along the same route in the evening, and stay at home until the morning of the next day. Consequently, we can predict the location and movement of a user based on the time of a day. Similarly, the spatial locality indicates that the paths along which a user can move are governed by geographical constraints, such as pathways, roads, or highways in the vicinity of the current location. The temporal and spatial locality information aggregated over time is referred to as the user mobility history. Then, some algorithms, such as one proposed in [59], are applied to the user mobility history to derive the most probable user mobility pattern (UMP). The UMP contains a list of cells/networks expected to be visited by mobile users starting from a given location/time.

Likewise, the service duration can be predicted in accordance with the service type, location, and time [60, 61]. Users access different types of services depending on their location and time. The mobile users commuting by the subways in Tokyo, for instance, mostly use only e-mail or web services, but do not use the voice service. Accordingly, different services have different call durations; video conferencing takes a longer amount of time than e-mail downloading, for

instance. If the user is downloading or streaming a video, the call duration depends on the file size and the average bandwidth of the network. Moreover, the call duration of the same type of services depends on the location/time of the service initiation and the correspondent host.

The user mobility history can be extended to include the type and duration of services accessed by the user at a given location or time [60]. Then, similar to the derivation of user mobility pattern, a user service (call) duration model would be constructed from the mobility history by using some statistical or heuristic approaches. Also, for some services, there are widely accepted call duration distributions such as an exponential distribution for voice services [56]. In such cases, we need to estimate only the mean call duration with respect to the user location and time coordinates. When making a service request, the user might tell the network selector what type of request it is making: telephony, web data, video, etc., so that the call duration can be predicted for that type of service, depending on the model derived from the history for the current location, time, and correspondent host.

We now describe the issues related to bandwidth estimation. In cellular networks, establishment of a connection guarantees the availability of a fixed amount of bandwidth for the whole connection duration. However, in a shared network such as wireless LAN, the available bandwidth depends on the number of users and their activities. In such cases, the bandwidth allocatable to a user depends on the bandwidth allocation policy [62], which guides how the network bandwidth is shared among users. The simplest scheme is to share the bandwidth equally among all users. That is, if B is the total bandwidth of a network shared by n number of users, the available bandwidth for each user is simply B/n. This scheme, however, does not work well when users use different types of applications requiring different amount of bandwidths. Other schemes can be used, such as sharing a portion of bandwidth among the users of the same type of services. In this scheme, the available bandwidth for a new user of a particular type of service is $b = B_i/n_i$, where B_i is the bandwidth share assigned to all the users belonging to service type i, and n_i is the number of users currently using the service, including the new one. B_i can be fixed or dynamically adjusted depending on user activities.

However, the bandwidth allocation mechanisms described above have a problem: they allow the bandwidth of already-connected users to fluctuate when a new user joins or an old user leaves the network. This problem can be solved by using an admission control and resource reservation [63] method. Admission control prevents new users from accessing the network in order to maintain the performance quality of the existing users. Similarly, a bandwidth reservation scheme assures a connection guarantee to some classes of services, irrespective of the users' movement across different cells of the network. Both of these mechanisms combined provide a notion that a connection once admitted would get the promised amount of bandwidth throughout the call duration. The promised bandwidth might be specified in terms of the sta-

tistical mean or worst-case minimum bandwidth. It is very logical to say that future shared networks will use these QoS assurance mechanisms because they are expected to support real-time applications. For instance, the wireless LAN is in the process of enhancement, in the form of the 802.11e standard, to support and maintain the network service quality.

5.5 Conclusion

In this chapter, our goal was to find a way to maximize user satisfaction in heterogeneous wireless overlay networks. We described the bandwidth utility functions for rigid, elastic, and adaptive applications; we used this bandwidth utility function and handover latency to derive the user satisfaction function. We found that selecting a high bandwidth access network does not guarantee higher satisfaction if the user happens to perform a handover to a lower bandwidth network. It is observed that after getting an estimate of the user's movement, we should assign the network with most availability to the user's call request, so that the user can remain in the same network throughout the call duration. By doing this, we can prevent the user satisfaction from being degraded. In accordance with the evaluation, we proposed an access network selection algorithm that maximizes user satisfaction in heterogeneous networks.

Chapter 6

Graceful Vertical Handover

6.1 Introduction

This chapter presents a new scheme for upward vertical handover management, which we call the *graceful vertical handover* scheme. This scheme provides graceful degradation of performance by smoothing the throughput change and reducing the possible packet losses during a handover. To achieve such a gracefulness in performance change, this scheme implements its functionalities in the link layer, IP layer, and TCP or application layer.

This chapter is organized as follows. Section 6.2 gives an overview of the problem statement. Section 6.3 presents the proposed graceful vertical handover scheme. Section 6.4 outlines the simulation setup and discusses performance evaluation results. Finally, Section 6.5 concludes this chapter.

6.2 Problem Statement

6.2.1 Non-Seamless Vertical Handover

The non-seamless nature of an upward vertical handover can be explained by referring to Fig. 6.1. The curve (a) shows the throughput change when a mobile node (MN) suddenly looses its connection with a wireless local area network (WLAN) and starts accessing a wireless wide area network (WWAN), such as the 3G network. This is a possible scenario in wireless communications because the wireless channel quality is time and location dependent. If an obstacle, such as a building or cliff, happens to come in between the MN and the WLAN access point (AP), the connection between them is abruptly lost. In such cases, the ongoing communication session of the MN is interrupted for the handover duration (t_1-t_0), and resumed once the MN
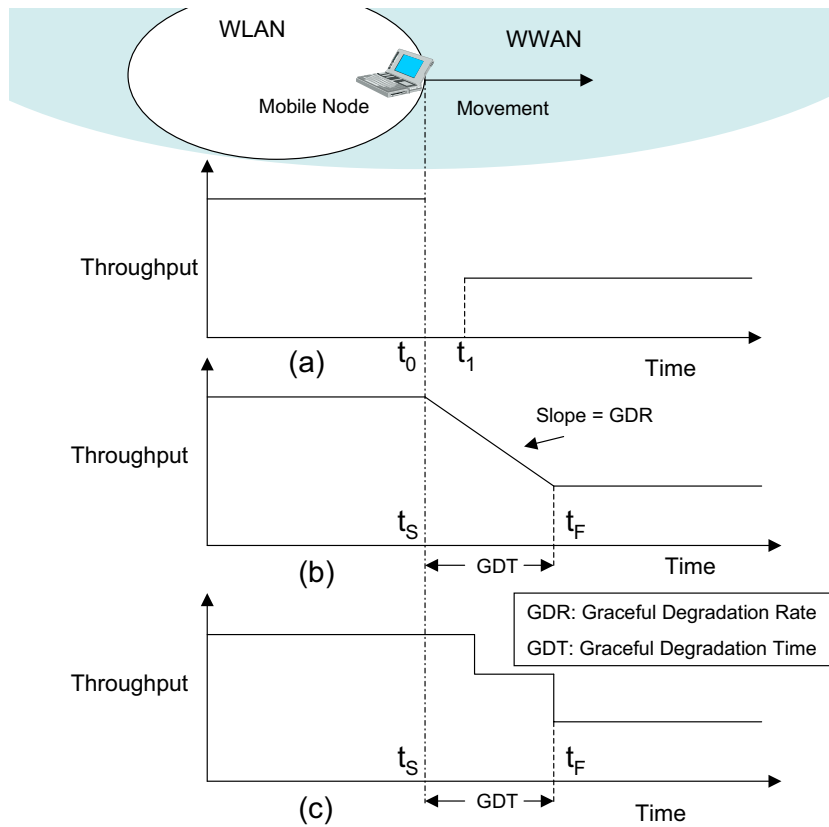
Figure 6.1: Throughput change during vertical handover when WLAN becomes suddenly inaccessible; (a) traditional handover, (b) graceful handover of TCP data application, and (c) graceful handover of adaptive multimedia application.

gets connected to the WWAN. Moreover, a large number of packets may be lost at the AP. This is one of the prohibitively worst situations, as it forces the service to be disrupted for some time and consequently affects user satisfaction obtained from the network service [11].

6.2.2 Need of Graceful Handover

The vertical handover fundamentally differs from the horizontal handover in several issues [66]. However, as in the horizontal handover, most of the previous researches on vertical handover are conducted with the purpose of reducing the handover latency [64] and packet losses [52]. We argue that the core difference between the horizontal and vertical handovers is the need of performing a handover. A mobile user performs a horizontal handover to a network that provides a better performance than (or at least the same performance as) the current one at the given location, whereas a mobile user performs an upward vertical handover to a lower performance network, when the better one becomes unavailable at the given location.

Therefore, it is natural to expect that mobile users wish to stay at a high performance network as much as possible, while preparing for an upward vertical handover to a low performance network. It is also logical to expect that mobile users prefer the gradual performance change with no disruption of network services during a handover. Some users even feel good to have them informed or interrogated in advance about the inevitable changes in the service quality after handover.

The above explanations clearly indicate that the longer vertical handover latency and gradual change of service quality are desirable for the better performance of the network service during an upward vertical handover. Taking into consideration these facts, we propose a graceful vertical handover scheme. The graceful handover is designed to achieve these objectives as depicted by curves (b) and (c) in Fig. 6.1 for two types of applications: TCP data applications and adaptive multimedia applications, respectively. For TCP applications, it is possible to gradually decrease the data rate without severely affecting the performance. The curve (b) shows that the data rate of the TCP application is decreased at a graceful degradation rate (GDR) over the graceful degradation time (GDT). The GDT is the graceful handover latency, which starts from the instance (indicated by t_S) when the data rate begins decreasing to the instance (indicated by t_F) when the user gets data packets from the low performance network. On the other hand, because multimedia applications inherently require a minimum bandwidth to maintain their quality, their data rates cannot be decreased gradually. In such cases, as depicted by curve (c), the data rate can be rather decreased in steps by adjusting the application to the lower quality or type. For instance, a high quality videophone service can be stepped down to a low quality videophone or voice phone. Note that unlike in a horizontal handover the longer handover la-

tency of the graceful handover scheme does not cause any packet loss. Instead, it reduces the packet loss and smoothens the throughput change during the handover time.

6.3   Description of Graceful Handover Scheme

In this section, we describe a scheme that attains a graceful handover in the heterogeneous networks consisting of WLAN and WWAN overlays. In particular, we exploit the infrastructure and ad hoc modes of operations of the WLAN system [67, 68]. In the infrastructure mode, MNs communicate through a fixed AP, whereas in the ad hoc mode, they communicate among themselves without going through the AP. The network layout used to describe our approach to a graceful handover is shown in Fig. 6.2. Here, both the WLAN and WWAN are available inside the circle, whereas only the WWAN is available in the areas outside the circle. When the MN is abruptly disconnected from an AP and no other APs are available from the MN's current location, the MN immediately enters the ad hoc mode and selects an intermediate node, called a relay node, to forward its data to/from the AP. After resuming a connection to the AP via the relay node (RN), the MN gradually performs a graceful handover to the WWAN. We use the Mobile IPv6 protocol [4] for managing mobility of the MN. Implementation of the graceful handover mechanism requires modifications to the operations of the link layer, IP layer, and transport or application layer, which we describe one by one in the following subsections.

6.3.1   Link-layer Operation

When the MN is in the WLAN coverage, it has a direct association with an AP, as shown by arrow (1) in Fig. 6.2. When the MN goes out of the WLAN's transmission range, it initiates a relay node search mechanism to find the best suitable RN for relaying its packets to/from the AP. Either of the following two modes of the RN search mechanism can be used.

1. Proactive mode: The operation of the proactive mode is shown in Fig. 6.3. In this mode, the RN periodically advertises RelayAdvertisement messages containing information about the bandwidth and intended relay duration in the network. We assume that an RN estimates the bandwidth between itself and the AP according to some heuristic method, such as the one described by Lee et al. [67]. The MNs in the vicinity cache the RelayAdvertisement messages and use it to select an RN when they move out of the WLAN coverage. In case there are many RNs, the MN may use an algorithm to determine the best RN based on the parameters, such as the available bandwidth, location of RNs, and the intended duration of the relay. After selecting an RN, the MN sends a RelayRequest message to the RN. Upon receiving the RelayRequest, the RN consults the AP to check whether the MN is an authorized one or not, and then replies
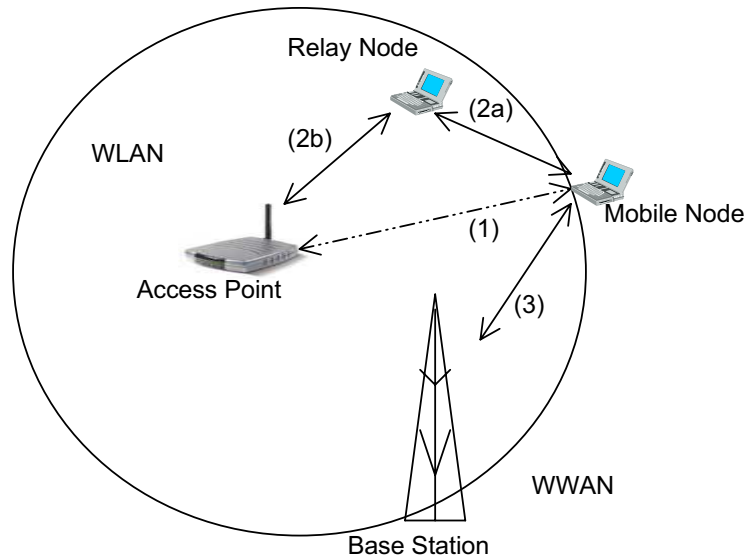
Figure 6.2: Operations of graceful vertical handover. Arrow (1) shows the direct link between mobile node and access point, arrows (2a) and (2b) show indirect links between mobile node and access point via relay node, and arrow (3) shows the connection between mobile node and base station after the handover.
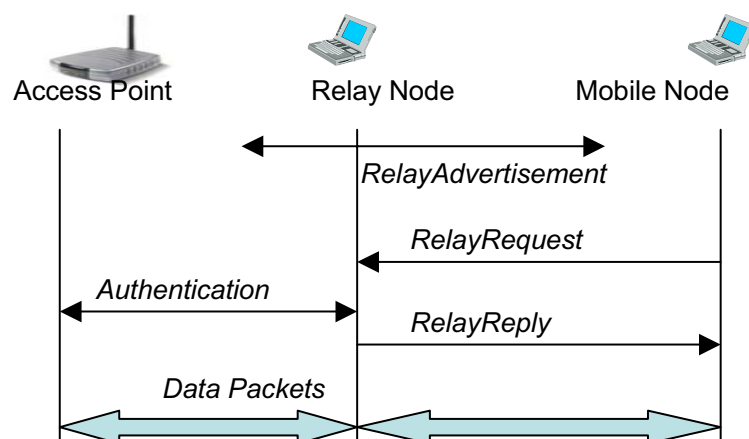

Figure 6.3: Relay node search mechanism: proactive mode.

to the MN with a RelayReply message.

2. Reactive mode: In the reactive mode, the RNs do not broadcast the unsolicited RelayAdvertisement, but send it to an MN that solicits the relaying function by broadcasting a RelaySolicitation message. All interested RNs reply with a RelayAdvertisement message, from which the MN selects the best RN. The remaining process is the same as that in the proactive mode.

Both of these modes of operation have advantages as well as disadvantages; the proactive mode increases the signaling traffic in the network, whereas the reactive mode increases the relay setup latency.

Some of the MNs that have abundant resources, such as battery power and available bandwidth, can act as relay nodes. Since an AP has an association with all the MNs located in its coverage, the AP can coordinate/manage the population of the RNs. An MN willing to function as an RN may send a request to the AP, which may accept or reject the request based on the information provided, such as the location, battery power, and channel condition. There may be a multi-hop relay system having an array of relay nodes in the communication path between the MN and the AP. However, it has been shown by Lee et al. [67] that single hop relay paths are sufficient for performance benefits, and the benefits of additional relay nodes on the path are similar or marginally worse because of the signaling overheads and channel interferences involved in managing the multi-hop relay system.

6.3.2   IP-layer Operation

After getting indirectly associated with the AP via an RN, the MN gradually starts the handover to the WWAN using the Mobile IPv6 operation [4]. The handover process is composed of two tasks: (a) IPv6 address configuration and (b) binding update. The MN configures an IPv6 care-of address from the network prefix advertised in the WWAN. After that, the MN performs a binding update (BU) with its home agent and correspondent nodes (CNs). We introduce an additional field, a graceful degradation time (GDT), in the BU message of the Mobile IPv6 protocol. This field indicates that the new care-of address is activated only after the GDT expires, so that the CN and home agent keep on sending data packets to the old care-of address until the GDT expires. If needed, the MN can update the CN or home agent on the GDT by sending another BU containing a new GDT value. The MN sends an update when it detects that it is moving faster or slower than the time indicated in the previous BU. Additionally, if no data packets arrive from the MN via its previous care-of address for some time, no matter what the GDT value is, the home agent and the CN immediately activate the new care-of address. This is done to minimize the data loss at the old AP, assuming that the MN has lost its connectivity

with the AP even via the RN.

6.3.3   Transport- or Application-layer Operation

The graceful vertical handover allows the throughput to degrade smoothly at a graceful degra-
dation rate (GDR) for the duration of the GDT. The GDR can be implemented either in the
transport layer or the application layer, depending on the application types as described below.

6.3.3.1   TCP Data Applications

TCP protocol employs a receiver-initiated flow control mechanism. For this purpose, the TCP
acknowledgement (Ack) packet has an "advertised window size" (*awnd*) field, which corre-
sponds to the amount of buffer space available in the receiver for additional data. To achieve a
graceful degradation of the TCP throughput during a vertical handover, we set the values of the
awnd to gradually decreasing numbers in such a way that after the GDT the congestion-window
size (*cwnd*) of the CN is reduced to the size proportional to the bandwidth of the WWAN. In pub-
lished literature, some similar approaches, such as Freeze-TCP [69] and TCP Westwood [70],
have been presented to tackle the problems of network heterogeneity. In Freeze-TCP, when a
mobile node is disconnected from a network, it sends an Ack with a zero *awnd* size. Upon
receiving such an Ack, the sender freezes all timers and interrupts the transmission until it re-
ceives an Ack signal with a non-zero *awnd*. Although this scheme reduces the packet losses, it
cannot smoothen the throughput during the handover. TCP Westwood is a small improvement
over the Freeze-TCP, in the sense that it estimates the available bandwidth to set the values of
the *cwnd* and slow start threshold. However, it is a sender-side modified scheme, and thus takes
a longer time to detect the change in the bandwidth of the access networks that happens in the
receiver side. We address this problem by having the receiver promptly detect the bandwidth
change and inform the sender.

   To address the above problems, we slightly modify the TCP implementations in the sender
and receiver sides. With this modification, not only a sender but also a receiver regularly esti-
mates the round-trip time (RTT) by sending some dummy packets piggibacked on Acks to the
sender. This RTT can further be divided into two parts: the time taken in the fixed (i.e., wired
part of) network and the time taken in the wireless network. If we assume that the round-trip
transmission time in the fixed network (RTT_f) is almost constant, the change in RTT due to the
change of the wireless links is mostly dominated by the transmission time in the wireless seg-
ment. That is, RTT = [RTT_f + (Dummy + Ack packet size)/wireless bandwidth]. The MN uses
the RTT to infer the value of the sender's (i.e., CN's) *cwnd* according to the following expres-
sion: *cwnd* = RTT*wireless bandwidth, assuming that the wireless bandwidth is the bottleneck.

During a handover, the MN estimates the expected RTT for the new network as $RTT_n = [RTT_f$ + (Dummy + Ack packet size)/ wireless bandwidth of new network], and uses it to estimate the expected value of the *cwnd* for the new network as: $cwnd_n = [RTT_n*$wireless bandwidth of new network]. To achieve the graceful degradation of the throughput, the MN then estimates the required reduction rate of the *cwnd* size as ($cwnd$ - $cwnd_n$)/GDT, and accordingly sets the value of the *awnd* in the consequent Ack packets. Upon receiving such Acks, the CN transmits the additional data segments as indicated by the *awnd* size. When the CN continuously receives a number of Acks with decreasing *awnd* values, it interrupts the TCP's normal operation by continuously setting the *cwnd* values equal to the received *awnd* values until it stops receiving such Acks, i.e., throughout the handover duration. When the handover process is complete, the MN's TCP operates normally, i.e., the *awnd* would be used to indicate the available buffer size, not the wireless bandwidth. Consequently, upon receiving a normal Ack (that usually has a higher *awnd* value) the CN sets the slow start threshold as half of its current *cwnd*, and then works normally as governed by the TCP protocol.

In this way, controlling the size of the congestion window during a handover would result in both the graceful degradation of the throughout and avoidance of packet losses.

6.3.3.2   Multimedia Streaming Applications

Multimedia applications, such as voice and video streaming, may not use TCP as a transport protocol; they use the UDP protocol instead, which does not employ a flow control mechanism. In addition, the vertical handover may trigger to change the media or application type to adapt to the capacity of the new network. In such cases, it is appropriate to use an application-layer signaling message to inform the CN of the desired application type suitable for the new network. For this purpose, the MN forms a signaling message and sends it to the CN. After getting the message, a media gateway or adaptation function of the CN initiates the new application.

6.4   Performance Evaluation

We have implemented the graceful handover mechanism in the ns-2 simulator [71]. We used a Mobiwan [72] patch to support the Mobile IPv6 protocol in the ns-2 simulator. We further extended it to enable the MN to support multiple interfaces for communicating with different access networks that provide different bandwidth services. In this work, we focused on the study of the TCP performance during a vertical handover from a high-bandwidth WLAN to a low-bandwidth WWAN, while the study of the multimedia application's performance will be carried out in our future work. The network layout used for the performance evaluation is shown
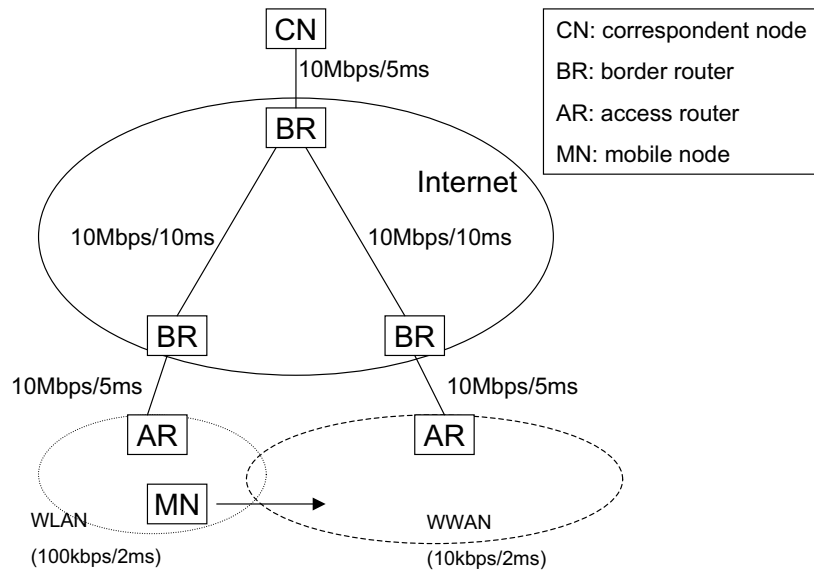
Figure 6.4: Network layout for simulation study.

in Fig. 6.4. The bandwidth and delay in each segment were set to the values as indicated in the figure. To enable us to trace the progress in the TCP packet sequence number, we used relatively low-bandwidth access networks (WLAN with 100 kbps and WWAN with 10 kbps). In addition, the buffer size of the AR's ingress interface is set to a low value of 5 packets. Setting a larger buffer size may help reduce the number of packet losses, however, at the cost of increasing delays in packet delivery. With this setup, we compared the performances of TCP Reno without and with graceful handover. In the both cases, the CN started a TCP session at 0.5 second from the beginning of the simulation, and, while having an active TCP session, the MN performed handover from the WLAN to the WWAN at 25th second.

Figure 6.5 shows the progress of the TCP sequence number of packets at the CN (i.e, the sender node) and the throughput observed by the MN at the TCP layer without the graceful handover mechanism. It shows that the throughput reduced to almost zero during the handover. If we enlarge this figure around the handover time, we get more insight into the effects of the handover on the TCP performance, shown in Fig. 6.6. This figure also includes the sequence number of the packets received by the MN and the corresponding acknowledgements. It shows that during the handover the packets that had already been sent by the CN but had not yet been received by the MN were lost in the network. Moreover, the CN kept sending packets to the old address of the MN until the CN received a BU message from the MN. The handover delay was about one second. After the handover, the performance further deteriorated because the CN had to wait for a TCP retransmission timeout (RTO) to resend the lost packets. The causes of the timeout were the injection of a large number of packets into the low-bandwidth WWAN, and
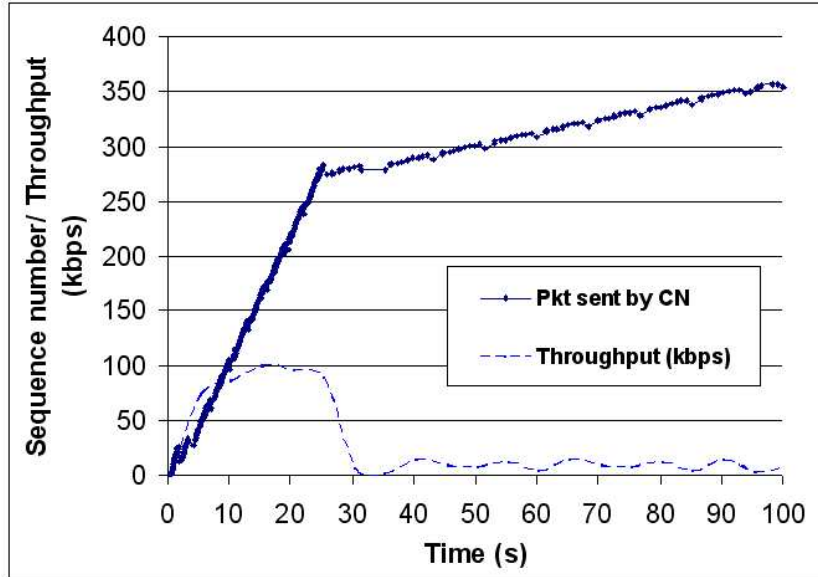
Figure 6.5: Progress in TCP packet sequence number at CN, and throughput observed by MN in general upward vertical handover (i.e, without graceful handover). The MN performed a handover after 25 seconds.
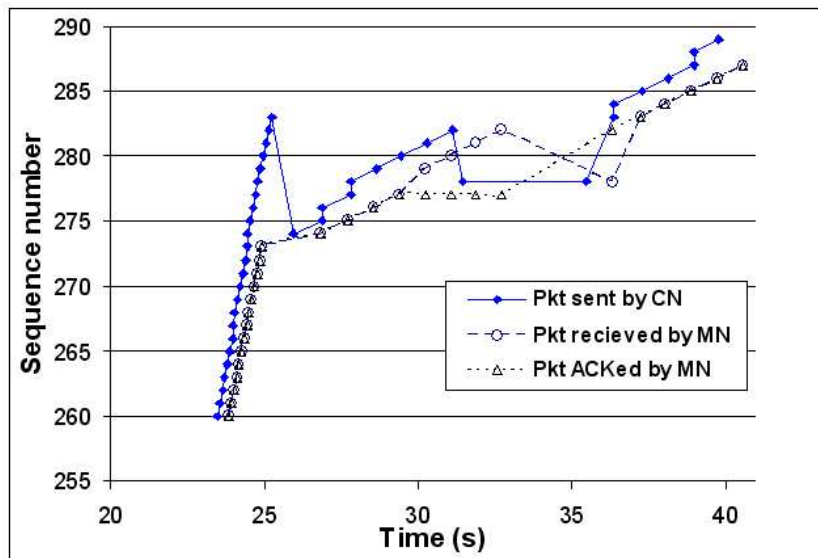

Figure 6.6: Sequence numbers of packets sent by CN, and received and acknowledged by MN in upward vertical handover (i.e., without graceful handover).
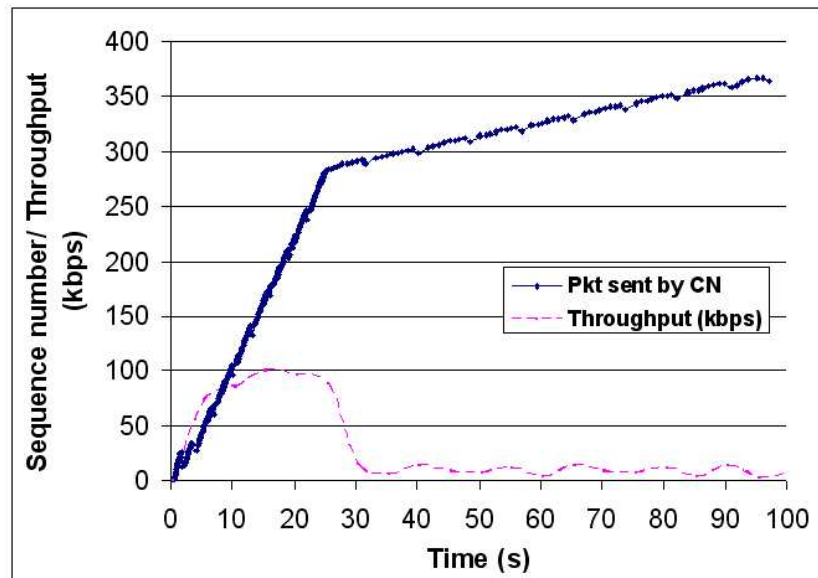
Figure 6.7: Progress in TCP packet sequence number at CN, and throughput observed by MN in graceful vertical handover. The MN performed a handover after 25 s.

consequently repeated losses in the network. The CN sent many packets because it used the slow-start threshold value that was estimated before the handover for matching the capacity of the relatively high-bandwidth WLAN.

Based on the above evaluation, we can infer that, for a smooth handover, we need to control the congestion window size according to the capacity of the new network, i.e., the WWAN. The proposed graceful handover scheme does control the *cwnd* size and improves the TCP performance as described in the following paragraphs.

To evaluate the performance of the graceful handover mechanism, we used the proactive mode of the relay node search mechanism. Additionally, to simplify the mechanism, we designated some nodes as relay nodes in advance, so that the MN knew its relay node while it was still connected to the WLAN AP.

Figures 6.7 and 6.8 show the TCP performance of the graceful handover scheme. In this case too, the direct connection between the MN and the WLAN AP was broken at 25th second, and the MN then started the graceful handover operation by searching for an RN. The search mechanism did not take a longer time because of the availability of the pre-designated relay node. After establishing a connection to the AP via the RN, the MN obtained its packets accumulated at the AP. The MN then started the graceful operation of TCP by estimating the expected values of the *cwnd* for the WLAN and WWAN, and accordingly setting the *awnd* in Ack packets. Upon receiving such Acks the CN modified its *cwnd*, thus limiting the data packets injected into the network. The figure shows that the modification of the *cwnd* of the CN
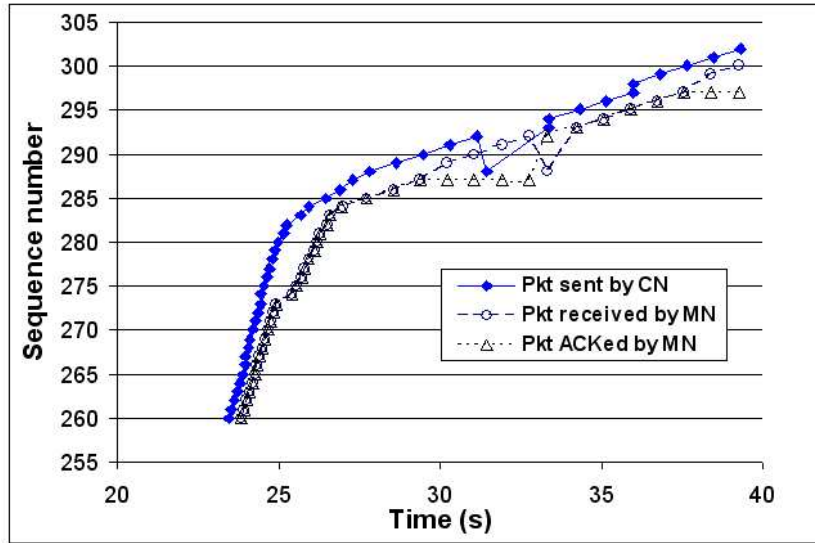
Figure 6.8: Sequence numbers of packets sent by CN, and received and acknowledged by MN in graceful vertical handover.

resulted in reducing the packet loss as well as smoothening the TCP throughput during the handover time (from 25 to 29 seconds). The graceful operation of TCP terminated at 29th second when the MN left the WLAN and got connected to the WWAN. In this case, there were no TCP retransmission timeouts during/after the handover. Thus, it is clear that the graceful handover scheme reduces the packet loss, while improving and smoothening the throughput.

6.5   Conclusion

In this chapter, we presented a graceful vertical handover scheme that provided graceful performance degradation during an upward vertical handover. Smoothening the throughput change and reducing the packet losses during a handover by modifying the operations of the link and higher layers produced the graceful handover. Through simulations, we demonstrated the improved TCP performance of the graceful handover.

Chapter 7

Summary and Future Work

7.1 Summary

In the traditional homogeneous wireless networks, mobility management operations are carried out by the link-layer specific mechanisms, which are usually not portable from one type of network to another. Therefore, mobility management in the newly emerging heterogeneous wireless networks is becoming a challenging issue for developing a ubiquitous computing system. An efficient mobility management scheme allows a mobile device to have always best connected (ABC) to the Internet. It facilitates the mobile device to connect to different types of access networks that optimally satisfy the resource requirements of desired communication services. In this dissertation, we addressed the mobility management issues of two types of access network environments: network mobility and heterogeneous overlapped networks. We optimized network mobility management by proposing new schemes for route optimization and handover management. For the heterogeneous overlapped networking environment, we addressed the problem of optimal network selection by defining a user satisfaction function that incorporates the user application requirements and network service characteristics. We also proposed a graceful vertical handover mechanism.

Network mobility (NEMO) management is concerned with managing mobility of an entire network, called mobile network. The mobile networks, or moving networks, are being designed to facilitate mobile devices to seamlessly access Internet by grouping the mobile devices and hiding their mobility on the Internet topology as long as they stay in the mobile networks. The NEMO Basic Support protocol, developed by the Internet Engineering Task Force (IETF), outlines the basic operation of mobility management of mobile networks. This protocol reduces the volume of signaling messages flowing out of the mobile network, however, at the cost of increasing overheads and delays in data packets delivery. We reduced these overheads and de-

lays by developing a mobile router assisted route optimization (MoRaRo) scheme. In addition, we proposed a cooperative mobile router-based handover optimization (CoMoRoHo) scheme to carry out a seamless handover of mobile networks.

The MoRaRo scheme is simple to implement because it requires only a slight change in the implementation of the NEMO basic support protocol in the local components of a mobile network, such as mobile routers and mobile nodes. It does not require any change in correspondent nodes, home agents, or any other network components located in domains beyond the mobile network. It enables a correspondent node to forward packets directly to the mobile network without any tunneling, thus solves the problem of pinball routing and reduces packet delays and encapsulation overheads in the core networks. To evaluate the MoRaRo scheme, we developed analytical models for estimating packet delivery overheads, efficiency, and delay. We also performed an analysis and simulation to assess performance during a handover in terms of the correspondent binding update latency and overheads. We found that the scheme improved the packet delivery efficiency of the network mobility support protocol by about 100%. It significantly reduced the packet delivery delays while using only a few signaling messages for establishing a direct route between a correspondent node and a mobile node moving in the mobile network. The reduction in packet delivery overheads increases the overall system throughput by effectively utilizing the network resources for transporting data packets. Similarly, reducing delays in packet delivery helps to efficiently support time critical multimedia applications in mobile networks. This scheme also eliminates the unnecessary home agents traversal by data packets, thus improves system reliability by reducing the excessive computational burden on the home agents.

The CoMoRoHo scheme achieves a seamless handover of long, vehicular-multihomed mobile networks. This scheme exploits the advantage of a multihoming environment of a long vehicular mobile network, which can have two or more mobile routers spatially separated by a certain distance. The CoMoRoHo scheme enables different mobile routers to access different subnets during a handover and cooperatively receive packets destined for each other. That is, when a mobile router is performing a handover to a new subnet, it requests the access router of the old subnet to tunnel packets to some other mobile router that is still located in the domain of the old subnet. The other mobile router then forwards the packets to the mobile router that has requested the access router for tunneling. Similarly, when the mobile router completes its handover to the new subnet, it helps the other mobile router to perform a lossless handover to the new subnet. This scheme thus reduces packet losses by establishing a local tunnel between the access router of the old network and the mobile router by using only one signaling message. To evaluate the CoMoRoHo scheme, we carried out performance modeling in regard to the handover latency, packet loss, signaling overhead, and packet delivery overhead in access networks.

The analysis results showed that the CoMoRoHo scheme outperformed other schemes such as Fast Handover for Mobile IPv6 by reducing the packet losses as well as signaling overheads by more than 50%. Consequently, the reduction in packet losses improves the quality of service provided by mobile networks to user applications. Moreover, the reduction in signaling and packet-delivery overheads in the access network enhances the scalability of network mobility management by keeping the performance intact even when the network is overloaded. Our contribution regarding network mobility is thus the design and evaluation of route optimization and handover management schemes to improve the overall system performance of mobile networks by reducing overheads, delays, and losses.

In recent years, a wide variety of wireless access networks that support multimedia services have been emerging. The service areas of many of these networks overlap so that a mobile user can have an access to any network that supports the user's application. A mobile user can take advantage of the availability of such heterogeneous wireless networks only when the mobile device is equipped with a mechanism that can select the most appropriate service or network type depending on the user's current location and preferences. In this dissertation, we addressed the issue of optimal network selection by defining a user satisfaction function, which is a user centric measurement of the network service quality, and evaluating the impact that the mobility from one network to another has on the user satisfaction function. In addition, we designed and evaluated a graceful vertical handover scheme to smoothly transfer connections from one network to another.

We quantified user satisfaction as a function of the bandwidth utility and handover latency. The bandwidth utility of a network service, in turn, is a measure of the degree of network capability to fulfill the bandwidth requirement of user applications. We took into consideration three types of applications: rigid, adaptive, and elastic, and formulated their bandwidth utility functions. We analyzed how user satisfaction changes when the user moves from a high bandwidth network to a low bandwidth network, under two types of movements: uniform random and directional movements. In the case of a random movement, we have no idea about the direction in which the user will move in the near future. To model such a random system, we used the well-known and widely used random-walk mobility model. Our evaluation showed that in the random movement, the change in user satisfaction due to a handover to a lower bandwidth network was not significant. However, when we considered a directional movement, where the movement probability to a low bandwidth network was much higher, i.e., the user used a high bandwidth network for short time and handed over to a low bandwidth network to complete the call, the satisfaction degradation was significant. Therefore, to avoid such degradation, we proposed an access network selection algorithm that uses a prediction of user mobility pattern. Limitation of this study is that we considered only a few very important parameters such as

bandwidth and handover latency for defining the user satisfaction function. There might be a lot of other parameters, such as network congestion level, latency, losses, reliability, and cost.

If an optimal network selected by the above algorithm becomes suboptimal because of increasing network congestion or user mobility, a mobile user has to perform a vertical handover to a new network, in most cases, to an inferior one. Since there are disparities in the overlay networks' such characteristics as bandwidths, latencies, and loss rates, the handover may result in the abrupt change in throughput, a large number of packet losses, and disruption of the communication services. This may ultimately impair user satisfaction obtained from the network service. To address this issue, we presented a graceful vertical handover scheme. This scheme provides a graceful degradation of performance by smoothing the throughput change and reducing the possible packet losses during a handover. To achieve such a gracefulness in performance change, this scheme implements its functionalities in the link layer, IP layer, and TCP or application layer. We carried out simulations to study the significance of the scheme's performance on the TCP-layer throughput and losses. We found that the scheme greatly improved the TCP throughput by smoothing the packet sending rate and reducing losses during a graceful handover to a low bandwidth network. In this way, our research on the optimal network selection and graceful handover mechanisms helps to improve the overall system performance of heterogeneous wireless networks.

7.2   Future Works

Both the network mobility and heterogeneous wireless networks are newly emerging elements of the future multimedia communication systems. Besides the issues addressed in this dissertation, there are many other issues to be resolved for the full-fledge deployment of network mobility and heterogeneous networks.

Network mobility decomposes the wireless connection between mobile nodes and access routers into two segments: one between the mobile nodes and mobile routers and the other between the mobile routers and access routers. This enables optimizing these segments independently. The mobile nodes can enjoy better wireless connectivity because of a shorter distance between them and mobile routers. Similarly, the mobile routers can have a high bandwidth dedicated wireless connection with the access routers. The mobile nodes residing in a mobile network can communicate to one another. However, the current NEMO protocols do not provide an efficient way of such communications. This is an issue of local route optimization, which may enable mobile nodes to forward packets via an optimal path when communicating with other mobile nodes of the same mobile network. Further extension of local route optimization may result in supporting a standalone mobile network, which will enable mobile nodes to

establish direct connections between them, even when the mobile network is not attached to the Internet infrastructure. This will, in fact, help in providing location based services to mobile nodes from the local servers located in the mobile network. Hence, local route optimization and supports for a standalone mobile network worth further study and research.

Another issue relating network mobility management is multihoming of mobile networks. Multihoming indicates that a mobile network can establish multiple logical or physical connections with access routers of the Internet infrastructure. A multihomed mobile network can have many home agents, many mobile router, or a mobile router with many egress interfaces. Multihoming can provide solutions to optimize network mobility management in many ways. In our CoMoRoHo scheme, we took into consideration a multihomed vehicular mobile network to achieve a seamless handover. Similarly, since the wireless connection between a mobile router and an access router may be a bottleneck in providing wideband multimedia services to many mobile nodes residing in the mobile network, multihoming can solve this problem by having multiple connections. Mulithoming can also improve the fault-tolerance capacity of the mobile network. Therefore, multihoming is a good research topic for realizing reliable, efficient, widely usable mobile networks.

In the optimal network selection section, our user satisfaction function included only the bandwidth and handover of latency. However, besides these two parameters, user satisfaction may depend on many other parameters such as loss rate, delays, cost, power consumption, and reliability. Therefore, while defining an effective user satisfaction function, we need to determine the adequate number of parameters and nature of their impact on the user satisfaction function. As a first step toward these issues, in [14], we included the available bandwidth, delay, and loss rate to define a user-centric performance. Similarly, we included the price of the network service as well as the power consumption of a mobile device to estimate a user-centric cost. The user satisfaction function can be a function of user-centric performances and costs. Hence, designing a universal function for quantifying user satisfaction worths further researches. Moreover, since user satisfaction may vary from person to person, time to time, or service to service, imparting an adaptiveness in the user satisfaction function may also be an important research topic.

The chapter on the graceful handover presented just an initiation to address the problem of abrupt performance changes and packet losses during a vertical handover. There are yet many issues to be solved, such as signaling complexity in choosing a relay node, delay in re-establishing a connection via a relay node, relay node coordination and configuration, security in relaying mechanisms, and prediction of mobility behavior to set the graceful degradation time.

References

[1] E. A. Brewer, R. H. Katz, Y. Chawathe, S. D. Gribble, T. Hodes, G. Nguyen, M. Stemm, T. Henderson, E. Amir, H. Balakrishanan, A. Fox, V. N. Padmanabhan, and S. Seshan, "A network architecture for heterogeneous mobile computing," IEEE Personal Communications, vol. 5, no. 5, pp. 8-24, October 1998.

[2] G. Wu, M. Mizuno, and P. J. Havinga, "MIRAI architecture for heterogeneous networks," IEEE Communications Magazine, vol. 40, no. 2, pp. 126-134, February 2002.

[3] C. Perkins, "IP Mobility support for IPv4," IETF RFC 3344, August 2002.

[4] D. Johnson, C. Perkins, and J. Arkko, "Mobility support in IPv6," IETF RFC 3775, June 2004.

[5] H. Soliman, C. Catelluccia, K. E. Malki, and L. Bellier, "Hierarchical mobile IPv6 mobility management (HMIPv6)," IETF RFC 4140, August 2005.

[6] R. Koodli, "Fast handover for mobile IPv6," IETF RFC 4068, July 2005.

[7] A. Conta and S. Deering, "Generic packet tunneling in IPv6 specification," IETF RFC 2473, December 1998.

[8] V. P. Kafle, E. Kamioka, and S. Yamada, "MoRaRo: Mobile router-assisted route optimization for network mobility (NEMO) support," IEICE Transactions on Information and Systems, vol. E89-D, no. 1, pp. 158-170, January 2006.

[9] V. P. Kafle, E. Kamioka, and S. Yamada, "CoMoRoHo: Cooperative mobile router-based handover scheme for long-vehicular multihomed networks," IEICE Transactions on Communications, vol. E89-B, no. 10, October 2006.

[10] V. P. Kafle, S. Pack, Y. Choi, E. Kamioka, and S. Yamada, "IIPP: Integrated IP paging protocol with a power save mechanism," Wireless Communications and Mobile Computing (WCMC), 2007.

[11] V. P. Kafle, E. Kamioka, and S. Yamada, "Maximizing user satisfaction based on mobility on heterogeneous mobile multimedia communication networks," IEICE Transactions on Communications, vol. E88-B, no. 7, pp. 2709-2717, July 2005.

[12] V. P. Kafle, E. Kamioka, and S. Yamada, "A scheme for graceful handover in heterogeneous overlay networks," in Proc. IEEE Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2006), pp. 343-348, June 2006.

[13] V. P. Kafle, E. Kamioka, and S. Yamada, "Extended correspondent registration scheme for reducing handover delay in Mobile IPv6," in Proc. International Workshop on Future Mobile and Ubiquitous Information Technology (FMUIT 2006 - held in conjunction with MDM 2006), p. 110, May 2006.

[14] V. P. Kafle, E. Kamioka, and S. Yamada, "User-centric performance and cost analysis for selecting access networks in heterogeneous overlay systems," in Proc. IFIP/IEEE Conference on Management of Multimedia Networks and Services (MMNS 2005), LNCS 3754, pp. 277-288, October 2005.

[15] V. P. Kafle, E. Kamioka, and S. Yamada, "Mobility and its impact on user satisfaction function in heterogeneous wireless networks," in Proc. Conference on Communications and Computer Networks (CCN2004), pp.397-402, November 2004.

[16] 3GPP TR 22.978 V7.1.0 (2005-06), "All-IP network (AIPN) feasibility study (Release 7)," Technical Specification Group Service and Systems Aspects, 3rd Generation Partnership Projects, 2005.

[17] G. Lee, P. Faratin, S. Bauer, and J. Wroclawski, "A user-guided cognitive agent for network service selection in pervasive computing environments," in Proc. IEEE PerCom 2004, pp. 219-228.

[18] J. Altmann, H. Daanen, H. Oliver, and A. S.-B. Suarez, "How to market-manage a QoS network," in Proc. IEEE INFOCOM 2002, vol. 1, pp. 284-293.

[19] E. Perera, V. Sivaraman, and A. Seneviratne, "Survey of network mobility support," ACM Mobile Computing and Communication Review, vol. 8, no. 2, pp. 7-19, April 2004.

[20] IETF Network Mobility (NEMO) Charter. http://www.ietf.org/html.charters/nemo-charter.html

[21] H.-Y. Lach and T. Ernst, "Network mobility support terminology," IETF Internet Draft, draft-ietf-nemo-terminology-02.txt, February 2004.

[22] V. Devarapalli, R. Wakikawa, A. Pertruscu, and P. Thubert, "Network mobility (NEMO) basic support protocol," IETF RFC 3963, January 2005.

[23] V. P. Kafle, E. Kamioka, and S. Yamada, "Performance evaluation of network mobility (NEMO) basic support protocol and route optimization," IEICE Technical Report, vol. 104, no. 41, pp. 53-58, MoMuC2004-74, November 2004.

[24] C. Ng, P. Thubert, H. Ohnishi, and E. Paik, "Taxonomy of route optimization models in the NEMO context," IETF Internet Draft, draft-thubert-nemo-ro-taxonomy-04.txt, February 2005.

[25] P. Thubert and M. Molteni, "IPv6 reverse routing header and its application to mobile networks," IETF Internet Draft, draft-thubert-nemo-reverse-routing-header-05.txt, June 2004.

[26] J. Na, S. Cho, C. Kim, S. Lee, H. Kang, and C. Koo, "Route optimization scheme based on path control header," IETF Internet Draft, draft-na-nemo-path-control-header-00.txt, April 2004.

[27] H. Ohnishi, K. Sakitani, and Y. Takagi, "HMIP based route optimization method in a mobile network," IETF Internet Draft, draft-ohnishi-nemo-ro-hmip-00.txt, October 2003.

[28] E. Perera, R. Hsieh, and A. Senevirante, "Extended network mobility support," IETF Internet Draft, draft-perera-nemo-extended-00.txt, July 2003.

[29] T. Ernst, C. Castelluccia, and H.-Y. Lach, "Extending mobile-IPv6 with multicast to support mobile networks in IPv6," in Proc. IEEE ECUMN 2000, pp. 114-121, October 2000.

[30] R. Wakikawa and M. Watari, "Optimized route cache protocol (ORC)," IETF Internet Draft, draft-wakikawa-NEMO-orc-00.txt, July 2004.

[31] S. Kent and R. Atkinson, "Security architecture for the Internet protocol," IETF RFC 2401, November 1998.

[32] J. Cuellar, J. Morris, D. Mulligan, J. Peterson, and J. Polk, "Geopriv requirements," IETF RFC 3693, February 2004.

[33] S. Yamada and E. Kamioka, "Access control for security and privacy in ubiquitous computing environments," IEICE Transactions on Communications, vol. E88-B, no. 3, pp. 846-956, March 2005.

[34] S.-C. Lo, G. Lee, W.-T. Chen, and J.-C. Liu, "Architecture for mobility and QoS support in all-IP wireless networks," IEEE Journal on Selected Areas in Communications, vol. 22, no. 4, pp. 691-705, May 2004.

[35] E. Kamioka, S. Yamada, and T. Sanda, "Proposal for context-aware information delivery and personal communication network architectures with preliminary evaluations of their performance," IEICE Transactions on Communications, vol. E87-B, no. 9, pp. 2672-2681, September 2004.

[36] D. Gross and C. M. Harris, Fundamentals of Queuing Theory, Wiley Series in Probability and Statistics, John Wiley & Sons, 1998.

[37] A. Campbell, J. Gomez, S. Kim, and C. Wan, "Comparison of IP micromobility protocols," IEEE Wireless Communications, vol. 9, no. 1, pp. 72-82, 2002.

[38] R. Ramjee, K. Varadhan, L. Salgarelli, S. Thuel, S. Wang, and T. Porta, "HAWAII: A domain-based approach for supporting mobility in wide-area wireless network," IEEE/ACM Transactions on Networking, vol. 10, no. 2, pp. 396-410, 2002.

[39] V. P. Kafle, E. Kamioka, and S. Yamada, "A handover scheme for long-vehicular multi-homed mobile networks," IEICE Technical Report, vol. 105, no. 490, MoMuC2005-71, pp.75-80, November 2005.

[40] C. Ng, E. Paik, and T. Ernst, "Analysis of multihoming in network mobility support," IETF Internet draft, draft-ietf-nemo-multihoming-issues-04.txt, October 2005.

[41] L. Zhang, S. Shenker, and D. D. Clark, "Observation of the dynamics of a congestion control algorithm: the effect of two-way traffic," in Proc. ACM SIGCOMM 1991, pp. 133-147.

[42] Y. Xu, H. C. J. Lee, and V. L. L. Thing, "A local mobility agent selection algorithm for mobile networks," in Proc. IEEE ICC 2003, pp. 1074-1079.

[43] S. Fu, M. Atiquzzaman, L. Ma, and Y.-J. Lee, "Signaling cost and performance of SIGMA: a seamless handover scheme for data networks," Wireless Communications and Mobile Computing (WCMC), vol. 5, pp. 825-845, October 2005.

[44] T. T. Kwon, M. Gerla, S. Das, and S. Das, "Mobility management for VoIP service: Mobile IP vs. SIP," IEEE Wireless Communications, vol. 9, no. 5, pp. 66-75, October 2002.

[45] N. Banerjee, W. Wu, S. Das, S. Dawkins, and J. Pathak, "Mobility support in wireless Internet," IEEE Wireless Communications, vol. 10, no. 5, pp. 54-61, October 2003.

[46] S. Shin, A. Forte, A. S. Rawat, and H. Schulzrine, "Reducing MAC layer handoff latency in IEEE 802.11 wireless LANs," in Proc. ACM MobiWac 2004, pp. 19-26.

[47] S. Pack and Y. Choi, "A study on performance of hierarchical mobile IPv6 in IP-based cellular networks," IEICE Transactions on Communications, vol. E87-B, no. 3, pp. 462-469, March 2004.

[48] J. Cao, L. Zhang, H. Chan, and S. K. Das, "Design and performance evaluation of an improved Mobile IP protocol," in Proc. IEEE INFOCOM 2004, pp. 319-329.

[49] T. Schmidt and M. Wahlisch, "Predictive versus reactive - analysis of handover performance and its implications on IPv6 and multicast mobility," Telecommunication Systems, vol. 30, no. 1-3, pp. 123-142, Springer, November 2005.

[50] T. Narten, E. Nordmak, and W. Simpson, "Neighbor discovery for IP version 6 (IPv6)," IETF RFC 2461, December 1998.

[51] J. Hoffmeyer, I.-P. Park, M. Majmundar, and S. Blust, "Radio software download for commercial wireless reconfigurable devices," IEEE Communications Magazine, vol. 42, no. 3, pp. 26-32, March 2003.

[52] J. W. Floroiu, R. Ruppelt, D. Saselem, and J. V. Stephanopoli, "Seamless handover in terrestrial radio access networks," IEEE Communications Magazine, vol. 41, no. 11, pp. 110-114, November 2003.

[53] L. Breslau and S. Shenker, "Best effort versus reservations: a simple comparative analysis," ACM Computer Communication Review, vol. 28, pp. 131-143, September 1998.

[54] S. Shenker, "Fundamental design issues for the future Internet," IEEE Journal on Selected Areas in Communications, vol. 13, no. 7, pp. 1176-1188, September 1995.

[55] Y. Hidaka and Y. Teshigawara, "Multimedia traffic management methods in consideration of end-user satisfaction," in Proc. Asia-Pacific Network Operations and Management Systems(APNOMS'97), pp. 275-286, 1997.

[56] J. S. Ho and I. F. Akyildiz, "Mobile user location update and paging under delay constraints," ACM/Baltzer Wireless Networks Journal, vol. 1, pp. 413-425, December 1995.

[57] J. Lie, Y. Pan, and X. Jia, "Analysis of dynamic location management for PCS networks," IEEE Transactions on Vehicular Technology, vol. 51, no. 5, pp. 1109-1119, September 2002.

[58] V. Pandey, D. Ghosal, and B. Mukherjee, "Exploiting user profiles to support differentiated services in next generation wireless networks," IEEE Network, vol. 18, no. 5, pp. 40-48, September 2004

[59] E. Cayirci and I. F. Akyildiz, "User mobility pattern scheme for location update and paging in wireless systems," IEEE Transactions on Mobile Computing, vol. 1, no. 5, pp. 236-247, July-September 2002.

[60] S. Tabbane, "Location management methods for third-generation mobile systems," IEEE Communications Magazine, vol. 35, no. 8, pp. 72-78, August 1997.

[61] D. Wischik and A. Greenberg, "Admission control from booking ahead shared resources," in Proc. IEEE INFOCOM 1998, vol. 1, pp. 873-882.

[62] A. Malla, M. El-Kadi, S. Olariu, and P. Todorova, "A fair resource allocation protocol for multimedia wireless networks," IEEE Transactions on Parallel and Distributed Systems, vol. 14, no. 1, pp. 63-71, January 2003.

[63] C. Oliveira, J. B. Kim, and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," IEEE Journal on Selected Areas in Communications, vol. 16, no. 6, pp. 858-874, August 1998.

[64] M. Stemm and R. H. Kartz, "Vertical handoffs in wireless overlay networks," ACM Mobile Network and Application (MONET), vol. 3, issue 4, pp. 335-350, 1998.

[65] C. Guo, Z. Guo, Q. Zhang, and W. Zhu, "A seamless and proactive end-to-end mobility solution for roaming across heterogeneous wireless networks," IEEE Journal on Selected Areas in Communications, vol. 22, no. 5, pp. 834-848, June 2004.

[66] J. McNair and F. Zhu, "Vertical handoffs in fourth generation multinetwork environments," IEEE Wireless Communications, vol. 11, no. 3, pp. 8-15, June 2004.

[67] S. Lee, S. Banerjee, and B. Bhattacharjee, "The case of a multi-hop wireless local area network," in Proc. IEEE INFOCOM 2004, vol. 2, pp. 894-905.

[68] J. Chen, S.-H. Chan, J. He, and S.-C. Liew, "Mixed mode WLAN: the integration of ad-hoc with wireless LAN infrastructure," in Proc. IEEE GLOBECOM 2003, pp. 2623-3627.

[69] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta, "Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments," in Proc. IEEE INFOCOM 2000, pp. 1537-1545.

[70] S. Mascolo, C. Casetti, M. Gerla, M. Y. Sanadidi, and R. Wang, "TCP Westwood: Bandwidth estimation for enhanced transport over wireless links," in Proc. ACM SIGMOBILE 2002, pp. 287-297.

[71] Network Simulator ns-2, http://www.isi.edu/nsnam/ns/

[72] http://www.inrialpes.fr/planete/mobiwan/

List of Publications

Transactions and Journals:

1. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "CoMoRoHo: Cooperative mobile router-based handover scheme for long-vehicular multihomed networks," IEICE Transactions on Communications, vol. E89-B, no. 10, October 2006.

2. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "MoRaRo: Mobile router-assisted route optimization for network mobility (NEMO) support," IEICE Transactions on Information and Systems, vol. E89-D, no. 1, pp. 158-170, January 2006.

3. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Maximizing user satisfaction based on mobility on heterogeneous mobile multimedia communication networks," IEICE Transactions on Communications, vol. E88-B, no. 7, pp. 2709-2717, July 2005.

4. Ved P. Kafle, Sangheon Pack, Yanghee Choi, Eiji Kamioka, and Shigeki Yamada, "IIPP: Integrated IP paging protocol with a power save mechanism," to appear in Wiley Wireless Communications and Mobile Computing (WCMC), (accepted, October 2005).

Referred Conferences:

1. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "A scheme for graceful handover in heterogeneous overlay networks," in Proceedings of IEEE International Conference on Wireless and Mobile Computer, Networking and Communications (WiMob 2006), pp. 343-348, Montreal, Canada, June 2006.

2. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Extended correspondent registration scheme for reducing handover delay in Mobile IPv6," in Proceedings of International Workshop on Future Mobile and Ubiquitous Information Technology (FMUIT 2006), held in conjunction with MDM 2006, p. 110, Nara, Japan, May 2006.

3. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "User-centric performance and cost analysis for selecting access networks in heterogeneous overlay systems," in Proceedings of 8th IFIP/IEEE International Conference on Management of Multimedia Networks and Services (MMNS 2005), LNCS 3754, pp. 277-288, Barcelona, Spain, October 2005.

4. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Mobile IPv6-type route optimization scheme for network mobility (NEMO) support," in Proceedings of 10th IFIP International

Conference on Personal Wireless Communications (PWC 2005), pp. 381-388, Colmar, France, August 2005.

5. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Mobility and its impact on user satisfaction in heterogeneous wireless networks," in Proceedings of 2nd IASTED International Conference on Communication and Computer Networks (CCN 2004), pp. 397-402, MIT, Cambridge, USA, November 2004.

6. Ved Kafle, Sangheon Pack, and Yanghee Choi, "Performance evaluation of IP paging with power save mechanisms," in Proceedings of 28th IEEE Conference on Local Computer Networks (LCN 2003), pp. 315-316, Bonn, Germany, October 2003.

7. Sangheon Pack, Ved Kafle, and Yanghee Choi, "An adaptive power Ssaving mechanism in IEEE 802.11 networks to support IP paging protocol," in Proceedings of 8th IEEE International Workshop on Mobile Multimedia Communications (MoMuC 2003), pp. 461-466, Munich, Germany, October 2003.

8. Sangheon Pack, Ved Kafle, and Yanghee Choi, "Performance analysis of IP paging protocol in IEEE 802.11 networks," in Proceedings of 28th IEEE Conference on Local Computer Networks (LCN 2003), pp. 673-680, Bonn, Germany, October 2003.

Presentations and Technical Reports:

1. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Schemes for Network Mobility Management," in Proceedings of IEICE Society Conference, Network Planning, Control, and Management, BS-15-20, Kanazawa, Japan, September 2006.

2. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Local route optimization for network mobility (NEMO) support," in Proceedings of IEICE General Conference, Technologies and Architectures for Ubiquitous Network Systems, BS-8-7, Tokyo, Japan, March 2006.

3. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "A handover scheme for long-vehicular multihomed mobile networks," IEICE Technical Report, vol. 105, no. 490, MoMuC2005-71, pp. 75-80, November 2005.

4. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "A framework of graceful vertical handover," IEICE Society Conference, Network Planning, Control and Management BS-10-17, Hokkaido, Japan, September 2005.

5. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Evaluation of user-centric performances to select an optimal access network in heterogeneous systems," IEICE Technical Report, vol. 105 no. 80, MoMuC2005-10, pp. 51-56, May 2005.

6. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Performance evaluation of network mobility (NEMO) basic support protocol and route optimization," IEICE Technical Report, vol. 104, no. 443, MoMuC2004-74, pp. 53-58, November 2004.

7. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "A study of overheads in network mobility (NEMO) basic support protocol," in Proceedings of the IEICE Society Conference, Network Planning, Control and Management BS-9-9, Tokushima, Japan, September 2004.

8. Ved P. Kafle, Eiji Kamioka, and Shigeki Yamada, "Effect of mobility on user satisfaction in heterogeneous wireless networks," IEICE Technical Report, vol. 104, no. 190, MoMuC2004-41, pp. 7-12, July 2004.