2006

MOBILITY SUPPORT ARCHITECTURES FOR NEXT-GENERATION WIRELESS NETWORKS

WANG, Ql

MOBILITY SUPPORT ARCHITECTURES FOR NEXT-GENERATION WIRELESS NETWORKS

by

QI WANG

B.Eng., Dalian Maritime University, 1995

M.Eng., Dalian Maritime University, 1998

A thesis submitted to the University of Plymouth

in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing, Communications and Electronics

Faculty of Technology

March 2006

MOBILITY SUPPORT ARCHITECTURES FOR NEXT-GENERATION WIRELESS NETWORKS

Qi Wang

Abstract

With the convergence of the wireless networks and the Internet and the booming demand for multimedia applications, the next-generation (beyond the third generation, or B3G) wireless systems are expected to be all IP-based and provide real-time and non-real-time mobile services anywhere and anytime. Powerful and efficient mobility support is thus the key enabler to fulfil such an attractive vision by supporting various mobility scenarios. This thesis contributes to this interesting while challenging topic.

After a literature review on mobility support architectures and protocols, the thesis starts presenting our contributions with a generic multi-layer mobility support framework, which provides a general approach to meet the challenges of handling comprehensive mobility issues. The cross-layer design methodology is introduced to coordinate the protocol layers for optimised system design. Particularly, a flexible and efficient cross-layer signalling scheme is proposed for interlayer interactions. The proposed generic framework is then narrowed down with several fundamental building blocks identified to be focused on as follows.

As widely adopted, we assume that the IP-based access networks are organised into administrative domains, which are inter-connected through a global IP-based wired core network. For a mobile user who roams from one domain to another, macro (inter-domain) mobility management should be in place for global location tracking and effective handoff support for both real-time and non-real-time applications. Mobile IP (MIP) and the Session

Initiation Protocol (SIP) are being adopted as the two dominant standard-based macro-mobility architectures, each of which has mobility entities and messages in its own right. The work explores the joint optimisations and interactions of MIP and SIP when utilising the complementary power of both protocols. Two distinctive integrated MIP-SIP architectures are designed and evaluated, compared with their hybrid alternatives and other approaches. The overall analytical and simulation results shown significant performance improvements in terms of cost-efficiency, among other metrics.

Subsequently, for the micro (intra-domain) mobility scenario where a mobile user moves across IP subnets within a domain, a micro mobility management architecture is needed to support fast handoffs and constrain signalling messaging loads incurred by intra-domain movements within the domain. The Hierarchical MIPv6 (HMIPv6) and the Fast Handovers for MIPv6 (FMIPv6) protocols are selected to fulfil the design requirements. The work proposes enhancements to these protocols and combines them in an optimised way, resulting in notably improved performances in contrast to a number of alternative approaches.

Keywords: Mobility Support (Management), Macro Mobility, Micro Mobility, Mobile IP, Session Initiation Protocol, Quality of Service, Next-Generation Wireless Networks

To

my parents, Yanbao Wang and Meiling Zhu,

my wife, Jianfeng Yuan,

and my brother, Long Wang.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1G	First Generation
2G	Second Generation
2.5G	Second Generation +
3G	Third Generation
3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project 2
AAA	Authentication, Authorisation, Accounting
AAAL	Local AAA Server
AAAH	Home AAA Server
ACK	Acknowledgement
AP	Access Point
AOR	Address-of-Record
AR	Access Router
ARn or ARp	New AR or Previous AR
ATM	Asynchronous Transfer Mode
B3G	Beyond the Third Generation
BA	Binding Acknowledgement
BR	Binding Request
BRR	Binding Refresh Request
BS	Base Station
BU	Binding Update
BW	Binding Warning

CDMA	Code Division Multiple Access
CH	Corresponding Host
CIP	Cellular IP
CLASS	Cross-LAyer Signalling Shortcut
CoA	Care-of Address
CoT	Care-of address Test
CoTI	Care-of address Test Initiation
CS	Circuit-Switched
DAD	Duplicate Address Detection
DHCP	Dynamic Host Configuration Protocol
DiffServ	Differentiated Services
ECRIT	Emergency Context Resolution with Internet Technologies
EPAD	Emergency Provider Access Directory
FA	Foreign Agent
FBA	Fast Binding Acknowledgement
FBU	Fast Binding Update
FNA	Fast Neighbour Advertisement
FAn or FAp	New FA or Previous FA
FS	Foreign Server
FSn or FSp	New FS or Previous FS
FDMA	Frequency Division Multiple Access
FMS	Foreign Mobility Server
FTP	File Transfer Protocol
FMIPv6	Fast Handovers for Mobile IPv6
GFA	Gateway FA

GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GRX	GPRS Roaming eXchange
GSM	Global System for Mobile Communication
GW	Gateway
HA	Home Agent
HFA	Hierarchical FA
HLR	Home Location Register
HMIPv6	Hierarchical Mobile IPv6
IEE	Institution of Electrical Engineers
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
HACK	Handover Acknowledgement
HI	(FMIPv6) Handover Initiation or
	(HIP) Host Identify
HIP	Host Identity Protocol
HMS	Home Mobility Server
HO	Handoff (Handover)
HoA	Home Address
HoT	Home Address Test
HoTI	Home Address Test Initiation
HR	Home Registration or
	Home Registrar
HRply	Handoff Reply

HRqst	Handoff Request
HS	Home Server
HTTP	Hypertext Transfer Protocol
HY-MIP-SIP	Hybrid MIP-SIP
IEPREP	Internet Emergency Preparedness
IntServ	Integrated Services
IPsec	Security architecture for IP
IR	Intermediate Router
ICMP	Internet Control Message Protocol
ID	Identifier
IP	Internet Protocol
IPv4	IP version 4
IPv6	IP version 6
ITU	International Telecommunications Union
L1	Layer-1 (Physical Layer)
L2	Layer-2 (Link Layer)
L3	Layer-3 (Network Layer)
L4	Layer-4 (Transport Layer)
L5	Layer-5 (Application Layer)
LA	Location Area
LAN	Local Area Network
LCoA	on-Link or Local Care-of Address
LI-MIP-SIP	Loosely Integrated MIP-SIP
LL-MIPv4	Low Latency MIPv4 Handoffs
LU	Location Update

MAC	Medium Access Control
MAHO	Mobile-Assisted Handoff
MAP	(HMIPv6) Mobility Anchor Point or (GSM) Mobile Application Part
MCHO	Mobile-Controlled Handoff
MIP	Mobile IP
MIPv4	Mobile IPv4
MIPv4-RO	Mobile IPv4 with Route Optimisation
MIPv4-RR	Mobile IPv4 Regional Registration
MIPv6	Mobile IPv6
MH	Mobile Host
MM	Mobility Management
MSC	Mobile Switching Centre
NAI	Network Access Identifier
NCHO	Network-Controlled Handoff
NS	Neighbour Solicitation
NSIS	Next Steps In Signalling
OSI	Open System Interconnection
PAN	Personal Area Network
PAVER	Prompt Address Verification and complEment Replacement
PDP	Packet Data Protocol
PFAN	Previous Foreign Agent Notification
PrRtAdv	Proxy Router Advertisement
PrRtSol	Proxy Router Solicitation
PS	Packet-Switched

PSAP	Public Safety Answering Point
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RA	Router Advertisement or
	(GPRS) Routing Area
RCoA	Regional Care-of Address
RE	Route Extension
REED	RE End Declaration
RegReply	Registration Reply
RegReq	Registration Request
RO	Route Optimisation
RR	Return Routability
RRC	Radio Resource Control
RRM	Radio Resource Management
RS	Router Solicitation
RSS	Received Signal Strength
RSVP	Resource reSerVation Protocol
RTP	Real-time Transport Protocol
SA	Security Association
SCTP	Stream Control Transmission Protocol
SDP	Session Description Protocol
SGSN	Serving GPRS Support Node
SIP	Session Initiation Protocol
SIPPING	Session Initiation Proposal Investigation
SNMP	Simple Network Management Protocol

SS	Session Setup
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TI-MIP-SIP	Tightly Integrated MIP-SIP
UA	User Agent
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunication System
URA	UTRAN Registration Area
URI	Uniform Resource Identifier
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visitor Location Register
VoIP	Voice over IP
WAN	Wide Area Network
WLAN	Wireless LAN
WWW	World Wide Web

Acknowledgments

I own many people a sincere thank you for their unforgettable aids to have helped me through this long, difficult yet rewarding process of my Ph.D. research.

Firstly, I would like to thank Dr. Mosa Ali Abu-Rgheff, my director of studies, for his consistent encouragement, enormous patience as well as professional supervision throughout the whole course. Particularly, I appreciate his guidance in publishing our work in prestigious international journals and conferences in a timely and regular way, and his efforts to arrange my attendance of all the involved conferences so that I could present the papers and benefit from peer reviews and feedback. I would also like to thank my other supervisors, Prof. Martin Tomlinson and Dr. Kit Reeve, for their valuable support.

Secondly, I am grateful to Dr. Steven Furnell for his detailed review of my preliminary work and constructive advice on the project. I have also enjoyed the company of the current and previous research group members, and value the useful discussions with them and other colleagues associated with other groups or universities. This list can hardly be exhaustive: Prof. Guangchun Zhou, Dr. Lingfen Sun, Dr. Keming Yu, Dr. Genhua Pan, Dr. Bingmei Yang, Steve Donohoe, Anne Donohoe, German Gonzalez Lopez, Dr. Ammad Akram, Dr. Yun Won Chung, Joerg Wolf, Dr. Hongxing Zhao, Dr. Yongming Dai, Dr. Lixin Cheng, Dr Renxiang Ding, Chengfei Sui, Jinbo Zhao, Fan Wu, Zizhi Qiao, Zhuoqun Li, Hu Pin, Jing Cai, Xin Xu, Qijun Zhan, Yuqing Du, Yun Zhou, Dr Paul Dowland, Dr. Brahim Hamadicharef, Dr. Yeonho Chung, Alan Simpson, Eduardo Coutinho, Vadim Tikhanoff, and Dr. Ali Muayyadi...

Furthermore, I would like to thank Prof. Phil Dyke, Mrs. Sue Kendall, Ms. Carole Watson, Ms. Sue Locke, Ms. Pat Blower, Mrs. Glynis Hockey, Nicola Westlake, Sarah Amador and other managerial staff for their great help and considerations.

Finally, my family deserve a particular recognition for their tremendous love and unconditional support psychologically and financially during the past years. Without them, this project would have been an impossible mission. I am deeply indebted to my parents Yanbao Wang and Meiling Zhu, my brother Long Wang and his family (my sister-in-law Maomao Liu and my lovely niece Runpeng Wang), my parents-in-laws Shengtang Yuan and Yuqin Zhao, my brother-in-law Jizhou Yuan and his wife Zhi Yao, and all the other supporting relatives. Surely, my most special love goes to my dear wife Jianfeng Yuan, who always loves me and supports me without any reservation, and shares my laughers and tears with greatest understanding in my everyday life throughout all these years.

I love you all!

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was funded in part with the aid of an ORS grant and a University bursary.

Publications:

Journal Papers

1. [Wang and Abu-Rgheff IJCS] Q. Wang and M. A. Abu-Rgheff, "Signalling Analysis of Cost-Efficient Mobility Support by Integrating Mobile IP and SIP in All IP Wireless Networks", *(Wiley) International Journal of Communication Systems*, vol. 19, No. 2, pp. 225-247, March 2006.
2. [Wang and Abu-Rgheff CE] Q. Wang and M. A. Abu-Rgheff, "Next-Generation Mobility Support", *IEE Communications Engineer*, vol. 1, no. 1, pp. 16-19, January/February 2003.

Conference Papers

3. [Wang and Abu-Rgheff 3G2005] Q. Wang and M. A. Abu-Rgheff, "IPv6-Based Architecture for Fast and Cost-Effective Micro-Mobility Management", in *Proc. of IEE 6th International Conference on 3G and Beyond (IEE 3G2005)*, London, UK, November 2005, pp. 101-105.
4. [Wang and Abu-Rgheff 3G2004] Q. Wang and M. A. Abu-Rgheff, "Interacting Mobile IP And SIP For Efficient Mobility Support In All IP Wireless Networks", in *Proc. of IEE 5th International Conference on 3G Mobile Communication Technologies (IEE 3G2004)*, London, UK, October 2004, pp. 664-668.
5. [Wang, Abu-Rgheff etc ICC04] Q. Wang, M. A. Abu-Rgheff and A. Akram, "Design and Evaluation of an Integrated Mobile IP and SIP Framework for Advanced Handoff

Management", in *Proc. of IEEE International Conference on Communications 2004 (IEEE ICC 2004)*, Paris, France, June 2004, pp. 3921-3925.

6. [Lopez etc QoS04] G. Gonzalez Lopez, Q. Wang, M. A. Abu-Rgheff and A. Akram, "A MIP-SIP Macro-Mobility Management Scheme for VoIP across Wired and Wireless Domains", in *Proc. of 2004 IEE Conference on Telecommunications Quality of Service: the Business of Success (IEE QoS 2004)*, London, UK, March 2004, pp. 114-118.

7. [Wang and Abu-Rgheff 3G2003] Q. Wang and M. A. Abu-Rgheff, "Integrated Mobile IP and SIP Approach for Advanced Location Management", in *Proc. of IEE 4th International Conference on 3G Mobile Communication Technologies (IEE 3G2003)*, London, UK, June 2003, pp. 206-210.

8. [Wang and Abu-Rgheff EPMCC03] Q. Wang and M. A. Abu-Rgheff, "A Multi-Layer Mobility Management Architecture Using Cross-Layer Signalling Interactions", in *Proc. of IEE 5th European Personal Mobile Communications Conference (IEE EPMCC 2003,* currently known as *European Wireless (EW))*, Glasgow, UK, April 2003, pp. 237-241.

9. [Wang and Abu-Rgheff WCNC03] Q. Wang and M. A. Abu-Rgheff, "Cross-Layer Signalling for Next-Generation Wireless Systems", in *Proc. of IEEE Wireless Communications and Networking Conference 2003 (IEEE WCNC 2003)*, New Orleans, USA, March 2003, Vol. 2, pp. 1084-1089.

10. [Wang and Abu-Rgheff LCS02] Q. Wang and M. A. Abu-Rgheff, "Towards a Complete Solution to Mobility Management for Next-Generation Wireless System", in *Proc. of London Communications Symposium 2002 (LCS 2002)*, London, UK, September 2002, pp. 281-284.

Word count of main body of thesis: 62,836

Signed ..Wang Qi.....

Date ..24 March 2006.....

Chapter 1

Introduction

We provide an introduction of the project and the thesis in this chapter, which is structured as follows. We start with the research motivations, followed by the aim and objectives. The major contributions are then summarised and the organisation of the thesis is outlined at last.

1.1 Motivations

The last decade has witnessed a tremendous boom of mobile communications despite the ups and downs in the business. According to a report by ITU (International Telecommunications Union) in 2002 [ITU2002], the number of mobile subscribers worldwide increased dramatically from 215 million in 1997 to 946 million in 2001, and it was predicated that the number would reach 1700 million by 2010. Nevertheless, this prediction turns out to be too conservative: the figure has reached 1800 million by the middle of 2005 and will reach 2140 million by the end of the same year according to a latest forecast [CE2005]. Meanwhile, the growth of the Internet access has experienced a similar striking process. There were 580 million Internet users worldwide in the summer of 2003, and the number will be 1350 million by 2007 [E-consultancy2005].

The already marvellous yet still fast-growing popularity of both the Internet and mobile communications necessitates the convergence of both technologies on a unified global network infrastructure with efficient and effective mobility support. Technically, the increasing prevalence of real-time and non-real-time applications based on the Internet

Protocol (IP) suite is a key driver for this convergence and facilitates the interworking of separate wireless platforms especially the third generation (3G) mobile systems being introduced and the wireless local area networks (WLAN) under rapid deployment. The convergence will glue heterogeneous access networks together over a uniform end-to-end IP platform collectively known as all IP networks, create a new communications paradigm sometimes referred to as mobile or wireless Internet and lead to a new communications era labelled as next-generation or beyond 3G (B3G) systems/networks.

Next-generation systems are being investigated in industry and academia, and in government and international standardisation bodies such as the Internet Engineering Task Force (IETF) and Third-Generation Partnership Project (3GPP). These activities reflect the fact that we are experiencing a significant change in communications paradigm and maybe life style. In the next-generation systems, it is expected that users will request higher-quality and higher-speed multimedia applications that are ubiquitous across geographical boundaries of heterogeneous networks and available across a range of devices using a single user-level identity for subscription convenience, among others. Such an increasing demand for "anywhere, anytime, multimedia" services is one of the fundamental challenges in the creation of the next-generation systems, and only advanced terminal and personal mobility support can enable users to obtain uninterrupted multimedia services independent of terminal type and point of attachment to the network.

To provide mobility support in such a context, numerous protocols have been proposed over the years. However, despite the achievements each of the proposals comes with its own disadvantages that hinder itself from satisfying all the requirements envisioned for the next-generation mobility support. Therefore, much more research is needed in this crucial area towards a more useful solution that is efficient in costs, effective in application performances and comprehensive in handling diverse mobility scenarios.

In particular, this project has been motivated by the following research questions:

What are the pros and cons of the existing and emerging mobility architectures and protocols? Is any of them sufficient to support mobility of diverse mobility scenarios expected in the next-generation all IP networks?

Can a single-layer mobility solution meet all the requirements of mobility support? If not, which layers should be involved, what contributions do these layers make, and how can the contributions be exploited in a uniform framework?

How can both real-time and non-real-time applications, despite their differentiations in traffic characteristics and QoS (Quality of Service) requirements, be supported efficiently and effectively in the same architecture?

What is (are) the most promising solution(s) to supporting global-scale terminal and personal mobility and preferably other mobility types as well?

Which is (are) the most promising micro-mobility protocol(s) considering the huge number of proposals that have already existed and are still emerging? Should a new protocol be designed or should the best candidate(s) be optimised for improved performances?

These questions were haunting in my mind during my start-up stage of this project, and I hope that my efforts have helped a clearer understanding to most of them.

1.2 Aim and Objectives of the Project

The main aim of this project is to explore and design efficient and effective mobility support architectures and protocols suitable for the vision of next-generation all IP wireless networks that are expected to deliver both real-time and non-real-time mobile communications in both global and regional scales.

The specific objectives are:

- To investigate the state-of-the-art work on IP mobility management and identify the advantages and disadvantages of existing and emerging architectures and protocols;
- To explore the contributions of protocol layers to mobility support, search cross-layer signalling methods for information exchanges along the protocol stack, and envision a multi-layer framework for complete mobility support;
- To design advanced architectures and protocols to support macro terminal and personal mobility regarding both real-time and non-real-time applications and facilitate other mobility types;
- To devise optimised architectures and protocols to support micro mobility, especially for high-mobility users with real-time applications in the IPv6 networking context;
- To evaluate the performances of the proposed mobility support architectures and protocols compared with existing approaches wherever appropriate through theoretical analyses and simulations.

1.3 Contributions of the Thesis

The major contributions of this thesis, including an introduction of the associated publications, are summarised as follows:

1. A novel vision of multi-layer mobility support is presented.
    - A critical review of related work on mobility support from both generation evolution and protocol stack perspectives. This review is an extended version of a publication [Wang and Abu-Rgheff LCS02], which also strongly indicates a close integration of Mobile IP (MIP) and the Session

Initiation Protocol (SIP) as a solution towards a complete mobility support for the next-generation networks based on a brief survey.

- Cross-layer signalling methods are explored and especially a new efficient and generic method called CLASS (Cross-LAyer Signalling Shortcut) is proposed. This topic had been barely addressed in the literature before the publication of the associated paper [Wang and Abu-Rgheff WCNC03], which has been highly referenced in the cross-layer design community.
- Contributions of each protocol layer to mobility support are identified and a multi-layer mobility support framework is envisioned. The novelty of this framework is that it attempts to exploit the contributions from multiple layers to advanced mobility support in contrast to the dominant single-layer approach in mobility protocol design. Cross-layer signalling methods are utilised as vehicles to exchange mobility-related information vertically across protocol stack. The associated publications are [Wang and Abu-Rgheff CE, EPMCC03].

2. Two macro-mobility architectures based on novel integrations of MIP and SIP are designed and evaluated.

- The first macro-mobility architecture, TI-MIP-SIP (Tightly Integrated MIP-SIP), is proposed based on a tight integration of MIP and SIP to achieve terminal and personal mobility for long-term cost-effectiveness. In the TI-MIP-SIP architecture, MIP and SIP mobility entities and procedures of similar functionality are merged to minimise the redundancies found in emerging hybrid MIP-SIP architectures and maximise the efficiency. Both real-time and non-real-time applications are effectively handled during handoffs, and the support for both terminal and personal mobility is

achieved. In addition, the designs are applicable to both IPv4 and IPv6. The associated publications are [Wang and Abu-Rgheff IJCS, 3G2003]. In [Wang and Abu-Rgheff 3G2003], the architectural integration methodology is specified and the reusing of MIP and SIP messages is proposed, among other broad discussions. In [Wang and Abu-Rgheff IJCS], the detailed designs are presented, and the analytical and simulation results are reported.

- The alternative macro-mobility architecture, LI-MIP-SIP (Loosely Integrated MIP-SIP), is built upon a loose integration of MIP and SIP to achieve a trade-off between performance improvements and deployment convenience. Unlike TI-MIP-SIP, LI-MIP-SIP establishes necessary interactions between MIP and SIP entities instead of fully integrating them physically. Two schemes are devised to achieve different yet similarly effective interactions and lead to different signalling designs. The preliminary design and analysis were published in [Wang and Abu-Rgheff 3G2004] whilst an updated and extended version is in preparation for a journal publication.
- Support for other mobility scenarios is facilitated. Though the design of the macro-mobility architectures is focused on terminal and personal mobility, the support for other mobility types is facilitated in the integrated MIP-SIP architectures. This topic is briefly discussed in the thesis and detailed research remains as a future work. In [Wang, Abu-Rgheff etc ICC04], an initial policy-based mobility table is proposed to automatically detect and execute the diverse mobility operations. The same paper also presents the

signalling integration philosophies of MIP and SIP with a route optimisation option proposed for more reliable and faster macro handoffs.

3. A new micro-mobility architecture combining the merits of hierarchical MIP and fast handoffs with optimisations is devised and assessed.

- This architecture is built upon a cost-driven optimised combination of HMIPv6 (Hierarchical MIPv6) and FMIPv6 (Fast Handovers for MIPv6) with a set of optimisation algorithm and mechanisms introduced, though many of these optimisations are applicable to the IPv4 version of hierarchical MIP and fast handoff protocols. The architecture is optimised for high-mobility users with real-time applications demanding explicit QoS support.
- The following optimisations are included in the architecture. Firstly, a new scheme called PAVER (Prompt Address Verification and complEment Replacement) is designed to accelerate IPv6 address auto-configuration by removing the bottleneck of handoff delays safely. Secondly, a dynamic interaction with IP QoS signalling protocols is explored to balance the costs of QoS route extension and QoS route optimisation. Thirdly, a mechanism called REED (Route Extension End Declaration) is introduced to eliminate out-of-sequence packets due to the joint use of HMIPv6 and FMIPv6. Lastly, an algorithm to combine the enhanced HMIPv6 and FMIPv6 with all the optimisation efforts incorporated.

The associated publication is [Wang and Abu-Rgheff 3G2005] and another journal publication is under preparation.

1.4 Organisation of the Thesis

The organisation of the thesis, together with the associated publications, is depicted in Figure 1.1 and described as follows.

In Chapter 2, we provide the background information on mobility management and detailed review of the state-of-the-art work. In the first half of the chapter, we present the reference protocol stack, the concepts of handoff and location management, the retrospect on mobility evolution and the overview of mobility management in all IP networks. In the second half, we scrutinise the existing and emerging architectures and protocols for mobility support in all IP networks. These architectures and protocols are largely classified into two categories: macro mobility and micro mobility, and the representative members of each category are expounded and compared in details. In addition, the cross-layer design methodology is introduced and existing cross-layer signalling methods are explored.

Chapter 3 presents the big picture of the proposed multi-layer framework for future-generation mobility support. The requirements and design challenges are identified and a cross-layer design methodology is advocated. A new cross-layer signalling method called CLASS is proposed as a generic and efficient scheme to facilitate cross-layer design. Subsequently, the contributions to mobility support from each protocol layer are specified, and a multi-layer mobility support framework is thus envisioned for exhaustive mobility support with CLASS and other cross-layer schemes utilised. Finally, the focus of this project is stated.

In Chapter 4, we propose TI-MIP-SIP, a macro-mobility support architecture that tightly integrated MIP and SIP. Firstly, we investigate the hybrid MIP-SIP architectures, which are emerging in parallel with this project, and identify their shortcomings. Secondly, the architectural integration issues are discussed, resulting in the design of a merged mobility server and a uniform address management. Thirdly, the protocol signalling

designs are presented for handoff and location management, respectively. Fourthly, the support for other mobility types is discussed. Finally, the costs of the proposed architecture are analysed and compared with other architectures including its hybrid counterpart, the pure SIP approach and the pure MIP approach through analytical and simulation results.

As an alternative macro-mobility support approach, a loosely integrated MIP-SIP architecture called LI-MIP-SIP is defined in Chapter 5. The enhancements to mobility servers and the introduced interactions between them are justified and presented. Subsequently, two schemes are proposed to establish the necessary interactions and the corresponding protocol signalling designs are expounded. Performances of LI-MIP-SIP, in contrast to TI-MIP-SIP and other architectures, are then evaluated under more metrics in addition to costs through analyses and simulations. The advantages and disadvantages of each joint MIP-SIP approach including TI-MIP-SIP, LI-MIP-SIP, their hybrid counterpart, MIP and its variants are summarised and discussed in the end. Notably, both IPv4 and IPv6 contexts are investigated in both TI-MIP-SIP and LI-MIP-SIP architectures.

To complement the macro-mobility proposals, we propose a micro-mobility architecture in Chapter 6. The chapter is focused on the IPv6 context, and it begins with the problem statement identifying the shortcomings of HMIPv6 and FMIPv6, and existing efforts to combine both protocols, among other related work. Next, an overview of the proposed system is presented, followed by the design details on the proposed optimisation and integration. The interactions of this micro-mobility architecture with the proposed macro-mobility ones are then discussed. At last, the analytical and simulation results confirm the performance improvements in contrast to the standard HMIPv6, FMIPv6, and a couple of other combination approaches. Note that much of the integration and optimisation methodologies employed for the HMIPv6 and FMIPv6 context could be applicable to their IPv4 counterparts.

Finally, Chapter 7 concludes the thesis. A summary is provided, followed by our contributions to knowledge. We also identify the limitations of the current work and future work directions. Conclusions are drawn in the end.

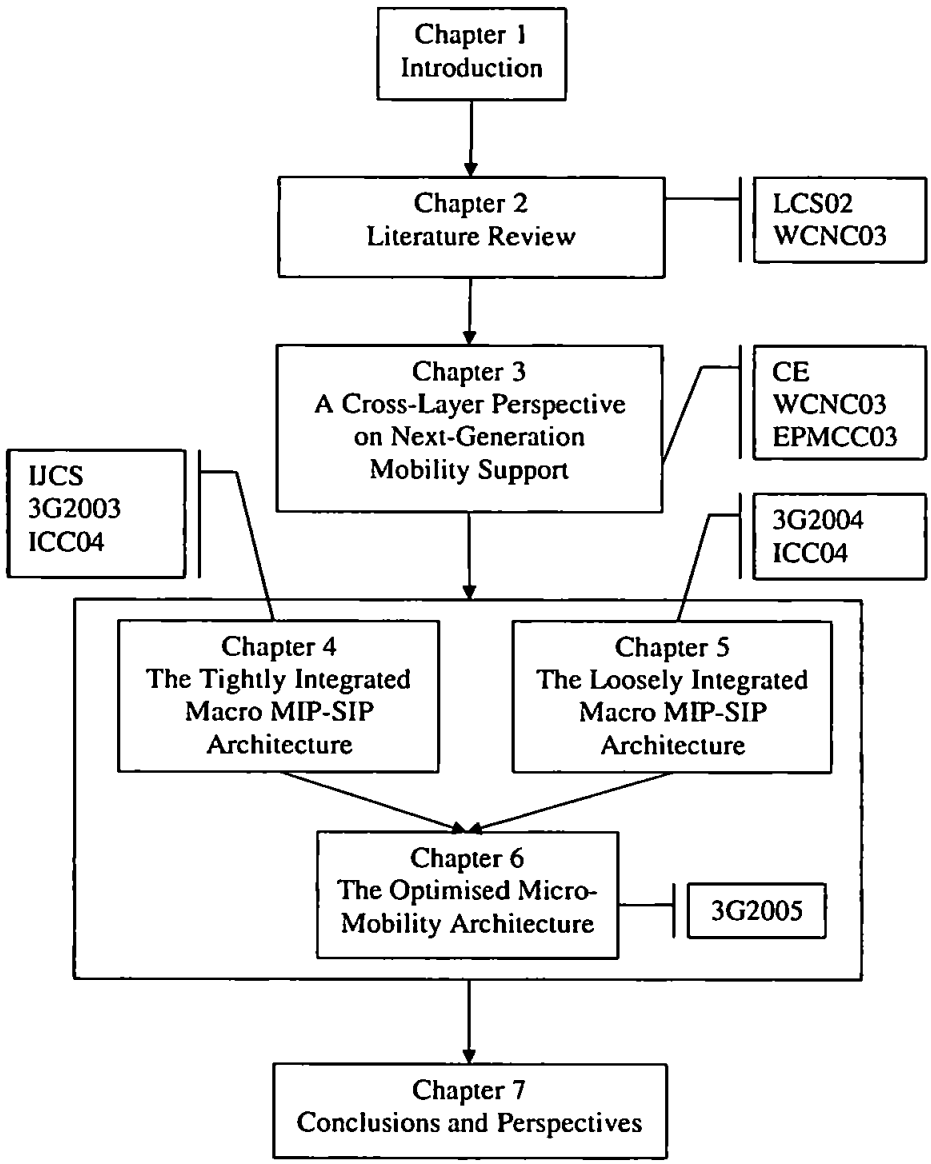


Figure 1.1 Organisation of the Thesis

Chapter 2

Literature Review

In this chapter, we critically review the literature on existing and emerging mobility support protocols and architectures. This chapter is partially based on two publications [Wang and Abu-Rgheff LCS02, WCNC03].

2.1 Introduction

This chapter presents the background work on mobility support, and the up-to-date versions of the protocols are reviewed at our best effort to reflect the state of the art. Additional relevant research emerging during the progress of the project is surveyed in the subsequent chapters wherever appropriate, mainly in the Related Work sections.

The remaining of the chapter is organised as follows. The reference protocol stack used in the thesis is presented in Section 2.2, followed by an overview of mobility management concepts in Section 2.3. In Section 2.4, we retrospect the evolutionary development of the mobility management in the past and current generations of wireless systems. In Section 2.5, we introduce the all-IP-based next generation wireless systems, and the concepts of macro- and micro-mobility management. The focus of this chapter is the survey of the typical macro- and micro-mobility management protocols, which are expounded in Sections 2.6 and 2.7, respectively. Additionally, in Section 2.8 the emerging cross-layer design methodology is introduced as a promising approach to tackle complex problems including mobility support in wireless networks. Finally, the summary is given in Section 2.9.

2.2 Reference Protocol Stack

We start with a generic reference protocol stack [Tanenbaum 1996], together with the major functions of each layer, as shown in Figure 2.1. This reference protocol stack comprises, from bottom to top, five layers: the physical layer (L1), the link layer (L2), the network layer (L3), the transport layer (L4) and the application layer (L5). The main functions of each layer are discussed as follows.

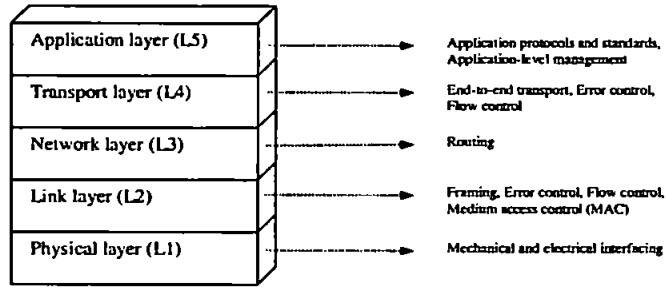


Figure 2.1 Reference protocol stack

The physical layer is the bottom layer of the protocol stack. It defines the mechanical and electrical interfaces, and deals with the underlying physical transmission medium such as copper wire, fibre optics and wireless links. The information unit is bit ('0' or '1') in this layer as the data streams are transparent to it.

The link layer delimits the input bits into frames whose sizes are usually a few hundred bytes. Once the framing is done, the link layer can perform error correction and flow control to ensure correct frame delivery at an appropriate speed between the sender and the adjacent receiver. Moreover, the MAC (medium access control) sub-layer in L2 is needed to allocate multi-access channels. Examples of MAC protocols are CDMA (Code Division Multiple Access), TDMA (Time Division Multiple Access), FDMA (Frequency Division Multiple Access), etc.

The main function of the network layer is routing packets from the source to the destination across multiple hops. In the Internet context, the Internet Protocol (IP) [RFC791 for IPv4, RFC2460 for IPv6] enables global routing with well-defined packet header and routing table in the routers. Each IP packet header contains the source IP address and the destination IP address, among other information. Each router along a packet's journey consults its routing table to determine the next hop of the packet. To control and optimise the basic IP routing, Internet traffic engineering is introduced and expected to steer traffic through the network in the most effective way [RFC3272].

As aforementioned, the transport layer provides end-to-end transport and flow control, compared with the L2 point-to-point functions. Protocols in this layer provide either a "reliable" or an "unreliable" transport service to the applications running in the upper layer. In the Internet-protocol suite, the two dominant IP-based transport protocols are TCP (Transmission Control Protocol) [RFC793] and UDP (User Datagram Protocol) [RFC768]. TCP is a reliable transport protocol, attempting to deliver correct, complete and in- order packets to the applications running over it. Through TCP, lost packets are retransmitted, corrupted packets are either corrected or retransmitted, and duplicate copies are eliminated. Furthermore, TCP has built-in flow control using packet loss as the indication of congestion and throttles its sending rate to alliviate congestion. These features allow TCP appeal to non-real-time applications, which usually requires reliable transmissions. However, such features can become disadvantages for real-time applications especially in error-prone environments like a wireless network. In contrast, UDP is an unreliable transport protocol because it does not verify that packets have reached their destination, and offers no guarantee that they will arrive in order. If an application requires these guarantees, it should provide them itself or use TCP if no addon mechanisms are available. UDP, together with RTP (Real-time Transport Protocol) [RFC3550], is typically adopted

by real-time applications such as audio and video, where the delay or jitter caused by TCP retransmission, re-ordering or flow control would render TCP unusable. For either TCP or UDP, the applications at any given IP address are distinguished by their TCP or UDP Port Number. By convention certain well known ports are associated with specific applications.

Finally, we move to the top layer, the application layer, where most common network programs reside. Popular Internet programs and their corresponding protocols include the Hypertext Transfer Protocol (HTTP) [RFC2616] for the World Wide Web (WWW), the File Transfer Protocol (FTP) [RFC959] for network file copying, the Simple Mail Transfer Protocol (SMTP) [RFC821] for Email, etc. Also running here are application-level management protocols, such as the Session Initiation Protocol (SIP) [RFC3261] for managing realtime applicatins especially VoIP (Voice over IP). In addition, the multimedia source-coding standards, such as the MPEG-x [MPEG], the G.7xx and H.26x [ITU-T] series, belong to this layer.

2.3 Overview of Mobility Management

Mobility is a unique and the most important characteristic of wireless mobile systems, distinguishing themselves attractively from wired systems like PSTN (Public Switched Telephone Network). Consequently, mobility management is fundamental to the proper operation of wireless systems. Mobility management includes two essential tasks, namely location management and handoff (or handover) management [Akyildiz etc 1998], corresponding to the idle mode and the active (or busy) mode of a mobile host (MH)[1], respectively. An MH is in the idle mode when it is powered on whereas not involved in any ongoing sessions (or calls in conventional voice-centric wireless systems). On the other hand, an MH is in the active mode when it is powered on and involved in one or

[1] A mobile host (MH) is also referred to as a mobile terminal (MT), a mobile station (MS), a mobile node (MN), or a user equipment (UE). These terms can be used interchangeably.

more sessions in progress. Either location or handoff management generally involves a number of procedures to fulfil their tasks. Figure 2.2 presents the mobility management modes and the involved procedures to be discussed as follows.

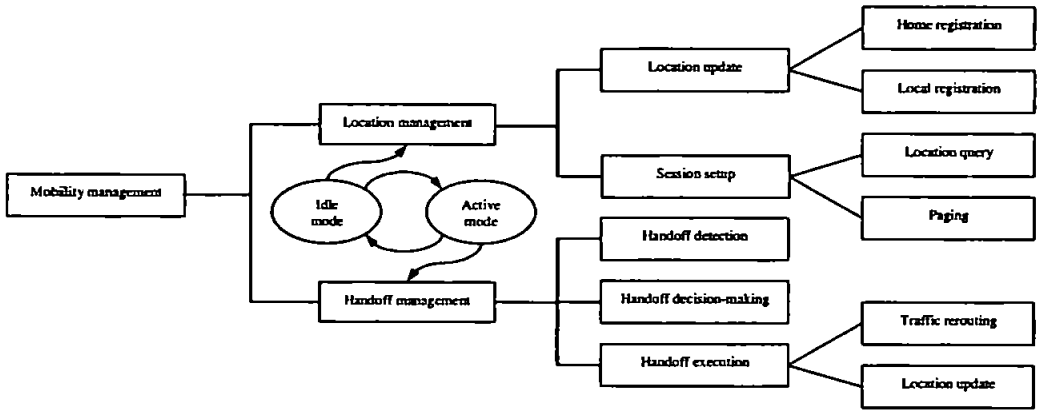


Figure 2.2 Mobility management modes and procedures

2.3.1. Location Management

Location management tracks and locates an MH for the delivery of incoming sessions, and thus involves a location update (or registration) procedure at the involved mobility server(s) for tracking the registered MHs and a session setup procedure for session delivery. The location information, among other information, is usually stored in hierarchical databases of the mobility servers in the home network and the foreign (or visited) networks. The location update at a home or a local mobility server is thus also known as home or local (or regional) registration, respectively. For session setup, the location databases are enquired to locate a targeted MH. Paging is often needed when the network only maintains approximate location information of the invited MH. Generally, the more frequently an MH performs location updates, the more accurately the network can track the MH. Therefore, there is a trade-off between location-update costs and paging costs, and different algorithms have been designed to minimise the overall costs of location management [Wong and Leung 2000]. Notably, a link-layer paging procedure is usually

available in wireless systems, and thus an upper-layer location-management protocol may simply utilise the existing L2 procedure rather than introduce a new one.

2.3.2. Handoff Management

The task of handoff management is to enable an ongoing session to continue as the MH changes its network attachment (e.g., base station, access point, or access router) or channel. Typical criteria to trigger a handoff include deterioration in quality of the signal and user movement. A handoff consists of a series of processes: handoff detection, handoff decision-making and handoff execution. In the handoff detection stage, measurements (or other monitoring) are taken periodically to compare the signal quality and detect movement. Based on these measurements, a handoff decision is made in the next stage. Once a handoff is determined, the handoff execution process is initiated. In this process, the traffic of the on going sessions is re-routed to the new attachment or channel, and subsequently location updates may be conducted if the re-routing is not achieved through such location updates. Note that a handoff in progress may be aborted due to lack of resources in the targeted network attachment, repeated signalling retransmission failure, user termination etc. For discussion brevity, we usually assume that a handoff is not aborted unless otherwise specified.

There are several perspectives to classify handoffs, as shown in Figure 2.3. Firstly, a handoff can be performed between two channels in a cell (intra-cell handoff) or between two adjacent cells (inter-cell handoff). In the latter category, further handoff types can be specified depending on the locations of the old and the new cells. For instance, a handoff can be an intra-subnet handoff (the two cells belong to the same subnet), inter-subnet or intra-domain handoff (between two subnets within an administrative domain), or inter-domain handoff (between two domains) in a hierarchical system. In addition, a handoff

across two systems of different radio technologies is called inter-system (or vertical) handoff, and the complementary scenario is called intra-system (or horizontal) handoff.

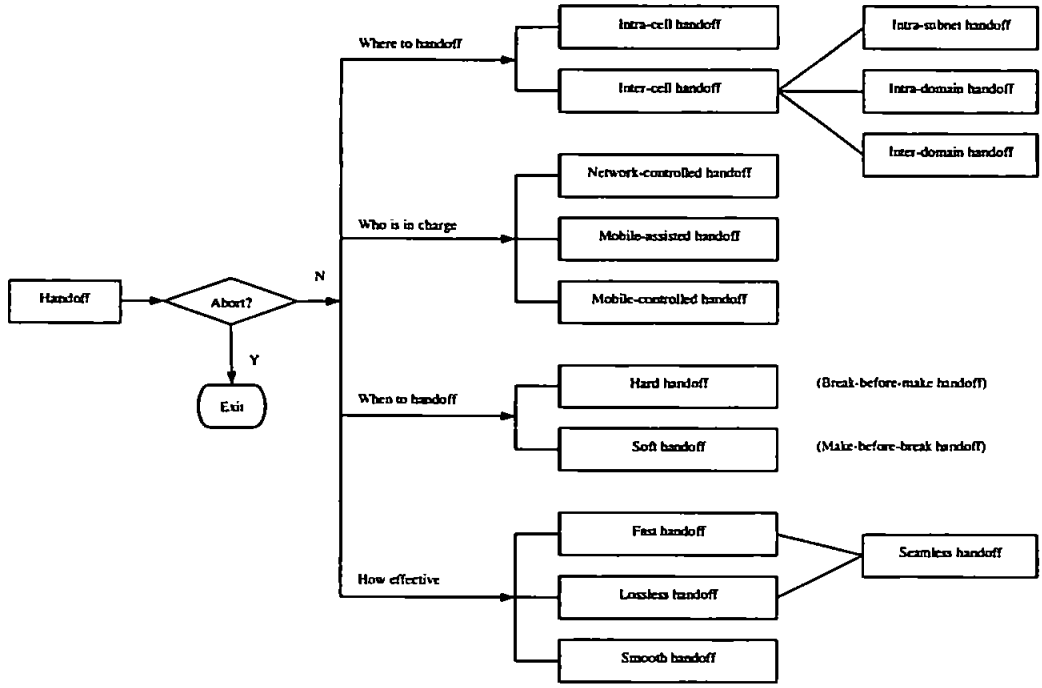


Figure 2.3 Handoff classifications

Secondly, a handoff can be controlled by the network, or the MH, or both; accordingly, there exist network-controlled handoff (NCHO), mobile-assisted handoff (MAHO), and mobile-controlled handoff (MCHO), following the order that the handoff decision-making responsibility is decentralised [Tripathi etc 1998]. In a system that adopts an NCHO protocol, information about the signal quality for all MHs is centralised in the network, and the network performs handoff detection and makes handoff decision. In an MAHO protocol, an MH detects a handoff whilst the network makes the decision. For an MCHO, an MH completely controls the handoff processes, and thus can handle frequent handoffs more promptly. With such a growing degree of handoff decentralisation, the time required to execute a handoff request decreases though the available information for handoff decision also decreases.

Thirdly, in certain systems an MH can be enabled to communicate with one or more candidate network attachments in addition to the current one simultaneously, so that the ongoing sessions are not interrupted at all upon a handoff. This is known as soft handoff, which is a kind of make-before-break handoff. On the contrary, if the connection to the new network attachment can only be setup after the old one is released, i.e., break before make, the handoff is called hard handoff. Note that soft handoff is commonly found in systems based on CDMA link-layer technology [Wong and Lim 1997], though the make-before-break behaviour could be mimicked at upper layers in other systems.

Finally, handoffs can be categorised according to their effectiveness. In a fast handoff, the time that an MH is unable to receive incoming session traffic at its new attachment is minimised. If no traffic is lost during a handoff, it is a lossless handoff. Note that a lossless handoff is not necessarily a fast handoff, and vice versa. If a handoff is both lossless and fast, it is called a seamless handoff. For real-time applications, when the traffic loss is low enough and the handoff is fast enough the end user may hardly notice the service disruption or degradation during the handoff. For this reason, this kind of handoff is sometimes also referred to as a seamless handoff [Malki etc 2004]. At last, a smooth handoff is achieved if the session traffic during the handoff is buffered at the old attachment and transferred to the new one so that traffic loss is minimised.

2.4 Evolution of Mobility Management

Wireless systems have been developed in an evolutionary way generation by generation over the last twenty years or so. The first-generation (1G) systems are of diminishing importance. The dominant generations today are the second generation (2G) and their enhancement (2.5G) with the third generation (3G) under initial deployment. These generations are represented in Europe by GSM (Global System for Mobile

communications), GPRS (General Packet Radio Service) and UMTS (Universal Mobile Telecommunications System), respectively.

In 2G (e.g., GSM), a wireless network consists of multiple location areas (LAs), each of which comprises a set of cells. One or more LAs are under control of a Mobile Switching Centre (MSC) and a Visitor Location Register (VLR), which are usually co-located and collectively referred to as MSC/VLR. When an MH moves within an LA in the idle mode, no location update is needed. When it travels into a new LA, an MH report its new location to the serving VLR. If the VLR is also in charge of the new LA, no further location update is invoked; otherwise, the VLR performs a home location update at the MH's Home Location Register (HLR) on behalf of the MH. In a call delivery procedure, a calling correspondent host (CH) sends a call initiation message to its own MSC/VLR, which then requests a call setup between the MSC/VLR of the called MH and itself, through the help of the HLR of the called MH. Finally, the called MSC/VLR pages the called MH, and the MH replies to receive the call. These mobility management functions are achieved by the exchange of the MAP (Mobile Application Part) messages. Minimising the signalling traffic for location management is the focus of the related research [Akyildiz etc 1998]. For handoff management in GSM, an MH keeps on measuring the received signal strength (RSS) and requests for a handoff when the RSS is below a predefined threshold whilst the handoff decision is made by the network side (e.g., the MSC). Thus, the GSM employs an MAHO protocol.

The GPRS system is evolved from GSM by updating the existing GSM entities (HLR, MSC/VLR, the base station subsystem etc) as well as introducing a couple of new core-network entities called Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). Correspondingly, the core network is enhanced by a packet-switched (PS) domain in addition to the GSM circuit-switched (CS) domain. For PS services, a session

context based on the Packet Data Protocol (PDP) needs to be created, and SGSNs take over the role of MSCs for mobility management. The location area unit employed in GPRS is called routing area (RA), which is typically a subset of one (and only one) GSM LA. This smaller granularity allows for signalling and paging over smaller areas, and thereby achieves a better optimisation of radio resources. GPRS co-operates with the GSM LA-based location management, resulting in a more efficient paging mechanism for MHs that use GSM and GPRS simultaneously [Lin etc 2001]. By introducing the PS domain and services, GPRS paves the way towards the adoption of IP mobility. However, in principle, the mobility management of 2G and 2.5G are both link-layer based and for terminal mobility only.

Again, UMTS is built upon GPRS, though it is advancing towards an all-IP vision [Patel and Dennett 2000]. An IP Multimedia Subsystem (IMS) is introduced to support real-time multimedia IP services through SIP. Regarding mobility management, the link-layer solution is further improved. In contrast to the monopoly role of a GPRS SGSN, a UMTS SGSN shares mobility management with the UMTS Terrestrial Radio Access Network (UTRAN) under its control so that signalling loads can be distributed between the core network and the access network. For further thinning location management requirements, GPRS RAs are in turn partitioned into URAs (UTRAN Registration Areas) for pico cells. As a summary, we can express as follows the location area relationship between the three generations: GSM LA $\supseteq$ GPRS RA $\supseteq$ UMTS URA $\supseteq$ cell. In 3G, global roaming becomes more practical with GSM, GPRS and UMTS co-existing to cover a global area. The evolution approach of cellular generations, cumbersome as it is in a sense, facilitates the mobility management of the hybrid system. Inter-operator roaming is proposed by the GSM Association based on a framework called GPRS Roaming eXchange (GRX) [GSMA34]. In GRX, GPRS and UMTS networks are interconnected to create a tier

of the Internet, through which carriers can exchange IP traffic securely that is generated by roaming MHs.

Current research is focused on further reducing the signalling loads and improving the capabilities for handoffs among GSM, GPRS and UMTS (inter-system handoffs). In [Lin and Chlamtac 2001], a middleman is introduced between the home network and the visited network. It acts as HLR for the visited network and as VLR for the home network to reduce signalling response time and latter stage traffic. In [McNair etc 2000], an inter-system handoff scheme was proposed based on the analysis of the boundary cell region between systems. This scheme seems to bring the systems together quite naturally by taking advantage of their existing handoff procedures. However, it is an indirection solution and the extra signalling time introduced has a very large impact on the overall handoff time for pico and micro cells. Ref. [Kaaranen etc 2001] illustrated the UMTS-GSM handoff procedure based on some modifications of GSM to facilitate the discovery of UMTS. This approach leads to a more direct solution. Notably, all these mobility management capabilities are achieved through link-layer mechanism and signalling.

2.5 Mobility Management in All-IP Networks

In this section, we present the all-IP vision of the next-generation wireless system and the corresponding mobility management, followed by an introduction to the concepts of macro- and micro-mobility management in all-IP networks.

2.5.1. Vision of Next-Generation and All IP Mobility Management

In a wider context, in addition to the wide-area cellular systems there are a number of other emerging wireless systems, such as the wireless local area networks (WLANs), and

the personal area networks (PANs) using, e.g., the Bluetooth technology. Table 2.1 lists some typical wireless systems [Aretz etc 2001, Vidales etc 2004].

Table 2.1 Diversity in wireless systems

System	*Data rates*	*Range*	*Mobility*	*Cost*	*Deployment environment*
GSM/GPRS	9.6 to 384 kbps	≤ 35 Km	High	High	Wide area network
UMTS	≤ 2 Mbps	≤ 20 Km	High	High	(WAN)
IEEE 802.11b	≤ 11 Mbps	50 ~ 300 m	Medium	Low	Local area network
IEEE 802.11a	≤ 54 Mbps	50 ~ 300 m	Medium	Low	(LAN)
Bluetooth	≤ 721 Kbps	0.1 ~ 100 m	Low	Low	Personal area network (PAN)

As seen from Table 2.1, these systems are optimised for different ranges, different mobility scenarios, different deployment environments and different applications requiring various data rates at different costs. Therefore, there is a strong potential for them to co-operate in a complementary way so that multimedia mobile communications can be expected anytime and anywhere. It would be cost-effective to achieve this aim through convergence of various wireless systems. IP is widely recognised as the most suitable L3 technology to integrate all the different systems, each of which is featured with distinctive L2 technology. In another words, as these wireless systems evolving to carry both real-time and non-real-time services, an all-IP-based system akin to the Internet is likely to be the most favourable choice [Evans and McLaughlin 2000, Patel and Dennett 2000, Bos and Leroy 2001, Berezdivin etc 2002, Sami 2003, Mahonen etc 2004]. The motivations for choosing IP lie in the expectations that an all-IP-based system can be better suited to support rapidly growing mobile data and multimedia applications, to bring the successful Internet service paradigm to mobile providers and users, and to glue diverse radio access networks seamlessly and render services transparently across systems.

In such an all-IP vision, all kinds of wireless (and wired) access networks are centred on a common IP-based core network, e.g., the next-generation high-performance Internet. The degree towards all IP in access networks will increase over time [Wu etc 2002], as is

the case found in the phased 3G cellular systems, e.g., UMTS (terrestrial and satellite subsystems). The first pure all-IP wireless access networks are wireless local area networks (WLANs), which have emerged in public wireless networks. In addition, IP-based wired or fixed networks have been widely deployed. Across these IP-enabled ubiquitous and heterogeneous networks, a mobile user would be able to roam seamlessly with advanced IP-based mobility management. Figure 2.4 shows such a perspective based on [Kari 1999].

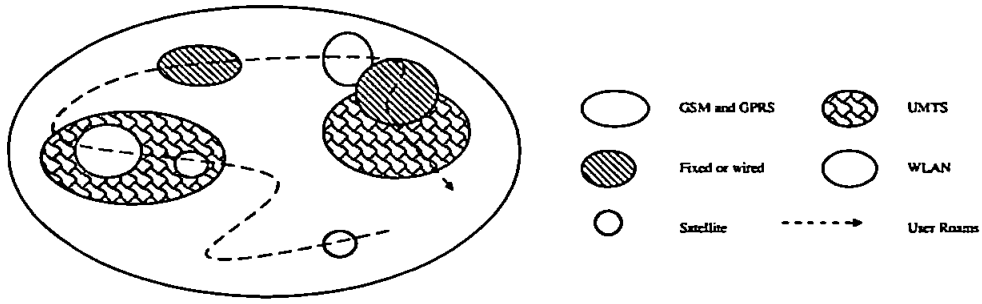


Figure 2.4 Seamless roam over ubiquitous and heterogeneous networks

2.5.2. Overview of Macro and Micro Mobility Management

For effective mobility management, the IP-based access networks are organised into domains, particularly administrative domains [Das etc 2000, Ramjee etc 2002 A], which are inter-connected through a global IP-based wired core network. A domain usually consists of several subnets, each of which is covered by an access router (AR). Zero or more L2 access points (APs) may be connected to an AR. Mobility between two APs under the same AR is handled by the system-specific L2 mobility protocol. IP mobility corresponds to movements between two ARs, though L2-L3 interactions may exist. Generally, IP-based mobility protocols fall in two broad categories: macro-mobility (or inter-domain) protocols and micro-mobility (or intra-domain) protocols, corresponding to macro mobility and micro mobility, respectively. Macro mobility refers to a movement of an MH from one AR belonging to one domain to another AR managed by another domain.

In contrast, micro mobility refers to a movement between ARs that are in the charge of a same domain. Figure 2.5 demonstrates the differentiation of both mobility scenarios under a generic network model.

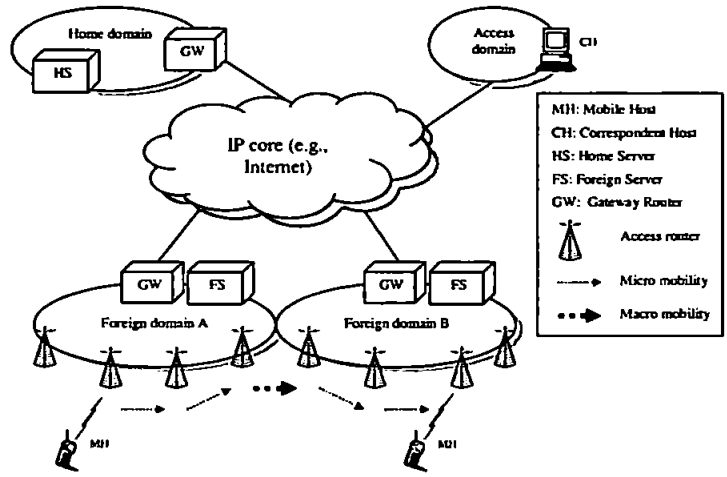


Figure 2.5 Differentiation of macro and micro mobility

Generally, an MH initially registers with a home server (HS) at its home domain. An HS is analogous to a HLR in GSM and maintains the up-to-date location, among other information possibly, of all the MHs it servers. This is achieved by location updates at the HS (home registrations) performed by an MH when changing the network attachment (i.e., AR) outside of the home domain. One of the major differences between a macro-mobility protocol and a micro-mobility one is when a home registration has to be triggered. In a macro-mobility protocol, an MH conducts home registration each time it changes an AR even within a domain. This incurs considerable global signalling overhead and handoff delay since a foreign domain is typically far away from the home domain. On the contrary, in a micro-mobility protocol (working together with a macro-mobility protocol), at least one foreign server (FS), acting as local HS, is introduced into a foreign domain to limit home registrations. On a micro mobility, an MH just reports its new location to the FS, which keeps tracking the MH as long as it moves within the domain. The HS merely

knows the domain-level address of the MH. An MH performs a home registration only on macro mobility, when the micro-mobility protocol triggers its macro-mobility partner. Note that the home and foreign servers usually have different names in different protocols, and they always refer to the concerned MH if not stated explicitly.

2.6 Macro-Mobility Protocols

In this section, we review a number of typical macro-mobility protocols, especially the two dominant approaches: the Mobile IP (MIP) family and the mobility support based on the Session Initiation Protocol (SIP). A few alternate protocols are also briefly discussed.

2.6.1. Mobile IP (MIP)

In this subsection, we present the detailed macro-mobility protocols under the Mobile IP umbrella, including Mobile IPv4 (MIPv4), Mobile IPv4 with Route Optimisation (MIPv4-RO), and Mobile IPv6 (MIPv6).

2.6.1.1. Mobile IPv4 (MIPv4)

Mobile IPv4 (MIPv4) [RFC3344] is the current de facto standard for IP mobility management. Two mobility-aware routers, called Home Agent (HA) and Foreign Agent (FA), are introduced to the home network and the foreign networks of an MH, respectively. An MH is assigned a long-term home address (HoA) on its home network and a dynamic care-of address (CoA) that is topologically correct in each foreign subnet. Every time it changes the CoA, an MH needs to perform a location update at its HA (home registration) by sending the new HoA-CoA binding to its HA, which maintains a built-in mobility binding list containing the up-to-date HoA-CoA bindings of all the MHs it is serving. In the following, we describe the MIPv4 mobility overview and the detailed signalling and data flows, which are also shown in Figure 2.6 and Figure 2.7, respectively.

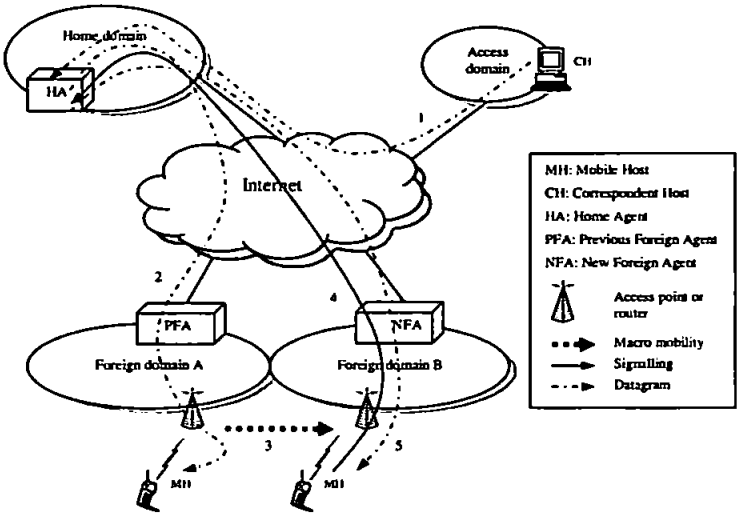


Figure 2.6 MIPv4 mobility overview

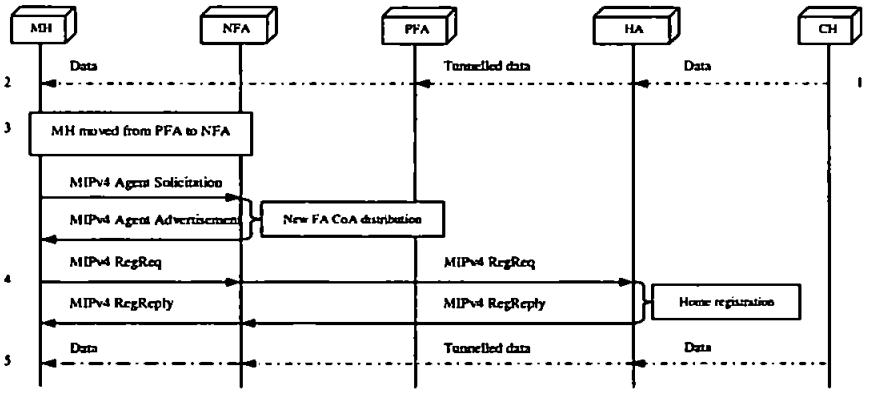


Figure 2.7 MIPv4 signalling and data flows

For a downlink communication (from a CH to an MH), a CH sends IP datagram to the HoA of the destination MH (Step 1 in Figure 2.6 and Figure 2.7) as an MH is known to all its CHs by its HoA. The HA of the MH then intercept the datagram, and check its binding list for the MH's current location. If the MH is in the home network, the HA simply forwards the datagram to the MH using standard IP routing. Otherwise, the HA tunnels the datagram to the MH's CoA using IP-in-IP encapsulation or other alternate methods (Step 2). Briefly, the downlink communication (from a CH to an MH) must pass by the MH's HA, and this is known as triangular routing. In contrast, for an uplink communication

(from an MH to a CH), generally the MH can send datagram to the CH directly using standard IP routing, though reverse tunnelling via the HA is also enabled.

On entering a new foreign subnet (belonging to the same or different foreign domain), an MH detects the movement on receiving an Agent Advertisement message multicast (or limitedly broadcast) periodically by the new FA (NFA). Optionally, an MH can send an Agent Solicitation message to the NFA, which also responds an Agent Solicitation with an Agent Advertisement (Step 3). Through either method, the NFA distributes an FA CoA (the new FA's own IP address) to an MH. Alternatively, a non-MIP mechanism such as a DHCPv4 (Dynamic Host Configuration Protocol version 4) [RFC2131] server can be used to configure a local IP address, called co-located CoA, for an MH. Upon obtaining the CoA, the MH performs home registration by sending a Registration Request (RegReq) message to its HA via the NFA if it uses an FA CoA (Step 4). After a successful home registration, the HA replies with a Registration Reply (RegReply) message and tunnels the subsequent datagram to the MH's current CoA (Step 5). The NFA or the MH itself then decapsulates the incoming tunnelled data, depending on the type of the CoA (FA CoA or co-located CoA). Figure 2.6 and Figure 2.7 show the scenarios when FA CoAs are used, and thus the previous FA (PFA) or the NFA detunnels the incoming packets for the MH in the previous and the current subnets, respectively. Note that the same signalling serves as the location update procedure when the MH is in the idle mode. In addition, as each mobility binding has a lifetime, an MH needs to send new RegReq messages of the same binding periodically to its HA to refresh the binding that is due to expire. When returning to its home domain, an MH needs to deregister its CoA at the HA.

In sum, the base MIPv4 is designed to support network-layer mobility in a transparent way so that the mobility (changes of IP addresses) can be hidden from the upper layers. Such mobility transparency is achieved by using a CoA for IP routing and delivering

datagram identified by the HoA to the upper layers. This is especially useful for TCP-based applications, which must maintain unchanged IP addresses during the sessions' life. Briefly, MIPv4 provides a simple and useful transparent mobility, though the triangular routing compromises the routing efficiency.

2.6.1.2. Mobile IPv4 with Route Optimisation (MIPv4-RO)

In the MIPv4 with route optimisation (MIPv4-RO) [Perkins and Johnson 2002, Perkins 1997], the base MIPv4 protocol is extended to support direct CH to MH transmission via binding cache management and to support smooth handoff between the previous and the new FAs. Figure 2.8 and Figure 2.9 illustrate the mobility overview and the detailed signalling and data flows in MIPv4-RO, respectively.

In MIPv4-RO, a CH maintains a binding cache containing the HoA-CoA binding of one or more MHs. Before sending an IP datagram to an MH, a CH checks its binding cache first. If a valid binding for the MH is available, the CH can tunnel the datagram directly to the MH's CoA. Otherwise, the CH sends the datagram to the MH's HoA as in MIPv4 (Step 1 as shown in Figure 2.8 and Figure 2.9). On intercepting the datagram, the MH's HA deduces that the CH does not have a valid binding for the MH and thus sends a Binding Update (BU) message to the CH, indicating the current HoA-CoA binding (Step 2). Meanwhile, the HA tunnels the received datagram to the MH's CoA (Step 2'). On receiving the BU, the CH creates a binding cache entry (or updates an existing entry) for the MH, and tunnels the subsequent datagram to the MH's CoA directly (Step 3).

On a handoff from the previous FA (PFA) to the new FA (NFA), the MH obtains a new CoA as defined in MIPv4 (Step 4). It then sends a Registration Request (RegReq) with an optional Previous Foreign Agent Notification (PFAN) extension to the new FA (Step 5), which in turn creates a BU and sends it to the previous FA on behalf of the MH if the PFAN extension is present (Step 5'). At the same time, as defined in MIPv4, the new

FA relays the RegReq (without the PFAN option) to the MH's HA, specified in the Home Agent field of the RegReq (also Step 5). On receiving the BU, the previous FA deletes the MH's visitor list entry and, if the new CoA of the MH is included in the BU, also creates a binding cache entry for the MH so that it can tunnel in-flight packets to the MH's new CoA (Step 6'). Through this mechanism, smooth handoff is enabled. The previous FA is also expected to return a BA to the MH's new CoA. The new FA detunnels the BA and sends it to the MH right away, or waits for the Registration Reply (RegReply) from the HA and then sends the RegReply with the BA carried in an undefined extension.

Regarding updating the binding cache entry at the CH on the handoff, there are a few means to fulfil this task for route optimisation. Firstly, the MH can place a Binding Warning extension in the RegReq (Step 5) so that the HA can send a BU to the CH(s) specified in this extension when processing the registration (Step 6). Secondly, the previous FA can send a Binding Warning (BW) message to the HA on receiving a packet towards the MH (Step 5"), and then the HA can send a BU to the CH specified in the BW (Step 6). Alternatively, when the destination MH is not in the visitor list or the binding cache the previous FA can send the BW to the CH, which then queries the HA by sending a Binding Request (BR) message and obtains the binding by receiving a BU from the HA.

Briefly, MIPv4-RO introduces a set of extensions to reduce the inefficiency of the triangular routing and packet loss during handoffs, and thus can effectively improve the performances compared with the base MIPv4.

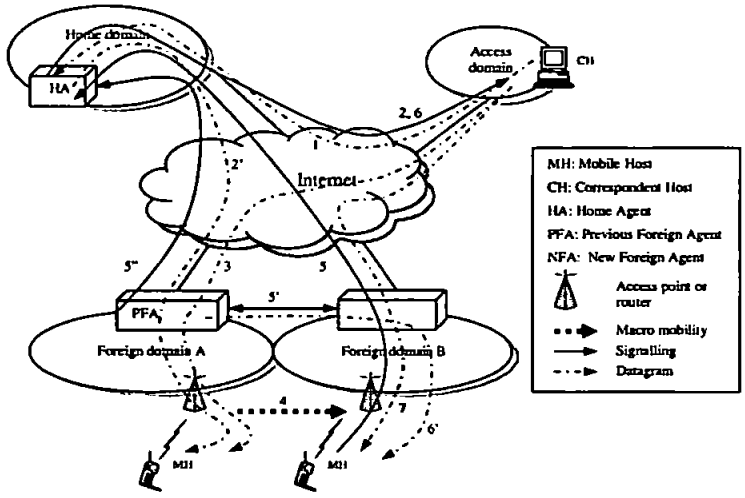


Figure 2.8 MIPv4-RO mobility overview

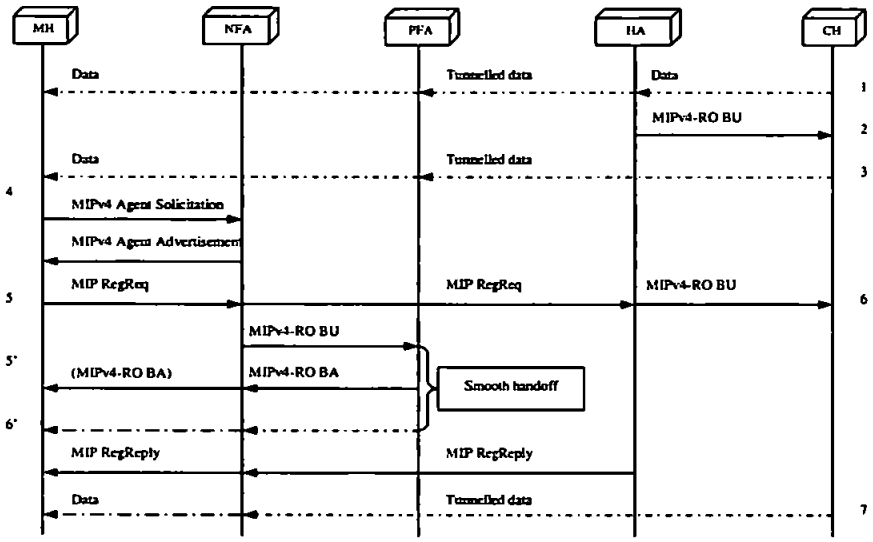


Figure 2.9 MIPv4-RO signalling and data flows

2.6.1.3. Mobile IPv6 (MIPv6)

For mobility in the IPv6 context, Mobile IPv6 (MIPv6) [RFC3775] is designed by utilising the development experiences in MIPv4 (and MIPv4-RO) and some new features offered by the IPv6 standard. Hence, MIPv6 shares much commonness with MIPv4, though their major differences are highlighted in Table 2.2. One of the most remarkable differences is that an end-to-end form of route optimisation is standardised as an integral

component in MIPv6, in contrast to the set of extensions proposed in MIPv4-RO. In fact, the MH is in charge of releasing up-to-date binding information to its CH(s) in MIPv6, whereas the MH's HA plays this role in MIPv4-RO.

Table 2.2 Major differences between MIPv4 and MIPv6

	MIPv4	*MIPv6*
Movement detection	Agent advertisement	IPv6 Neighbour Discovery
FA presence	Optional though usually present	Not needed
CoA distribution	FA or DHCPv4	IPv6 stateless auto-configuration or DHCPv6
CoA type	FA CoA or co-located CoA	co-located CoA
Home registration	Registration Request and Registration Reply, via FA if FA CoA used	Binding Update and Binding Acknowledgement
Route optimisation (RO) status	A set of work-in-progress extensions (MIPv4-RO)	A standardised integral and fundamental part; supported by all IPv6 nodes
RO setup	HA involved	Between MH and CH(s) directly
RO-enabled data delivery	Tunnelling between CH(s) and MH	Source routing with routing header
RO security	Pre-arranged mobility security associations between HA and CH(s)	No pre-arranged mobility security associations needed between MH and CH(s)

In the following, we present the detailed signalling and data flows in MIPv6, as illustrated in Figure 2.10.

As in MIPv4-RO, a CH looks up the destination MH in its binding cache when sending a datagram to the MH. If it does not have a valid binding, the CH sends the datagram to the HoA of the MH (Step 1 in Figure 2.10). The HA intercepts the datagram and tunnels it to the MH's CoA using IPv6 encapsulation if the MH is away from its home network, or forwards the datagram to the MH using standard IP routing if the MH is in the home network (Step 2). On receiving such a tunnelled datagram, the MH can deduce that there is no valid binding at the CH. Thus, the MH can update its binding at the CH if it likes. To do this, the MH initiates a procedure called Correspondent Registration (or CH binding, Step 3), which itself includes two processes referred to as Return Routability (RR) and CH binding. The RR process (Step 7) is discussed a little later. Regarding the CH binding (Step 8), the MH sends a Binding Update (BU) to the CH if the RR process succeeds. The CH then updates the MH binding cache entry and may reply a Binding Acknowledgement (BA) if needed. Afterwards, the CH can send the datagram to the MH's CoA directly with the MH's HoA placed in the IPv6 type-2 routing header (Step 4).

Through these steps, the triangular routing found in the base MIPv4 can be largely reduced. Surely, only Step 4 is needed if a valid binding is available when a CH is ready to send a datagram to an MH. In the uplink direction, an MH sends a datagram to its CH with its CoA placed in the Source Address and its HoA in the Home Address destination option.

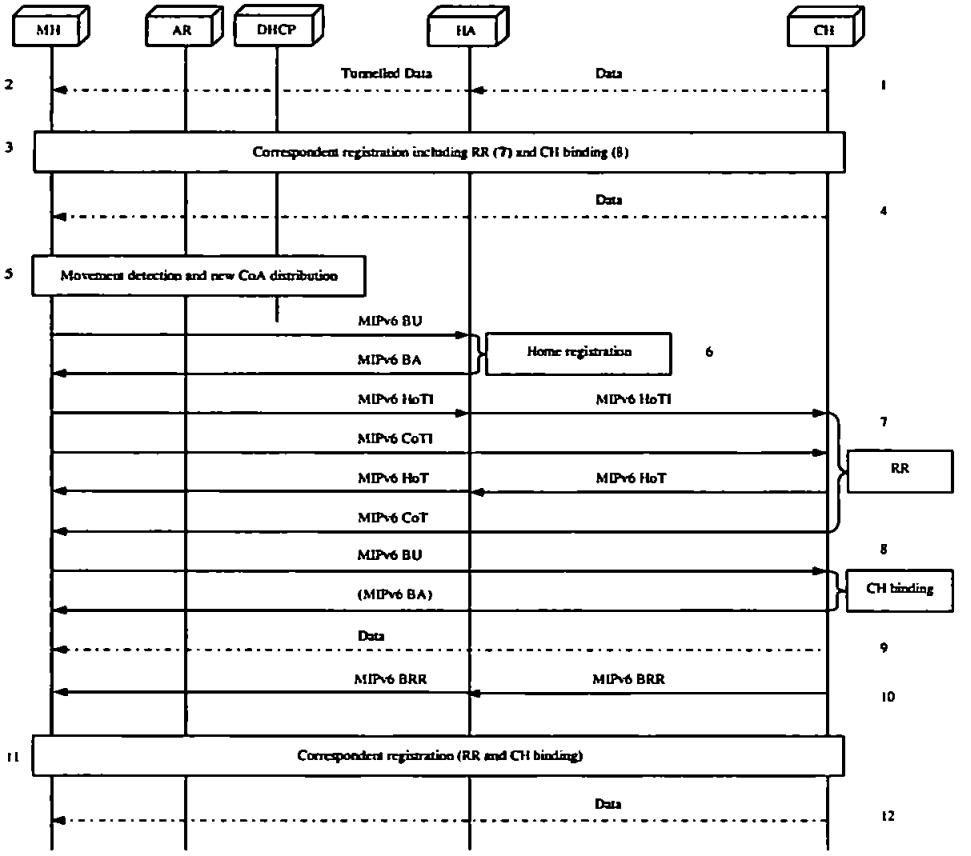


Figure 2.10 MIPv6 signalling and data flows

On a handoff from one access router (AR) to another in a foreign domain, an MH detects the movement with IPv6 Neighbour Discovery as the current default mechanism, and obtains a new IPv6 address (co-located CoA) through IPv6 stateless auto-configuration or DHCPv6 (Step 5). A process called Duplicate Address Detection (DAD) is involved to validate the new address. If Step 5 succeeds, the MH performs home registration by sending a BU to its HA to register the new CoA, and the HA then replies with a BA (Step

6). Having sent the BU to its HA, the MH can initiate the Correspondent Registration procedure (Step 7 and Step 8) for the CH(s). To accelerate the handoff, the MH may not wait for the BA from the MH before it starts the Correspondent Registration(s). After Step 8 completes successfully, the CH can send the subsequent datagram to the MH's new CoA using the routing header as aforementioned in Step 4 (Step 9).

In addition, when the binding cache is going to expire whilst the binding is still in active use, the CH can send a Binding Refresh Request (BRR) to the MH's HoA (Step 10). The BRR is then tunnelled by the HA to the MH, which may then start the Correspondent Registration (Step 11). This procedure completes with a BU sent to the CH if the RR succeeds. This enables the CH to further use the CoA for direct CH to MH transmission.

Finally, the involved RR process (Step 7) is introduced to enable a CH to assure that the MH is in fact addressable at its claimed CoA and HoA so that the subsequent BU from the MH can be authenticated and authorised. This is done by testing whether packets addressed to the two claimed addresses are routed to the MH. The MH can pass the test only if it is able to supply proof that it received the data that the CH sends to those addresses. The MH initiates the RR by sending a Home Test Init (HoTI) message and a Care-of Test Init (CoTI) message simultaneously to the CH. The HoTI is sent via the HA, which will tunnel the Home Test (HoT) message from the CH to the MH's new CoA later. This indicates that the home registration should have succeeded before the HA receives the HoT. When the MH has received both HoT and CoT, the RR is complete and the MH has the data it needs to generate a binding management key for the BU sent to the CH.

To sum up, MIPv6 combines the successful features of MIPv4 and MIPv4-RO whilst reusing standard IPv6 procedures wherever appropriate.

2.6.1.4. The Strength and Weakness of Mobile IP

A major strength of the MIP family protocols is the mobility transparency for TCP traffic by consistently identifying packets delivered to the upper layer using an unchanged HoA. Furthermore, as IP-layer protocols, the MIP family support macro mobility across homogenous as well as heterogeneous networks. MIP also boasts small message sizes from its compact binary codes. On the other hand, MIP may be unsuitable for real-time applications because of the following reasons. Firstly, MIPv4 triangular routing or MIPv6 RR usually leads to high handoff delays. Secondly, MIP imposes extra data delivery overhead to RTP/UDP packets by using MIPv4 tunnelling or MIPv6 type-2 routing header and the destination option. Note that the payloads of RTP/UDP packets are often featured by small size. For example, in VoIP, 20-byte payload is usual when G.729 codec is in use [Oouchi etc. 2002]. Thirdly, due to the network-layer constraints MIP is lack of advanced features specific to application- or user-level mobility requirements.

In addition, MIP support for paging is also proposed (e.g., [Zhang etc 2002, Ramjee etc 2002B]) but this functionality has not been standardised because of the questionable necessity for an IP-level paging [Kempf and Mutaf 2003], among other reasons.

2.6.2. Session Initiation Protocol (SIP)

The Session Initiation Protocol (SIP) [RFC3261] was originally designed for establishing, modifying and terminating IP multimedia sessions, especially RTP/UDP applications like voice over IP (VoIP). Operating in the application layer, SIP can resemble MIP mobility operations, and thus has been proposed to support both location and handoff management (for UDP applications) in addition to the built-in session setup capability [Schulzrinne and Wedlund 2000, Kwon 2002]. Furthermore, SIP can provide advanced mobility features such as session renegotiations for real-time applications on handoffs, and

hence would enable the ongoing sessions to adapt to mobility. In addition, SIP inherently supports user-level mobility with the help of SIP infrastructure and user-level identifiers.

SIP identifies users with URIs (Uniform Resource Identifiers). A SIP URI starts with "SIP:" or "SIPs:" ('s' indicates a secure URI), analogous to "http:" and "https:" for WWW, respectively. The "SIP:" or "SIPs:" is then followed by an email-like identifier string. The basic form of the string is a username appended by the "@" sign and a host name that is either a domain name or a numerical IP address (an IPv6 address needs to be placed in square parenthesis). Examples of SIP URIs are SIP:Alice@plymouth.ac.uk, SIP:Bob@141.163.7.212, and SIPs:Carol@[3ffe::200:86ff:fe76:9616]. Generally, each SIP user is publicly known by a long-term URI called Address-of-Record (AOR), analogous to the HoA in MIP. Moreover, a SIP contact address represents the current location of a SIP user, like the CoA in MIP. In SIP, the bindings of AORs and contact addresses are maintained in databases called location services.

SIP employs a client-server paradigm. The client module running in a host is called User Agent (UA), and SIP servers comprise proxy servers, redirect servers, and registrars. Note that these SIP entities are logical and thus can be implemented separately or collectively in a domain. In the following description, a SIP home server (HS) in an MH's home domain can act as a proxy or redirect server with a home registrar (HR) and a location service co-located. Figure 2.11 and Figure 2.12 display the SIP session setup procedures when the SIP HS serves as a proxy server and a redirect server, respectively.

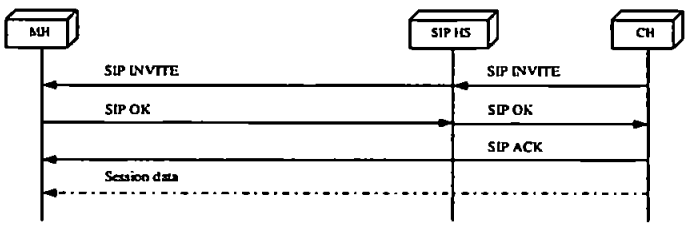


Figure 2.11 SIP mobility: session setup via proxy

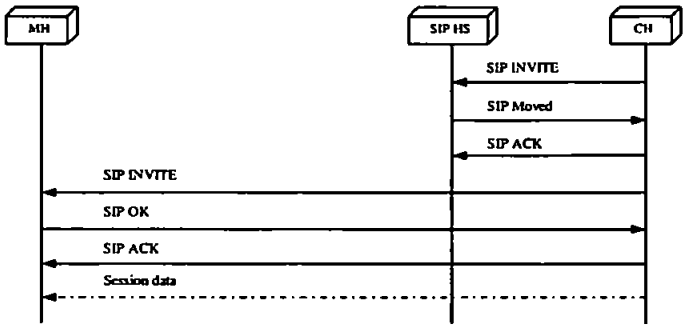


Figure 2.12 SIP mobility: session setup via redirection

In either case, a three-way handshake is applied and the INVITE, the OK and the Ack messages are essential whilst other provisional messages like Trying and Ringing can also be involved. When acting as a proxy server, the SIP HS forwards the INVITE from a CH to the MH's current location after a query at its associated location service, which maps the MH's AOR to its current contact address. When acting as a redirect server, the SIP HS returns the query result to the CH, which then generates a new INVITE and sends it to the MH's current location directly. During a SIP session setup, a CH and an MH negotiates the session parameters, which are described by the Session Description Protocol (SDP) [RFC2327] and are enclosed in the INVITE and OK messages.

Upon a handoff to a new subnet, an MH running SIP mobility obtains a new IP address typically from a DHCP server. This new IP address serves as the new SIP contact address. Then the MH initiates a MIPv6-style CH binding update of its new contact address with its AOR and session renegotiation by sending a re-INVITE message to the CH(s). This procedure succeeds with OK and Ack messages exchanged. The CH then redirects the subsequent session traffic towards the new contact address of the MH directly, and thus the triangular routing found in the base MIP is avoided. The MH also needs to perform location update (home registration) by sending a SIP REGISTER message to its HR, which updates the AOR-contact address binding of the MH at the location service. Like MIP, a SIP user needs to refresh its home registration periodically to keep the binding

valid. Figure 2.13 illustrates the SIP handoff signalling and data sequences with DHCPv4 messages for a new contact address.

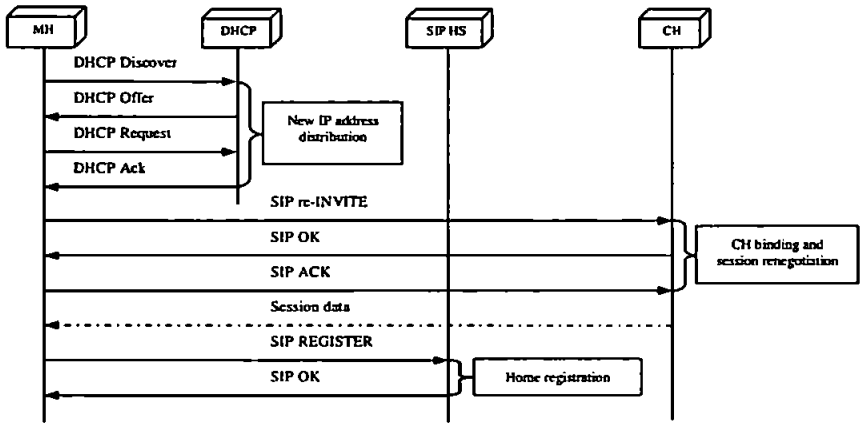


Figure 2.13 SIP mobility: handoff (including location update)

2.6.3. Comparison of MIP and SIP Mobility

As macro-mobility protocols, the MIP and SIP mobility procedures shares much similarity in principle, though these two approaches differ from each other in many details. Table 2.3 lists a comparison of MIPv4, MIPv4-RO, MIPv6, and SIP mobility in the architectural and signalling aspects.

Moreover, SIP mobility can achieve user-level mobility thanks to its application-layer approach. In contrast, it is difficult to extend MIP to fulfil such functions conveniently operating at the application layer. In addition, SIP seems more advantageous in supporting RTP/UDP-based real-time applications thanks to its powerfulness in session management.

However, SIP has its own disadvantages. Firstly, it is difficult and complex to extend SIP for tackling TCP mobility [Vakil etc 2001] as SIP is featured by mobility awareness to applications. Secondly, SIP is much more generous in message sizes since SIP messages are text based, which indicates that the pure SIP mobility approach would generate much higher signalling loads compared with MIP. Another drawback of SIP mobility is that SIP messages may incur additional processing delay in the application layer.

Table 2.3 Comparison of MIPv4, MIPv4-RO, MIPv6 and SIP mobility management

	MIPv4	*MIPv4-RO*	*MIPv6*	*SIP*
Home server	Home Agent (HA)	Home Agent (HA)	Home Agent (HA)	Home Proxy/Redirect Server Home Registrar (HR) DHCP Server
Foreign server	Foreign Agent (FA)	Foreign Agent (FA)	Access Router (AR)	Local Registrar and Proxy/Redirect Server
	DHCPv4 Server	DHCPv4 Server	DHCPv6 Server	DHCP Server
Host	Host part	Host part	Host part	User Agent (UA)
New IP address distribution (host configuration)	FA messages: Agent Solicitation Agent Advertisement	FA messages: Agent Solicitation Agent Advertisement	AR messages: Router Solicitation Router Advertisement	DHCPv4 messages: Discover, Offer, Request, Ack
	DHCPv4 messages	DHCPv4 messages	DHCPv6 messages	DHCPv6 messages: (Normal Mode) Solicit, Advertise, Request, Reply (Rapid Commit Mode) Solicit, Reply
Home registration (location update)	Registration Request Registration Reply	Registration Request Registration Reply	Binding Update Binding Acknowledgement	REGISTER OK
CH binding (route optimisation)	N/A	Binding Warning Binding Request Binding Update Binding Acknowledge	Binding Update Binding Acknowledgement Binding Request Refresh	re-INVITE OK ACK
Session setup	N/A (implicit)	N/A (implicit)	N/A (implicit)	INVITE OK ACK

2.6.4. Hybrid MIP-SIP Mobility Architectures

As discussed in Section 2.6.3, MIP is more efficient in supporting mobility of TCP-based non-real-time applications (TCP mobility) whereas SIP is more effective for RTP/UDP real-time applications (UDP mobility) and pre-session personal mobility (globally reach a user). Consequently, the joint MIP-SIP approach for mobility support appears to be a better solution than pure MIP or pure SIP approaches, and thus has gained increasing significance in recent years. In particular, a couple of hybrid MIP-SIP architectures [Politis etc 2004, Wong etc 2003] with specific designs are coming into being. Both hybrid architectures utilise MIP (or its variant) and SIP for TCP and UDP mobility, respectively; however, MIP and SIP operate in a rather independent way. In this subsection,

we investigate the two representative hybrid MIP-SIP architectures and identify their strength and weakness.

2.6.4.1. Typical Hybrid MIP-SIP Architectures

2.6.4.1.1. EVOLUTE

EVOLUTE [Politis etc 2004] ([Politis etc 2003] is a preliminary version) is the hybrid MIP-SIP architecture that is designed for the EVOLUTE project [EVOLUTE] and thus is named after the project. In the EVOLUTE architecture, IPv4 networking is considered, as shown in Figure 2.14.

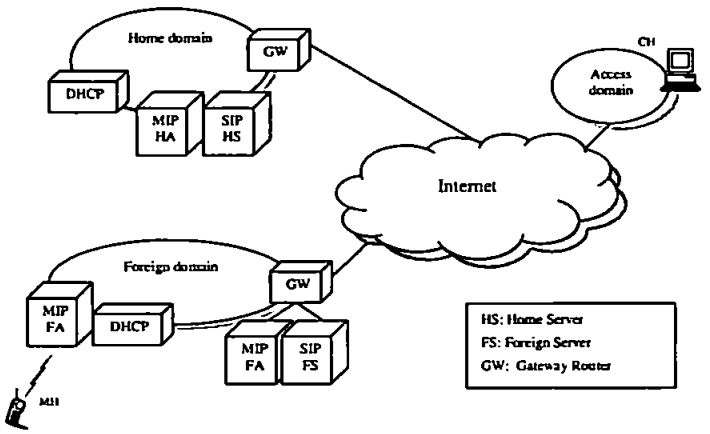


Figure 2.14 Network model in EVOLUTE

In the home domain of an MH, the MIP HA and the SIP HS (home server) coexist to handle MIP and SIP registrations, respectively. The SIP HS is a collection of the home SIP proxy or redirect server (depending on server configurations) and the SIP home registrar with the associated location service. In a foreign domain, a MIP FA and a SIP FS (foreign server) are collocated with the domain gateway router (GW), and collectively they are called the Enhanced Mobility Gateway (EMG).

Upon a macro handoff, the following handoff operations are performed, as shown in Figure 2.15.

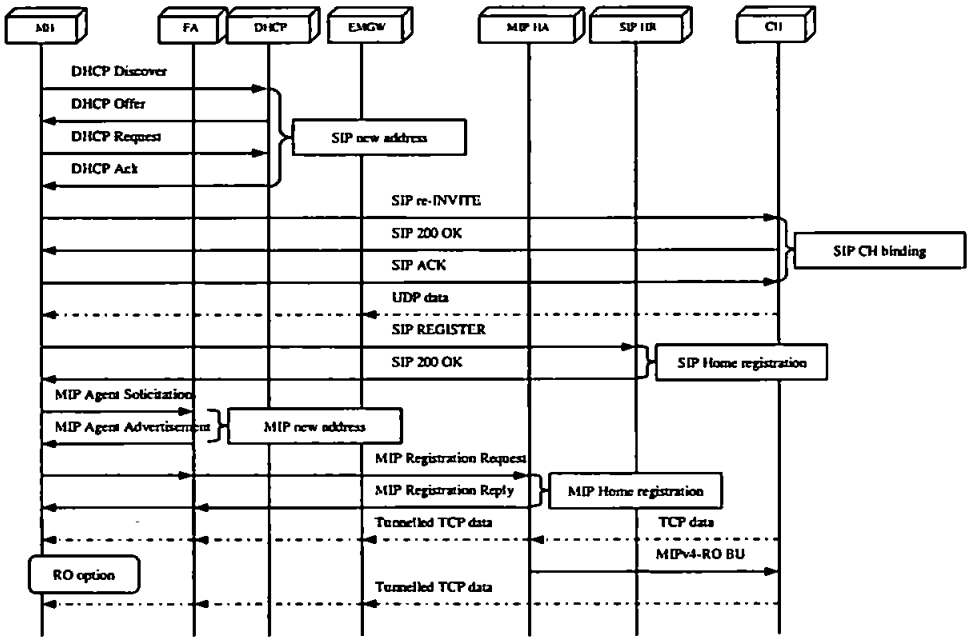


Figure 2.15 Handoff signalling and data sequences in EVOLUTE

If the ongoing session is a UDP application, an MH obtains a new IP address from a DHCPv4 [Droms 1997] server of the new domain and performs SIP binding in the CH and home registration in its SIP HS, respectively. On the other hand, if the ongoing session is a TCP application, the MH instead turns to MIP schemes by obtaining a CoA from the new FA and performing MIP home registration. In the standard basic MIPv4, the subsequent TCP data traffic is then intercepted by the MH's HA, which tunnels these packets to the FA. To handle the triangular routing, MIPv4-RO [Perkins and Johnson 2002] is used after home registration: when receiving the first data packet from the CH, the HA sends a MIPv4-RO Binding Update (BU) message to the CH, which can then tunnel the subsequent packets itself and send them to the FA directly. This process is referred to as RO option. The separation of TCP and UDP traffic takes place at the EMG.

2.6.4.1.2. MMM

Another representative hybrid MIP-SIP framework is called Multilayered Mobility Management (MMM) architecture [Wong etc 2003] ([Dutta etc 2001A] is a preliminary version). The handoff signalling and data sequences are shown in Figure 2.16.

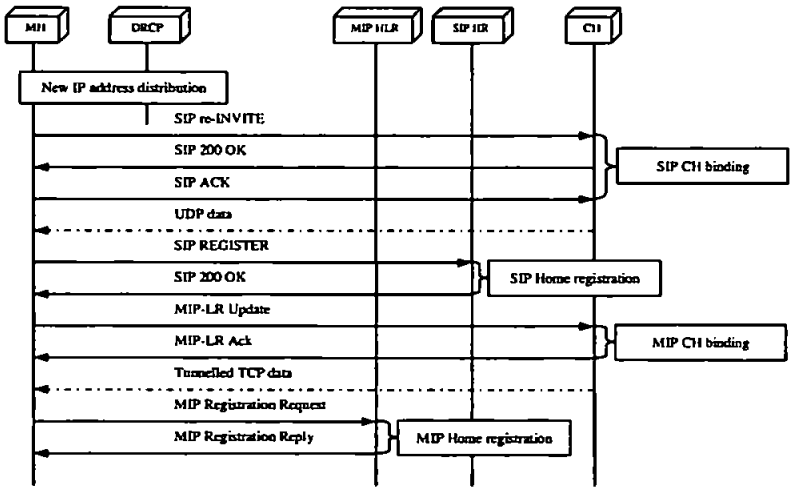


Figure 2.16 Handoff signalling and data sequences in MMM

MMM is similar to EVOLUTE with the following differences. Firstly, a MIPv4 variant called MIP-LR (MIP with Location Registers) [Jain etc 1999] is used instead of the standard MIPv4. MIP-LR resembles SIP (and MIPv6) in the direct CH registration for route optimisation upon handoffs, and the HA is also modified to a HLR (Home Location Register) to support some new features like location enquiry. Correspondingly, each domain has coexisting SIP server and MIP location register. Secondly, an MH itself is in charge of separating TCP and UDP packets as standard tools are available in modern operating systems to achieve this. Thirdly, the architecture adopts a variant of DHCP called DRCP (Dynamic and Rapid Configuration Protocol) [McAuley etc 2000] for uniform host auto-configuration (especially new IP address distribution), and thus the EVOLUTE double address distribution could be avoided. Note that a new IP address can serve as a MIP co-located CoA or a SIP new contact IP address.

2.6.4.1.3. Location Management Procedures

To provide a full picture of the hybrid MIP-SIP approach, Figure 2.17 and Figure 2.18 show the location update and session setup procedures in typical hybrid MIPv4-SIP architectures, especially the EVOLUTE architecture.

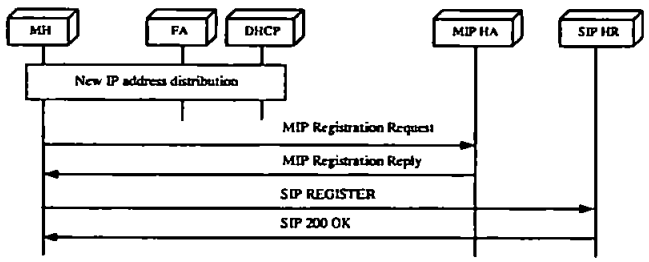


Figure 2.17 Hybrid MIP-SIP architecture: location update (home registration)

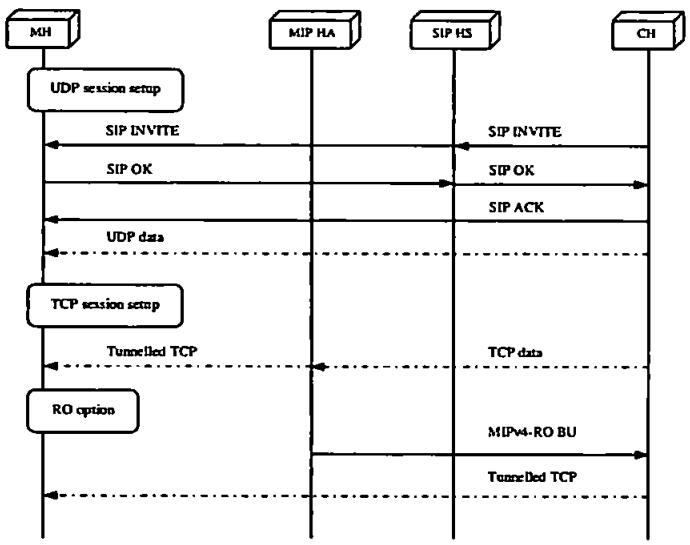


Figure 2.18 Hybrid MIP-SIP architecture: session setup

A uniform IP address distribution process is assumed. For location update, an idle-mode MH needs to perform both home registrations at the MIP HA (or MIP HLR) and the SIP HR using MIP and SIP messages simultaneously as explained in the next subsection. For session setup of a UDP session, SIP uses a three-way handshake between the CH and the MH with the help of the SIP HS (Figure 2.18 shows the proxy case). In the case of a TCP session, no explicit session-setup signalling is introduced and the CH simply sends

packets to the MH's HA and lets the HA tunnel the packets to the MH if the base MIPv4 is used. If the RO option proposed in MIPv4-RO is available, the HA may send a MIPv4-RO BU to the CH so that the CH can tunnel the subsequent packets to the MH's CoA directly. For the MMM architecture, additional non-MIP signalling may be found in the session setup as described in the MIP-LR [Jain etc 1999].

2.6.4.2. The Strength and Weakness of the Hybrid MIP-SIP Architectures

As shown in the above descriptions, in the hybrid architectures MIP and SIP support mobility jointly yet almost independently. In fact, MIP and SIP apply their own signalling protocols for location updates and handoff management, and in the EVOLUTE even for respective new or care-of IP addresses, requiring the presence of both an FA and a DHCP server. Only minimum interactions exist between MIP and SIP entities. The advantage of this approach lies in a relatively prompt deployment since the entities of MIP and SIP are separately adopted and the protocols operate almost independently. Especially in EVOLUTE, the existing MIP and SIP entities are kept almost intact.

However, the advantage would be seriously compromised by the following disadvantages. Above all, the system performances would be greatly deteriorated by significant unnecessary signalling overheads. Notably, both protocols are triggered simultaneously when an MH crosses each domain boundary in idle mode or active/busy mode. When crossing a domain boundary, if in the idle mode an MH needs to updates both of its MIP HA and SIP HR simultaneously using parallel MIP and SIP messages because otherwise the other home server would be unaware of the MH's location change and would result in misconducts due to unavailability of the MH's up-to-date location. For example, if only the SIP HR is updated of the new location, MIP HA would tunnel the packets of a non-real-time application from the CH to an old address, which is no longer valid to the MH, resulting in packet loss and communication failure. Similarly, if the MH only

performs location update at the MIP HA, the SIP HS would proxy or redirect control signalling of a real-time application to the invalid address. Thus, as long as a macro movement occurs, both home registrations should be performed regardless of the traffic type or the MH's mode. If in the active mode, handoffs are triggered. In addition to the above redundant global home registrations, for route optimisation independent MIP (the MIP-LR case in the MMM) and SIP binding updates are performed at the CH when both non-real-time and real-time applications are running between them on handoffs. Furthermore, no matter an MH moves or not, both MIP and SIP requires it to refresh its location binding periodically at the MIP HA and the SIP HR respectively to extend the lifetime of existing home registration. All these above redundant operations consume the valuable wireless bandwidth, the MH's limited battery in addition to posing the superfluous global traffic burden, and thus contribute to the whole system costs. Finally, the repetitive functionality in MIP and SIP entities (especially the MIP HA and the SIP HR) also tends to double the processing and maintenance costs, and thus corresponding optimisations are preferred wherever feasible. In addition, these hybrid architectures focus on IPv4 rather than IPv6. Figure 2.19 abstracts the generic mobility signalling (except session setup) blocks commonly found in typical hybrid MIP-SIP architectures.

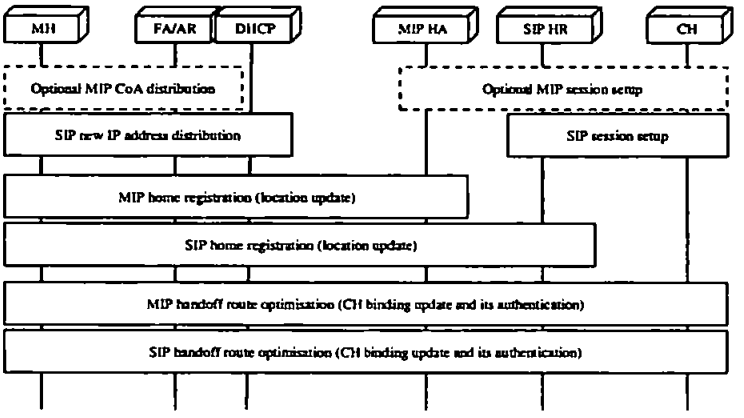


Figure 2.19 Generic mobility signalling block diagram in hybrid MIP-SIP architectures

2.6.5. Alternative Macro-Mobility Protocols

In addition to the network-layer MIP family and the application-layer SIP mobility protocols, there are a number of alternate macro-mobility protocols, which mainly operate in the transport layer. In the following, we briefly review a few representative protocols.

2.6.5.1. Migrate

The Migrate protocol [Snoeren and Balakrishnan 2000] provides an end-to-end TCP mobility solution without triangular routing or third parties (routers or servers). The protocol is TCP-connection-oriented and extends TCP with the proposed "Migrate options", i.e., introducing a connection token into the SYN (Synchronise Sequence Numbers) field of a TCP header. Connections with same tokens are considered as a same one, regardless of changed IP address and port number. After changing its IP address, an MH will restart the previous connection by sending a special SYN packet containing the token. The CH will then resynchronise the connection. Before the CH receives the special SYN packet, it continues to send packets to the old IP address. An initial test showed that achieved performance is similar to that of MIPv4-RO.

2.6.5.2. MSOCKS+

MSOCKS+ [Bhagwat etc 2002] is a proxy-based protocol targeted to achieve TCP mobility in a corporate domain context. It uses split-connection TCP, and the proposed proxy slices the two TCP segments to ensure that this approach does not violate the end-to-end semantics of the TCP protocol. A roaming MH communicates with a CH via the proxy, and thus the triangular routing like that in MIPv4 would happen.

2.6.5.3. Mobile SCTP

The Stream Control Transmission Protocol (SCTP) [RFC2960] is a new reliable transport protocol with enhanced features different from TCP and UDP. In particular, the

multi-homing feature of SCTP and the ADDIP extension [Stewart etc 2005] can allow SCTP to support end-to-end mobility without support of third parties. The multi-homing feature enables an SCTP host to support multiple IP addresses for an SCTP connection; the ADDIP extension enables an SCTP host to add a new IP address or delete an unnecessary IP address, and to change the primary IP address while an SCTP connection is active. Recent research on SCTP mobility, referred to as mobile SCTP, is mainly built upon this extension. Upon a handoff, the MH running mobile SCTP can notify the CH of the IP address change by sending an SCTP Address Configuration Change (ASCONF) packet. The MH can maintain the old IP address during the handoff, and thus a soft handoff is achievable [Koh etc 2004].

2.6.5.4. Host Identity Protocol (HIP)

Like mobile SCTP, another recent interesting research topic on mobility support is the Host Identity Protocol (HIP) [Moskowitz etc 2005], which introduces a new layer between the network layer and the transport layer. In HIP, upper layer sockets are bound to Host Identities (HIs) instead of IP addresses. In addition, the binding of these HIs to IP addresses is performed dynamically. When an MH changes location, it simply sends a HIP readdress (REA) packet and the connection continues uninterrupted. However, adding a new layer to the protocol stack is a huge modification that may cause an updating of numerous Internet applications.

2.6.5.5. The Strength and Weakness of Transport-Layer Mobility Protocols

Compared with network-layer mobility protocols, the transport-layer ones are more straightforward as end-to-end route optimization is naturally built-in. Thus, transmission is more efficient and it may be easier to implement end-to-end QoS and security schemes.

Nevertheless, this approach has the following disadvantages. Firstly, this approach is transport-protocol specific and thus other transport protocols may also have to be modified

to provide mobility support for applications. For example, mobile SCTP does not support TCP mobility, which the majority of the Internet applications need. Secondly, in the TCP-based protocols such as Migrate and MSocks+, modifying the well-established TCP protocol will cause significant back-compatibility problems. Thirdly, this approach appears useful only for handoff management and thus location management may rely on other mobility protocols like MIP [Goff and Phatak 2004]. Finally, without third parities, the end-to-end mobility approach can hardly support simultaneous movements of the MH and the CH.

2.7 Micro-Mobility Protocols

Micro-mobility protocols can be broadly classified into two categories: tunnelling-based or host-specific [Campbell and Castellanos 2000, Campbell etc 2002]. We focus on the former category and present a comparison of both approaches. Moreover, fast handoff protocols are described as enhancements to micro-mobility protocols. We further investigate the related work on enhancing HMIPv6 [RFC4410] and/or FMIPv6 [RFC4068], and integrating the two schemes. In addition, QoS extensions to micro-mobility and two-phased mobility management are discussed.

2.7.1. Tunnelling-Based Approach

In this subsection, we review two representative tunnelling-based micro-mobility protocols: MIPv4 Regional Registration (MIPv4-RR) and Hierarchical MIPv6 (HMIPv6), for IPv4 and IPv6 networking environments, respectively. Additional tunnelling-based micro-mobility protocols such as the TeleMIP (Telecommunications-Enhanced MIP) [Das etc 2000], the IDMP (Intra-Domain Mobility Management Protocol) [Das etc 2002] and

the BCMP (BRAIN Candidate Mobility Management Protocol) [Keszei etc 2001] are akin to MIPv4-RR in principle and thus are not discussed for brevity.

2.7.1.1. MIPv4 Regional Registrations (MIPv4-RR)

The MIPv4-RR protocol [Gustafsson etc 2004] employs a hierarchy of FAs to handle MIPv4 registrations locally. Typically, a two-level hierarchy is considered in a foreign domain where all regional FAs are connected to a Gateway FA (GFA), though a multiple-level hierarchy is also possible. The two domains *A* and *B* in Figure 2.20 demonstrate these two layouts, respectively. In the former case, direct tunnels connect the GFA to FAs that are located at access routers, whilst an intermediate hierarchy of FAs are deployed in the latter case (three-level hierarchy). In the following, we assume a two-level hierarchy and describe the protocol details as shown in Figure 2.21. Note that more than one GFA may coexist in a domain. An MH's changing GFA is a macro-mobility event, and thus a home registration must be triggered.

When first arriving at a foreign domain (e.g., domain *A* as shown in Figure 2.20), an MH obtains two CoAs: one is the address of the GFA (GFA CoA), and the other is a "local" CoA, which is typically an FA CoA. These CoAs are included in the extended Agent Advertisements from the serving FA (FA1). After this step (Step 1), the MH sends a MIPv4 Registration Request (RegReq) to FA1, which in turn relays the message with its own address included in the Hierarchical Foreign Agent (HFA) extension to the GFA specified in the care-of address field of the message. After creating an entry (the binding of the MH's "local" CoA and its HoA) for the MH in its visitor list, the GFA relays the message (without the HFA extension) to the MH's HA for home registration, registering the binding of the MH's HoA and its GFA CoA. The HA then replies with a MIPv4 Registration Reply (RegReply), which finally reaches the MH via the GFA and FA1 (Step 2). These registration messages establish a tunnel between GFA and FA1 along the path

between the GFA and the MH. Packets addressed to the MH from a CH (not shown here) travel in the tunnels, which can be viewed as a separate routing network overlay on top of IP (Step 3). Subsequently, as long as the MH roams between FAs in the same domain, only regional registrations towards the GFA are needed and no home registration is triggered. For instance, on detecting a movement from FA1 to FA2 in Foreign domain A (Step 4), the MH simply performs regional registrations using MIPv4-RR registration messages (Step 5). As a result, a new tunnel between the GFA and FA2 is then established and data traffic is redirected to the MH's new location (Step 6). In addition, smooth handoff between the previous and the new FAs as specified in MIPv4-RO (with minor modifications) is optional in MIPv4-RR. Note that an MH with a co-located CoA can also use this protocol, typically by exchanging MIPv4-RR registration messages between the GFA and itself directly. Paging extensions are proposed in [Haverinen and Malinen 2000]. The location of an MH, in terms of a paging area, is known by its HA. On receiving a packet addressed to an MH located in a foreign domain, the HA tunnels the packet to the paging FA, which then pages the MH to re-establish a path toward the current point of attachment.

To sum up, MIPv4-RR is a natural extension to MIPv4 by introducing a GFA as a regional HA so that home registrations are largely reduced.

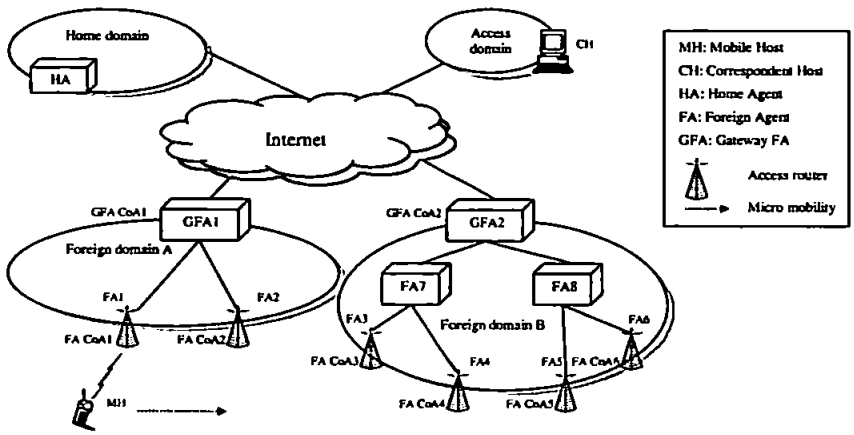


Figure 2.20 MIPv4-RR network model

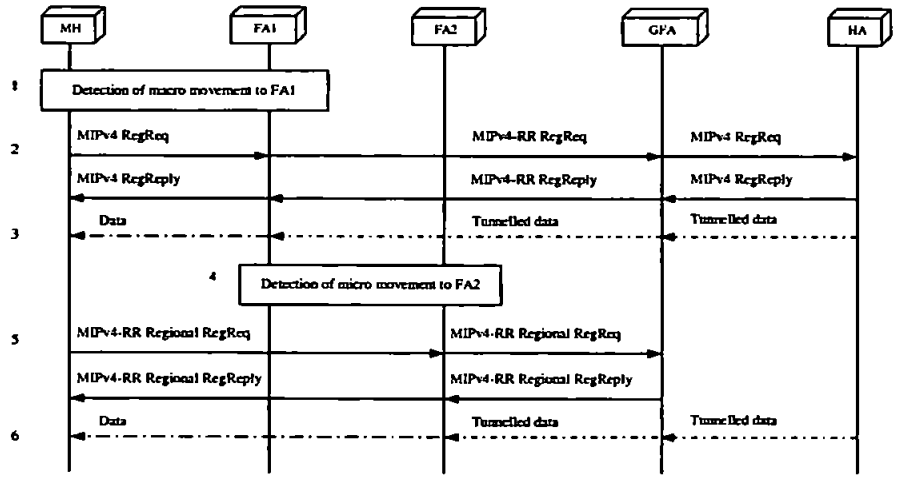


Figure 2.21 MIPv4-RR signalling and data flows

2.7.1.2. Hierarchical Mobile IPv6 (HMIPv6)

Like MIPv4-RR, HMIPv6 [RFC4140] is another tunnelling-based micro-mobility protocol being standardised in the IETF, though it is targeted at the IPv6 mobility. Analogous to the GFA and FAs in MIPv4-RR, HMIPv6 introduces a new MIPv6 entity, called Mobility Anchor Point (MAP). A MAP can be located at any level in a hierarchical network of IPv6 routers including an access router (AR). Multi-level MAPs can be deployed but are used independently. Figure 2.22 illustrates the protocol operation, explained as follows. An MH entering an HMIPv6 domain will receive Router Advertisements containing information on one or more MAPs from the serving AR. The MH can select the first-hop MAP or one further up in the hierarchy. Then, it creates a couple of CoAs, an on-link CoA (LCoA) and a Regional CoA (RCoA), by appending its interface address to the prefix of the AR and the MAP, respectively (Step 1). Upon successfully validating the LCoA (as validating a CoA in MIPv6) and forming the RCoA, the MH sends a local BU to the MAP to bind its LCoA with the RCoA, which is subject to the validation by the MAP first. If the RCoA is valid, a binding for the MH is created. Subsequently, the MAP returns a BA to the MH, indicating a successful binding with a

type-2 routing header that contains the MH's RCoA (Step 2). The involved address validations are through the standard Duplicate Address Detection (DAD) process. Following a successful registration with the MAP, a bi-directional tunnel between the MH and the MAP is established. Acting as a local HA, the MAP will receive all packets on behalf of the MH it is serving and will encapsulate and forward them directly to the current address of the MH. After registering with the MAP, the MH must register its RCoA with its HA by sending a BU that specifies the binding (RCoA, HoA) as in MIPv6 (Step 2). The HoA is used in the home address option and the RCoA is used as the CoA in the source address field. The MH should wait for the BA from the MAP before registering with its HA. The MH may also send a similar BU (i.e. that specifies the binding between the HoA and the RCoA) to its current CH(s) after the MIPv6 RR process if the route optimisation is preferred (Step 3) so that the CH can send datagram to the MH's RCoA directly. The MAP intercepts and tunnels the incoming datagram to the MH directly. AR1 simply relays the tunnelled datagram to the MH (Step 4).

If the MH changes its current address (LCoA) within a local MAP domain (Step 5), it only needs to register the new LCoA with the MAP (Step 6) so that the incoming datagram tunnelled by the MAP can be redirected towards the new LCoA (Step 7). The RCoA does not change as long as the MH moves within a MAP domain. This makes the mobility of the MH transparent to the CH(s) with which it is communicating. In addition, the MH may send a BU containing its LCoA (instead of its RCoA) to its CH(s), connected to its same link. Packets will then be routed directly without going through the MAP.

Upon a handoff to a new MAP that belongs to the same domain where the previous MAP is located, smooth handoff is recommended in HMIPv6 to speed up the handoff and reduce packet loss. For this purpose, the MH sends a BU to its previous MAP specifying its new LCoA. Packets in transit that reach the previous MAP are then forwarded to the

new LCoA. Nevertheless, the MH has to perform a home registration to register its new RCoA. In another words, an inter-MAP movement, like an inter-GFA movement in MIPv4-RR, triggers macro-mobility operations.

Briefly, despite some differences HMIPv6 appears to be an IPv6 version of MIPv4-RR in that both introduce local registrations at virtual home agent(s) in a foreign domain, and establish tunnels between an MH and the virtual home agent(s) for packet delivery. In addition, IP tunnelling is on top of IP routing and thus the tunnelling-based protocols are sometimes referred to be "L3.5" protocols.

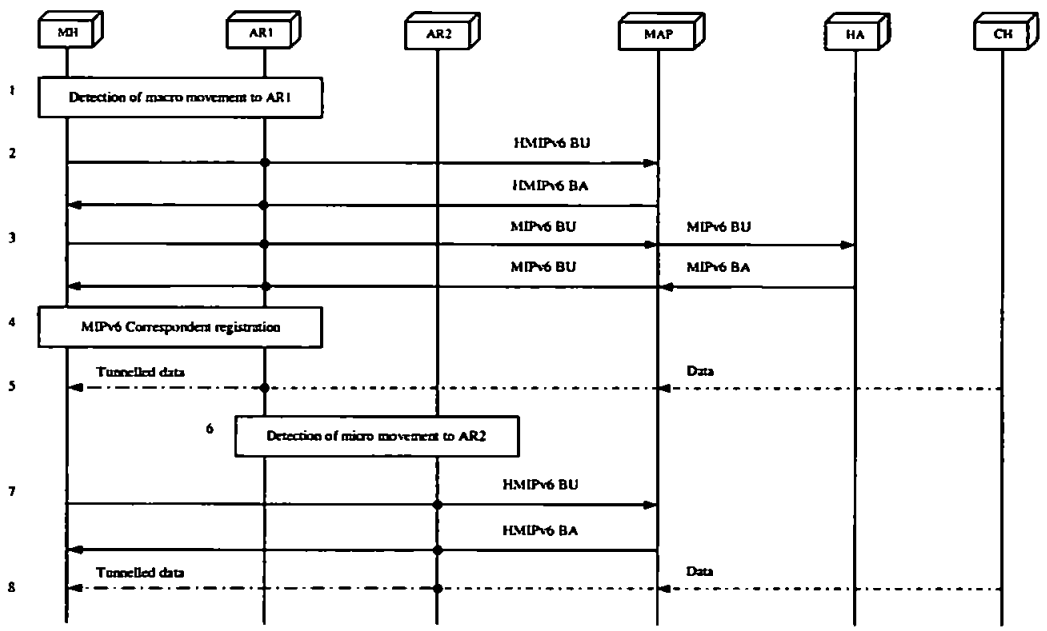


Figure 2.22 HMIPv6 signalling and data flows

2.7.2. Host-Specific Approach

In this subsection, we outline a couple of typical host-specific micro-mobility protocols, Cellular IP (CIP) and HAWAII. Both protocols were originally designed for IPv4 micro mobility, though much of the design principles may be applicable to the IPv6 context.

2.7.2.1. Cellular IP (CIP)

Location and handoff management are integrated with routing in Cellular IP access networks [Campbell etc 2000]. To minimise control messaging, regular data packets transmitted by MHs are used to update and refresh host location information at CIP-aware entities (called CIP nodes) so that the path routes are maintained. Hence, CIP nodes monitor MH-originated packets and maintain a distributed, hop-by-hop location database that is used to route packets to MHs. CIP uses the HoA to identify an MH. CIP supports both hard handoffs and semi-soft handoffs. During a semi-soft handoff, the crossover CIP node bi-casts incoming packets to both previous and current access points. IP paging is also supported in CIP. When packets need to be sent to an idle mobile host, the host is paged using a limited scope broadcast and in-band signalling.

Furthermore, an IPv6 version of CIP (CIPv6) [Shelby etc 2001] has been proposed to update CIP with IPv6 capability. For instance, an MH obtains its CoA through IPv6 stateless auto-configuration and it is identified by its CoA instead of HoA as in the original CIP (CIPv4).

2.7.2.2. HAWAII (Handoff-Aware Wireless Access Internet Infrastructure)

Another representative host-specific micro-mobility protocol is HAWAII [Ramjee etc 2002 A]. On entering a new FA domain, an MH receives a co-located CoA and retains it unchanged while moving within this domain. Nodes in a Hawaii network execute a generic IP routing protocol and maintain mobility-specific routing information as per host routes added to legacy routing tables. In this sense, HAWAII nodes can be considered enhanced IP routers, where the existing packet forwarding function is reused. Location information (i.e., mobile-specific routing entries) is created, updated, and modified by explicit signalling messages sent by mobile hosts. HAWAII defines four alternative path setup schemes that control handoffs between access points. The appropriate path setup scheme is

selected depending on the operator's priorities between eliminating packet loss, minimising handoff latency, and maintaining packet ordering. HAWAII also uses IP multicasting to page idle mobile hosts when incoming data packets arrive at an access network and no recent routing information is available [Ramjee etc 2002 B].

2.7.2.3. Comparison of Micro-Mobility Protocols

The Table 2.4, based on [Campbell and Castellanos 2000, Campbell etc 2002, Akyildiz etc 2004], compares the operation details and deployment considerations of Cellular IPv4, HAWAII, MIPv4-RR and HMIPv6. According to [Campbell and Castellanos 2000, Campbell etc 2002], the performance differences among CIPv4, HAWAII and MIPv4-RR are not significant. Therefore, deployment considerations are more important when implementation choices are available.

Table 2.4 Comparison of micro-mobility protocols

	Host-specific protocols		Tunnelling-based protocols	
	CIPv4	HAWAII	MIPv4-RR	HMIPv6
Layer	L3	L3	"L3.5"	"L3.5"
Mobility-aware entities	All CIP nodes: intermediate and access	All routers: intermediate and access routers (ARs)	FA(s) and GFA(s)	MAP(s)
Top-level mobility-aware entity in use	Gateway	Domain root router	GFA	The selected MAP
Mobile host ID	HoA	CoA	HoA	LCoA
Intermediate nodes	L2 switches	L2 switches	L3 routers	L3 routers
Means of location update	Data packets	Signalling messages	Signalling messages	Signalling messages
Location update direction	Towards the Gateway	Towards the previous router	Towards the GFA	Towards the MAP
Paging	Implicit	Explicit	Explicit (an unofficial extension)	Not officially defined
Tunnelling	No	No	Yes	Yes
Smooth handoff or variants	Semi-soft handoff by bi-casting	Yes, between ARs	Optional, between FAs	Yes, between MAPs
Fast handoff	Optional	Optional	Optional	Optional
MIP messaging	No	Yes	Yes	Yes
Additional cost	Propagating host-specific routing information in mobility-aware nodes		Tunnelling overhead in mobility-aware nodes of each hierarchy	
Reliability	Rely on root (gateway) router		Rely on mobility agents (FAs or MAPs) at each hierarchy	
Gradual deployment	Difficult		Easy	

In summary, the host-specific protocols are advantageous in avoiding the tunnelling overhead incurred in the tunnelling-based ones (yet at the cost of propagating host-specific routing information). However, the host-specific approach replaces the standard IP routing with host-specific routing and the intermediate nodes can only be L2 entities. This

indicates that each router in this approach has to be mobility aware. Such a requirement is really a huge deployment problem. In contrast, the intermediate nodes in the tunnelling-based protocols are just standard routers. This is a great deployment advantage as only selected entities are mobility aware.

2.7.3. Fast Handoff Protocols

Most of the above macro- and micro-mobility protocols, such as the MIP family, were originally designed without any assumptions about the underlying link layers over which they would operate so that they could have the widest possible applicability. This approach has the advantage of facilitating a clean separation between L2 and L3 of the protocol stack, but it has negative consequences for handoff delay (or latency). Therefore, fast handoff protocols have been proposed to utilise L2 triggers to accelerate L3 handoffs. L2 triggers refer to the information from L2 that informs L3 of particular events, such as the forthcoming start of an L2 handoff and the notification on the completion of an L2 handoff.

Fast handoff protocols are building blocks to mobility management and may interact with either macro- or micro-mobility protocols (or even both), depending on specific designs. In this subsection, we survey two typical fast handoff protocols for IPv4 and IPv6, respectively. Although they are extensions to the MIP family protocols, much of the design principle may be applicable to non-MIP protocols.

2.7.3.1. Low Latency Handoffs in MIPv4 (LL-MIPv4)

In [Malki 2004], three methods are proposed to achieve low latency MIPv4 handoffs (LL-MIPv4): pre-registration handoff method, post-registration handoff method, and combined handoff method. The pre-registration handoff method enables anticipated IP-layer handoffs, where an MH is assisted by the network in performing an L3 handoff before it completes the L2 handoff. The L3 handoff can be initiated by the MH or the

network. Accordingly, L2 triggers are used both in the MH and in the FA to trigger particular L3 handoff events. The pre-registration method coupled to L2 mobility helps to achieve seamless handoffs between FAs. No new MIPv4 messages are proposed, except for an extension to the Agent Solicitation message in the mobile-initiated case as shown in Figure 2.23. In this case, the MH receives an L2 trigger containing the IP address (or equivalent information) of the new FA (NFA) before an imminent L2 handoff from the current FA (though referred to as the previous FA or PFA after the L2 handoff) to the NFA. Then, the MH sends an LL-MIPv4 Proxy Router Solicitation (PrRtSol) message (with the IP address of the NFA enclosed) to the PFA, which replies with an LL-MIPv4 Proxy Router Advertisement (PrRtAdv) message. This PrRtAdv is a cached copy of the one actually sent by the NFA beforehand. On receiving the PrRtAdv, the MH has enough information for home or regional registrations, via the PFA and the NFA, depending on whether the MIPv4-RR protocol is in use.

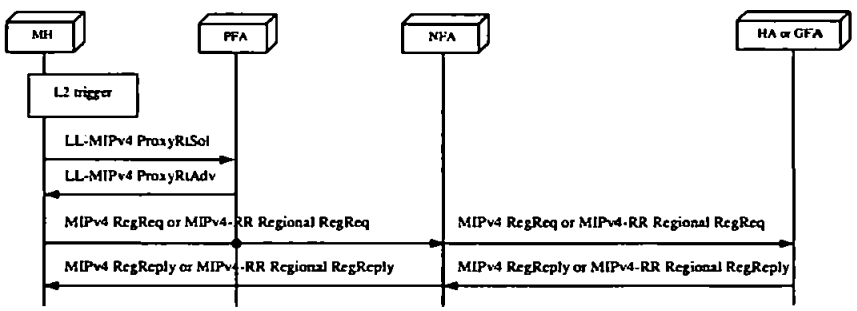


Figure 2.23 Pre-registration mode in LL-MIPv4

The post-registration handoff method proposes extensions to MIPv4 to allow the PFA and NFA to utilize L2 triggers to set up a bi-directional tunnel between PFA and NFA that allows the MH to continue using its PFA while on NFA's subnet. This enables a rapid service re-establishment at NFA. The MH eventually performs a MIPv4 registration after L2 communication with the new FA is established, but this can be delayed as required by the MH or FA. Until the MH performs registration, the FAs will setup and move bi-

directional tunnels as required to give the MH continued connectivity. Figure 2.24 illustrates the case where an L2 trigger at the PFA are utilised. On the L2 trigger, the PFA sends an LL-MIPv4 Handoff Request (HRqst) message to the NFA, which acknowledges with an LL-MIPv4 Handoff Reply (HRply) message. After the L2 handoff, the MH performs the home or regional registration via the NFA directly.

The combined method involves running a pre-registration and a post-registration handoff in parallel. If the pre-registration handoff can be performed before the L2 handoff completes, the combined method resolves to a pre-registration handoff. However, if the pre-registration handoff does not complete within an access technology dependent time, the PFA starts forwarding traffic for the MH to the NFA as specified in the post-registration handoff method. This provides for a useful backup mechanism when completion of a pre-registration handoff cannot always be guaranteed before the L2 handoff completion.

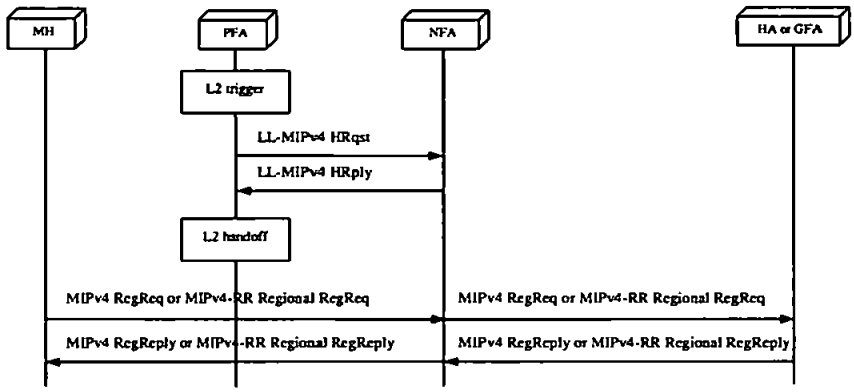


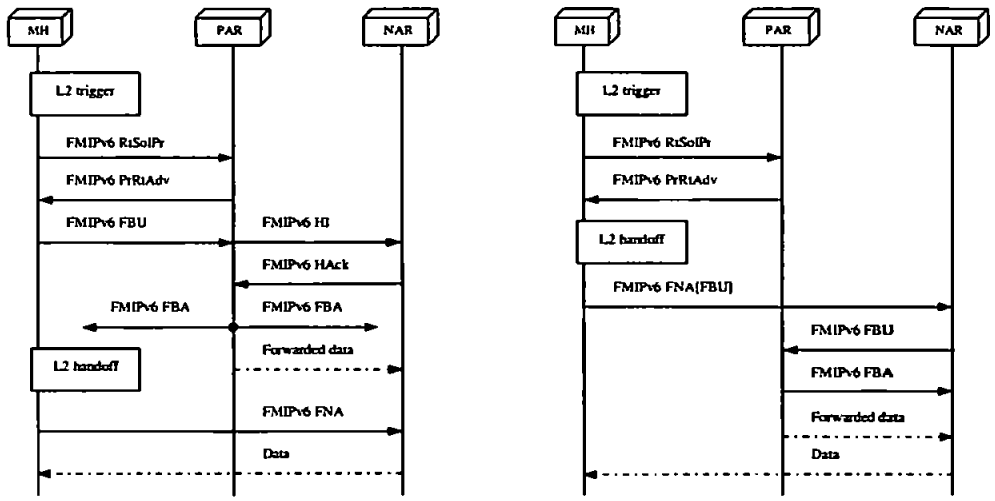
Figure 2.24 Post-registration mode in LL-MIPv4

2.7.3.2. Fast Handoffs for MIPv6 (FMIPv6)

Similar to LL-MIPv4, the Fast Handoffs for MIPv6 (FMIPv6) [RFC4068] are proposed to minimise handoff delays in the IPv6 context.

The protocol begins when an MH sends an FMIPv6 RtSolPr to its previous access router (PAR) to resolve one or more Access Point Identifiers to subnet-specific information.

In response, the PAR sends an FMIPv6 PrRtAdv message containing one or more [AP-ID, AR-Info] tuples. The MH may send an RtSolPr as a response to an L2 trigger or simply after performing router discovery. However, the expectation is that prior to sending an RtSolPr, the MH has discovered the available APs by link-specific methods. With the information provided in the PrRtAdv, the MH formulates a prospective new CoA and sends an FMIPv6 Fast Binding Update (FBU) message. The purpose of FBU is to authorise PAR to bind the previous CoA to the new CoA, so that arriving packets can be tunnelled to the new location of the MH. The FBU should be sent from PAR's link whenever feasible. For instance, an L2 trigger could enable the FBU transmission from the previous link. When it is not feasible, FBU is sent from the new link. Care must be taken to ensure that the new CoA used in an FBU does not conflict with an address already in use by some other node on link. For this, the FBU is encapsulated within an FMIPv6 Fast Neighbour Advertisement (FNA) message and is used when FBU is sent from the link of the new AR (NAR). Depending on whether an FMIPv6 Fast Binding Acknowledgement (FBA) is received or not on the previous link, which depends on whether an FBU was sent in the first place, there are two modes of operation. The scenario in which an MH sends an FBU and receives an FBA on PAR's link is referred to as predictive mode. The scenario in which the MH sends an FBU from NAR's link is called reactive mode. Note that the reactive mode also includes the case when an FBU has been sent from PAR's link but an FBA has not been received yet. When the FBU is sent from PAR's link, the PAR sends an FMIPv6 Handover Initiation (HI) to the NAR, and the NAR checks the validity of the proposed new CoA through DAD and returns the results to the PAR via an FMIPv6 Handover Acknowledge (HAck), which then sends an FBA to the MH. The signalling and data flows of these two modes are illustrated in Figure 2.25 (a) and (b), respectively.



(a) Predicative mode (b) Reactive mode

Figure 2.25 FMIPv6 signalling and data flows

2.7.4. Weakness of HMIPv6 and FMIPv6

Compared with MIPv6, HMIPv6 and FMIPv6 improve system performance in terms of reduced global signalling traffic and expedited handoff management, respectively. However, they still suffer from some shortcomings.

Firstly, on each time changing an AR (inter-AR movement), an MH has to make sure that a unique HMIPv6 LCoA (on-link CoA) or FMIPv6 CoA is obtained through stateless address auto-configuration. A tentative LCoA/CoA (TLCoA or TCoA) is constructed by appending the MH's interface identifier to the subnet prefix advertised by the new AR. Though the probability that this address is invalid (duplicate) is very low, the MH needs to perform the IPv6 Duplicate Address Detection (DAD) to check its uniqueness. In each DAD, the MH sends at least one multicast NS (Neighbour Solicitation) message containing the tentative LCoA as the target and listens to responses from other nodes for a pre-defined time (at least 1000 ms). If no reply is received after this period, the tentative LCoA becomes a valid LCoA. Otherwise, the MH needs to generate another tentative LCoA manually or using an alternate interface identifier. Obviously, the DAD is the dominating

time consumer in the standard auto-configuration procedure, and thus significantly increases the handoff delay. The oDAD (optimistic DAD) scheme [Moore 2005] advocates to skip the DAD process at the risk of address collisions. Thus, it is not an ideal solution.

Secondly, each inter-AR movement involves the HA or MAP for location update, QoS route reservation, which generate scalability concerns. HMIPv6 addresses this problem by introducing multiple MAPs in a domain and allowing different MHs select different MAPs, however, any inter-MAP movement is a macro event that incurs a home registration. The increase of such macro messaging actually contradicts the micro-mobility design goals.

Additionally, when FMIPv6 operates over HMIPv6, the packets forwarded from the previous AR and those from the new AR directly arrive at the MH in an interleaved way, resulting in out-of-order packets and hence QoS degradation. Let alone that no IP QoS support has been defined in the specifications.

2.7.5. Integration of HMIPv6 and FMIPv6

Despite these shortcomings, FMIPv6 and HMIPv6 could cooperate to support improved IPv6 micro mobility compared with the cases when either of them is applied alone. Therefore, research towards an optimal integration of both protocols has gained growing importance.

Ref. [Hsieh etc 2002] studied the superimposition case of FMIPv6 over HMIPv6, where the HMIPv6 operations follow FMIPv6 signalling directly on every micro handoff. In [Hsieh etc 2003], the direct FMIPv6 over HMIPv6 operations are enhanced with a movement tracking technique for seamless low-mobility handoffs in indoor large open space environment. For this purpose, six new kinds of additional messages are defined in the so-called S-MIP (Seamless-MIP) architecture to the existing HMIPv6 and FMIPv6 ones, and thus considerable signalling costs are incurred. Another architecture called F-

HMIPv6 (Fast-HMIPv6) [Jung and Koh 2004] was proposed to optimise the FMIPv6 over HMIPv6 operations in the networking scenario where the MAP is the crossover router of the previous and the new ARs. Under that circumstance, the FMIPv6 HI (Handover Initiation) and HAck (Handover Acknowledge) messages are signalled between the MAP and the new AR other than between the previous and the new ARs as defined in FMIPv6 [RFC4068]. A similar approach is also discussed in the HMIPv6 specification [RFC4410]. However, this optimisation hinders the smooth context transfer from the previous AR to the new AR via the HI message and new context transfer messages and protocols have to be defined [Dimopoulou etc 2004]. Moreover, it is not efficient when the MAP is located far from the ARs. In these studies, the DAD effects are either disregarded or omitted by skipping the process and the QoS signalling and out-of-sequence packets problem are barely addressed.

2.7.6. QoS Support with Micro-Mobility Extensions

Regarding IP QoS management, existing architectures are mainly based on the Differentiated Services (DiffServ) model [RFC2475] and/or the Integrated Services (IntServ) model [RFC1633] coupled with the Resource ReserVation Protocol (RSVP) [RFC2210] for explicit QoS signalling. Due to their complementary characteristics in QoS control granularity and scalability, IntServ/RSVP and DiffServ are suitable for access network and core network respectively; and thus IntServ/RSVP operating over DiffServ [RFC2998] (or Aggregated RSVP [RFC3175]) could provide an end-to-end QoS control with proper mobility extensions for mobile systems.

A crucial issue in QoS signalling in mobile networks is the interfacing between the QoS signalling and mobility protocols. An independent operation of QoS and mobility signalling could lead to ambiguities and even interoperability problems. Therefore,

interactions between the two protocols (e.g., how the mobility protocol triggers the transfer of signalling messages) should be clearly defined. In recent years, a number of QoS paradigms have been proposed for QoS management in mobile environments, especially by extending standard IP QoS signalling protocols to cooperate with MIP [Moon and Aghvami 2001 and 2003, Taha etc 2005].

As far as RSVP with IP micro-mobility extension is concerned, two major recent approaches are RSVP with pointer forwarding (RSVP-PF) [Lee etc 2001] and RSVP with crossover router re-routing (RSVP-CR) [Moon and Aghvami 2004]. In the RSVP-PF scheme, the QoS route is simply extended from the old access router to the new one upon each handoff. This approach leads to smooth handoff at low cost in every single operation. However, the cumulative consequence after a series of operations results in a triangular routing with additional network resources consumed and additional application end-to-end delay incurred. On the contrary, the RSVP-CR scheme tends to seek an optimised route by means of rebuilding the QoS route from a crossover router to the new access router. This approach often results in longer service disruption time due to searching the appropriate crossover router and rerouting, and requires that the chosen crossover routers be mobility aware and thus imposes significant deployment costs. The involved signalling costs are also significantly larger than those in the first approach are. Therefore, a trade-off between these two approaches is desirable. We target to achieve this objective through a two-phased dynamic optimisation design. In addition, [Paskalis etc 2003] defines the operations regarding address translation when RSVP is interworking with HMIPv6.

2.7.7. Two-Phased Mobility Management

The two-phased mobility management has been proposed in the context of ATM (Asynchronous Transfer Mode) networks [Wong etc 2000 and 2001]. The emerging work

has shown an increasing interest to apply such a methodology into the IP world though focused on MIPv4 so far. The philosophy of a two-phased scheme is to conduct a consecutive inexpensive yet effective operations in the first phase at the price of additional data delivery cost, and to initiate a more expensive operation in the second phase to eliminate the cumulative first-phase negative impacts with larger signalling costs incurred. Note that such a philosophy may be applicable to both macro and micro mobility scenarios.

An essential design issue is to determine when to trigger the second-phase operation. The trigger threshold can be predefined, e.g. [Lo etc 2004] use movement-based thresholds such as the number of handoffs or a change of paging/cluster area. This approach is easy to implement but high cost-effectiveness can hardly be guaranteed. Thus, a better approach should consider the total system costs trade-off. Cost-driven route optimisation algorithms are investigated in [Wong etc 2001] and [Lee and Akyildiz 2003] for macro ATM and MIPv4 handoffs, respectively, without considering a micro-mobility scheme. Thus, their algorithms involve global-area variables that can hardly be available or even easily estimated (unless with over-simplified assumptions) due to the targeted macro mobility. Moreover, QoS signalling is not taken into account either, let alone the trade-off between different QoS signalling approaches. Last but not the least, owing to the distinguishing MIPv6 mobility features, e.g., addresses setting and management, IPv6-based micro mobility architecture demands a more careful design.

2.8 The Cross-Layer Design Methodology

2.8.1. Introduction

Layering is the dominating design methodology of communication protocol stacks. An essential feature of the layering principle is layer-independence (or modularity), and

thus in a strict layered protocol stack, cross-layer communications are considered as violation. However, keeping the strict layering all the time can be cumbersome and may result in an inefficient implementation of a protocol suite. Therefore, the cross-layer design methodology [Haas 2001] has been introduced. The extreme implementation of the methodology is to merge all the interested layers into one flat single layer. This is absolutely orthogonal to the strict layered structure. Between the two extremes, cross-layer signalling can be introduced to a protocol stack to facilitate cross-layer interactions, and this approach would only require limited modifications to the existing protocol stack. Therefore, this approach is preferred in most cross-layer designs.

We argue that the cross-layer design methodology can play an important role for the next-generation wireless system, featured by all IP-based protocol stack, heterogeneous access networks, and multimedia data traffic [Evans and McLaughlin 2000]. We have seen that L2 handoff notifications are crucial to IP-based L3 fast handoffs. In fact, the cross-layer design can be fully justified in the wireless and mobile networks:

Firstly, the assumptions in the wired IP stack are inadequately suitable for the wireless networking. For example, one of the well-known assumptions in TCP protocol is that packet loss is caused by network congestion. However, in wireless systems, packet loss often occurs due to corruption. The congestion avoidance procedure can only make things worse. Exposing the packet corruption rather than congestion in the signalling from the link layer to the transport layer will facilitate an easy solution to this problem [Balakrishnan etc 1997, Balakrishnan 1998].

Secondly, the heterogeneity of network and traffic calls for a coordinated adaptation from multiple layers. Introducing a single collocated layer for various adaptation tasks would be too complex and heavy. The QoS adaptation even requires participation of all

layers [Haas 2001]. Therefore, a co-operation of multiple layers' adaptation would lead to a simpler and more flexible approach.

Thirdly, the rare radio resource and the limited power necessitate the optimisation of network performance; such optimisation can hardly be met in the sub-optimal wired architecture with strict layering. For example, error correction schemes are provided in both the link layer and the transport layer. In wireless systems, these schemes have to be invoked much more frequently to combat the errors due to unreliable channels. A co-ordination of the two layers can thus result in a more efficient solution [Wu etc 1999].

Fourthly, the emerging short-range networks such as ad hoc network and PAN entail an integrated design approach. For instance, in traditional networks the link layer is for point-to-point communications, while the transport layer is for end-to-end communications across various links. In short-range networks, the peer-to-peer communications mostly take place in the point-to-point level. By cross-layer design, duplicate efforts from each related layer can be avoided [Chen etc 2002].

In the cross-layer design methodology, two essential issues deserve further investigation. One is "What information should be exchanged across layers?" The answer to this question is certainly mission-oriented and algorithm-specific, as indicated in the above cases, among a lot more others. The other one is "How should such information exchange be performed?" The answer to this question is crucial to an efficient and effective cross-layer design. Nevertheless, research on cross-layer signalling methods lags behind in the cross-layer design methodology, and the remaining of this section contributes to this topic. The nature of cross-layer signalling is twofold. For one thing, the choice of a cross-layer signalling method largely depends on the mission and the corresponding implementation is often protocol-specific. For another, despite such dependency cross-

layer signalling methods, as vehicles for information exchange, can be generally classified and possibly standardised.

2.8.2. Cross-Layer Signalling Schemes

2.8.2.1. Method 1- Packet Headers

In IPv6, optional network-layer information can be encoded in additional headers. The Interlayer Signalling Pipe (ISP) briefed in [Wu etc 1999] takes advantage of this new feature by storing cross-layer information in the Wireless Extension Header (WEH) as shown in Figure 2.26. This method makes use of IP data packets as in-band message carriers with no need to use a dedicated internal message protocol.

However, normally an IP packet can only be processed layer by layer, and the conceptual top-to-bottom "signalling pipe" seems excessive in most cases. Moreover, an IP protocol stack only allows a header to be inserted into a packet delivered from a higher layer to a lower layer (downwards); therefore, this method is hardly applicable to upward cross-layer signalling. Furthermore, an extension header is usually placed between the IPv6 header and the transport-layer header in a packet, and this indicates that the WEH mainly facilitates network-layer information to be populated to lower layers. The latter two restrictions are relieved in [Gao etc 2004], where the WEH is generalised into a data structure called Cross-Layer Tag (CLT) and upward signalling is enabled by using a shared memory area. Another restriction in this method is that a lower layer may find it difficult (sometimes impossible) to access to the cross-layer information when header encryption or compression techniques are applied. All these restrictions affect the usability of this method. Finally, defining an extension header for cross-layer signalling should adhere to the IPv6 recommendations, e.g., the size of an extension header must be an integer

multiple of 8 bytes, and the new header's ordering constraint relative to the existing extension headers must be specified to facilitate the processing.

In addition to extension headers, reserved bits in existing headers can also be exploited. In [Balakrishnan etc 1997, Balakrishnan 1998], only one bit in the TCP packet header was used for Explicit Loss Notification (ELN) by a link-layer software agent Snoop in the Base Station. When Snoop is aware of a packet loss due to corruption, it sets the ELN bit in the TCP header and generates the in-band signalling as a feedback to the MH. This scheme suits a simple Boolean notification but does not scale well to bear complex control information.

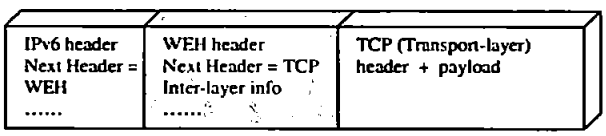


Figure 2.26 Bear cross-layer information with extension header

2.8.2.2. Method 2- ICMP Messages

ICMP (Internet Control Message Protocol, [RFC792] for ICMPv4 and [RFC1885] for ICMPv6, respectively) is a widely deployed signalling protocol in IP-based networks. Compared to the "pipe" described above, Method 2 [Sudame and Badrinath 2001] is to "punch holes in the protocol stack" and propagate information across layers by using ICMP messages as shown in Figure 2.27.

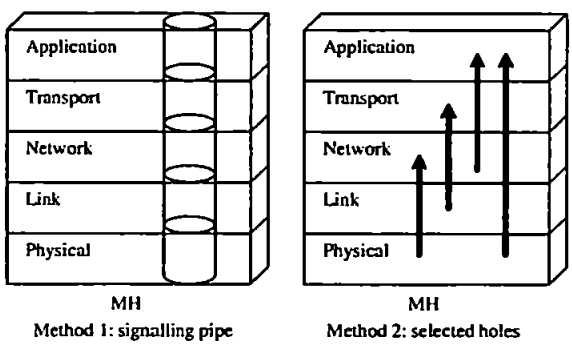


Figure 2.27 Comparison of cross-layer Method 1 and Method 2

In this scheme, desired information is abstracted to parameters, measured by corresponding layers wherever convenient. A new ICMP message is generated only when a parameter changed beyond the thresholds. Since cross-layer communications are carried out through selected "holes" not a general "pipe", this method seems more flexible and efficient. Furthermore, Method 2 is more mature since it has been implemented on Linux operating system (OS) with APIs (Application Program Interfaces) developed. However, an ICMP message is always encapsulated in an IP packet, and this indicates that the message has to pass by the network layer even if the signalling is only desired between the link layer and application layer.

2.8.2.3. Method 3- Network Service

In [Kim 2001], a specific access network service called Wireless Channel Information (WCI) was proposed. In this scheme, channel and link states from the physical layer and the link layer are gathered, abstracted and managed by third parties, the distributed WCI servers. Interested applications then access to the WCI for their required parameters from the lowest two layers as shown in Figure 2.28. Although it is not a cross-layer signalling scheme within an MH, we can deem it complementary to the former two schemes, as further implementation problems are considered in parameter definition, abstraction, coding, and decoding. However, any intensive use of this method would introduce considerable signalling overhead and delay over a radio access network.

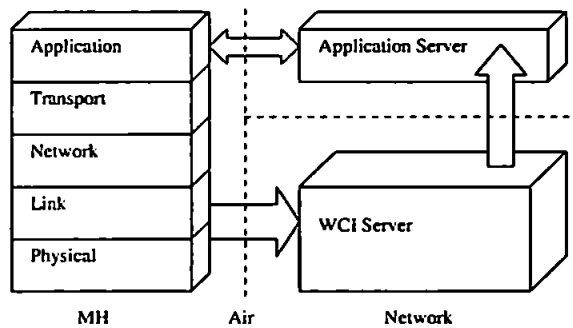


Figure 2.28 Concept model of cross-layer Method 3 (network service)

2.8.2.4. Method 4- Local Profiles

In [Chen etc 2002], local profiles are used to store periodically updating information for an MH in an ad hoc network as illustrated in Figure 2.29. Cross-layer information is abstracted from each necessary layer respectively and stored in separate profiles within the MH. Other interested layer(s) can then select the profile(s) to fetch the desired information. Seemingly, this method looks like Method 3, which stores the cross-layer information separately and keeps it ready for future use. However, in this method, internal profiles rather than external servers are applied. Analogically, Methods 1 and 2 store cross-layer information in memory basically, Method 3 stores the information in a network server, while Method 4 does this in local hard disk. Method 4 is flexible since profile formats can be tailored to specific applications, and the interested layers or applications can access the desired information directly. However, it is not ideal for time-stringent tasks.

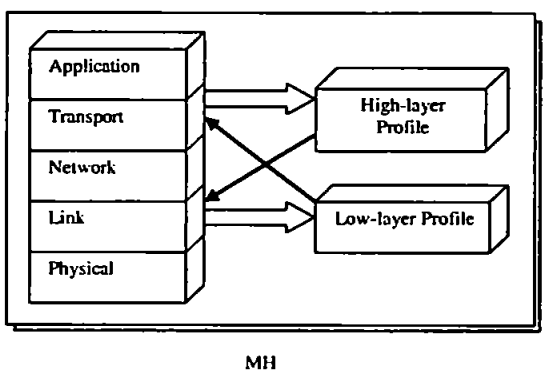


Figure 2.29 Concept model of cross-layer Method 4 (local profiles)

2.8.3. Shortcomings of the Existing Schemes

From the above discussion, a couple of major drawbacks of the existing methods can be identified. First, the signalling propagation paths across the protocol stack are not efficient. The layer-by-layer propagation approach just follows the data propagation mode. Consequently, the intermediate layers have to be involved even if only the source layer and

the destination layer are actually targeted. This will cause unnecessary processing overhead and propagation latency. Second, the signalling message formats are either not flexible enough for active signalling in both upward and downward directions, or not optimised for different signalling inside and outside the MH respectively. Furthermore, the desired message formats should be scalable enough for rich signalling more than cross-layer hints and notifications [Larzon etc 2002].

2.9 Summary

In this chapter, we have surveyed the background and the state-of-the-art work on mobility support, including handoff and location management, and related work. The reference protocol stack comprises physical, link, network, transport and application layers. The mobility protocols in the existing generations (before the late stage of 3G) are link-layer based, and thus optimised for mobility management in homogeneous systems. For 3G and beyond, an all-IP vision is widely acclaimed and IP-based mobility protocols are expected to support seamless mobility across heterogeneous networks.

In an all-IP system, access networks are commonly organised into administrative domains, interconnected to each other through a core IP network, e.g., the evolved Internet. Macro-mobility protocols are suitable for inter-domain mobility management whilst micro-mobility protocols can support intra-domain mobility more efficiently and effectively. The dominant macro-mobility protocols are the Mobile IP (MIP) family and the Session Initiation Protocol (SIP), running in the network and the application layer, respectively. Hybrid MIP-SIP architectures are emerging to exploit the complementary merits of MIP and SIP. Alternate macro-mobility protocols mainly operate around the transport layer. Regarding micro-mobility protocols, tunnelling and host-specific routing are the two major approaches. Additionally, fast handoff protocols are designed to utilise the link-layer

mobility information to accelerate the IP-based handoff operations. For IPv6 micro mobility, the combined HMIPv6 and FMIPv6 approach with QoS support is promising. The principle operations of these macro- and micro-mobility protocols are outlined and the detailed signalling is illustrated for the protocols essential in this thesis. Furthermore, their strengths and the weaknesses are discussed.

In addition, the cross-layer design methodology has justified its application in the wireless networks, e.g., IP-based fast handoffs usually rely on L2 handoff notifications. Cross-layer signalling schemes are a key enabler to many cross-layer designs. Several cross-layer signalling schemes have been proposed sparsely in the literature, and may be applied in specific context despite their shortcomings.

To sum up, based on the extensive literature review presented in this chapter, we can find that much more research is still needed towards a more useful mobility support for next-generation wireless networks.

Chapter 3

A Cross-Layer Perspective on Next-Generation Mobility Support

In this chapter, we present an overview of the proposed mobility support framework. We start with an investigation of the requirements and challenges in the next-generation (Beyond 3G or B3G) mobility support as the problem statement and outline the project roadmap in Section 3.1. In Section 3.2, we propose a new generic, efficient and flexible cross-layer signalling method. Subsequently, we envision a generic multi-layer framework for comprehensive mobility support with cross-layer interactions in Section 3.3, and then explain the crucial building blocks in the framework and specify our design emphasis in Section 3.4. Finally, Section 3.5 summarises the chapter. This chapter is partially based on three publications [Wang and Abu-Rgheff CE, WCNC03 and EPMCC03].

3.1 Problem Statement: Next-Generation Mobility Support Requirements and Challenges

3.1.1. Requirements of Next-Generation Mobility Support

3.1.1.1. The Necessity of Distinguishing Macro and Micro Mobility

As aforementioned in Chapter 2, to minimise the global signalling loads between an MH and its peer entities (home mobility server and CHs) and expedite the responses to intra-domain movements, preferably micro-mobility schemes should be introduced to

complement the macro-mobility protocols such as Mobile IP (MIP) and the Session Initiation Protocol (SIP), the two dominant approaches [Kwon etc 2002]. The separation of macro and micro mobility allows us to investigate optimisations catered to each distinct scenario and provide most appropriate solutions. Notably, both macro and micro mobility schemes should cooperate smoothly in a uniform networking platform.

3.1.1.2. The Necessity of Distinguishing Real-Time and Non-Real-Time Mobility

As an important vision of the next generation, mobility support for multimedia applications is desired. Due to the distinguished traffic characteristics, QoS requirements and underlying protocols, multimedia applications, real-time or non-real-time, should be treated separately for effective mobility support. In the all IP context, real-time and non-real-time applications usually run over RTP/UDP and TCP, respectively; and thus the separation of TCP and UDP traffic mobility support should be emphasised.

3.1.1.3. The Capability of Interworking with QoS Schemes

From the users' perspective, they expect equivalent or similar service quality in wired and wireless networks, even in the presence of mobility. Mobility management schemes ensure correct routing of packets to or from an MH as it changes its point of attachment to a wireless network yet without any QoS commitments. Therefore, it is desired to incorporate QoS management schemes with mobility extensions to support QoS-sensitive services, especially UDP real-time applications [RFC3583]. These QoS schemes should be able to interact with both macro and micro mobility schemes.

3.1.1.4. The Capability of Supporting Advanced Mobility

So far, we have been focused on the IP terminal mobility, referring to the capability to enable an MH to send and receive packets regardless of network attachment. Terminal mobility is a conception extended from current and previous generations of wireless systems, and must be supported in the next-generation networks. In the mean time, we

have noticed that the complexity and requirements of mobility management are growing with the evolution. The well-established 2G mobility procedures were designed only for terminal mobility of voice and a bit of data in a homogeneous system. Nevertheless, in the next generation, a mobile user may access to all IP-based heterogeneous networks for various services and multimedia sessions via a set of personal devices anywhere and anytime [Fasbender etc 1999]. Consequently, more mobility types are emerging, and selected ones are defined as follows based on [Schulzrinne and Wedlund 2000]:

Personal mobility refers to the capability of the network to reach a user globally using his or her unique personal ID (identifier) and the capability of a user to originate or receive a session by access to any authorised terminals. Session mobility is the ability that a user can maintain an ongoing session while changing terminals, say, from mobile phone to laptop PC, especially within a personal area network (PAN).

Furthermore, we define a couple of additional mobility types. Firstly, Ad Hoc mobility refers to the scenario where in an ad hoc network MHs can communicate with one another without a fixed infrastructure. Any of them can act as a router to relay a session for others. A caller and the callee can also directly establish a session if near enough. Secondly, mode mobility is the capability that an MH can switch between the infrastructure mode and the ad hoc mode, i.e., communicate with each other via the fixed network or the ad hoc network. The IETF network mobility (NEMO) [RFC3963] can be deemed as a special case of the mode mobility. In NEMO, an ad hoc network moves together as if a single node and interacts with the infrastructure network through a common gateway called mobile router, a host selected from the ad hoc network.

These mobility types, together with the terminal mobility, can be categorised as high-level mobility (personal and session mobility) and low-level mobility (terminal, ad hoc and mode mobility). Preferably, all these mobility types should be supported or facilitated in a

uniform framework, and wherever appropriate some advanced mobility features may be introduced as integral parts into the macro and micro mobility architecture, which is centred on IP-based terminal mobility. Existing projects such as Mobile People [Maniatis etc 1999] and ICEBERG [Wang etc 2000] handle part of the user-level mobility using proprietary protocols, though a standard-based protocol such as SIP is preferred with proper extensions [Schulzrinne and Wedlund 2000].

3.1.2. Design Challenges of Mobility Support Schemes

3.1.2.1. Powerfulness and Flexibility

Considering the complex requirements posed by the next-generation mobility support and the evolutionary development of wireless systems, we argue that the proposed mobility support framework should be designed to be both powerful and flexible to meet the challenging requirements in a progressive way. The framework should be capable to handle both macro and micro terminal mobility, support both real-time and non-real-time applications, allow QoS commitments for real-time applications, and facilitate various advanced mobility types. Meanwhile, the design methodology should be geared towards incremental development and deployment of these capabilities, be open to other mobility-related add-on designs like QoS adaptations and be compatible with infrastructure expansions.

3.1.2.2. Based on Standards

"Standard is king" in the ICT (Information and Communication Technology) world. Thus, the proposed mobility support architectures should be designed based on the protocols that have been standardised (e.g., the IETF RFCs) or being standardised as the most promising candidates (e.g., the IETF Internet drafts that are regularly discussed and updated in the standard track). Note that this requirement does not mean to rule out useful

optimisations of the involved standard schemes, which themselves are actually being evolved in the standardisation bodies like IETF, e.g., from a "Proposed Standard" to a "Draft Standard".

3.1.2.3. Cost-Effectiveness

No powerful architectures can be established without prices and little advances can be made without additional costs. However, the proposed architectures should be optimised to reduce the known costs incurred in similar architectures for common mobility procedures. The most interested costs are the signalling costs generated by mobility messages, which impose a significantly burden on the whole system and have attracted a great deal of research in the past decades (e.g., [Pollini etc 1995, La Porta etc 1996, Akyildiz and Wang 2002, Lo etc 2004]). Thus, minimising signalling costs can greatly improve the efficiency of the system.

3.1.2.4. Handoff Performance

Moreover, the cost optimisations should be achieved without sacrificing the mobility performance under major metrics; instead, with other enhancements the proposed architecture is expected to lead to improved performance. In particular, handoff performance is important for effective mobility support perceptible to mobile users. Therefore, compared with existing schemes, the proposed mobility framework is desired to accomplish superior performance in terms of handoff delay and handoff packet loss, among other criteria, during handoffs.

3.1.3. Project Roadmap

With the requirements and challenges for next-generation mobility identified, we outline the roadmap of the project. Figure 3.1 depicts the big picture of the proposed

architectures and their relationships under the umbrella of the expected comprehensive mobility support framework.

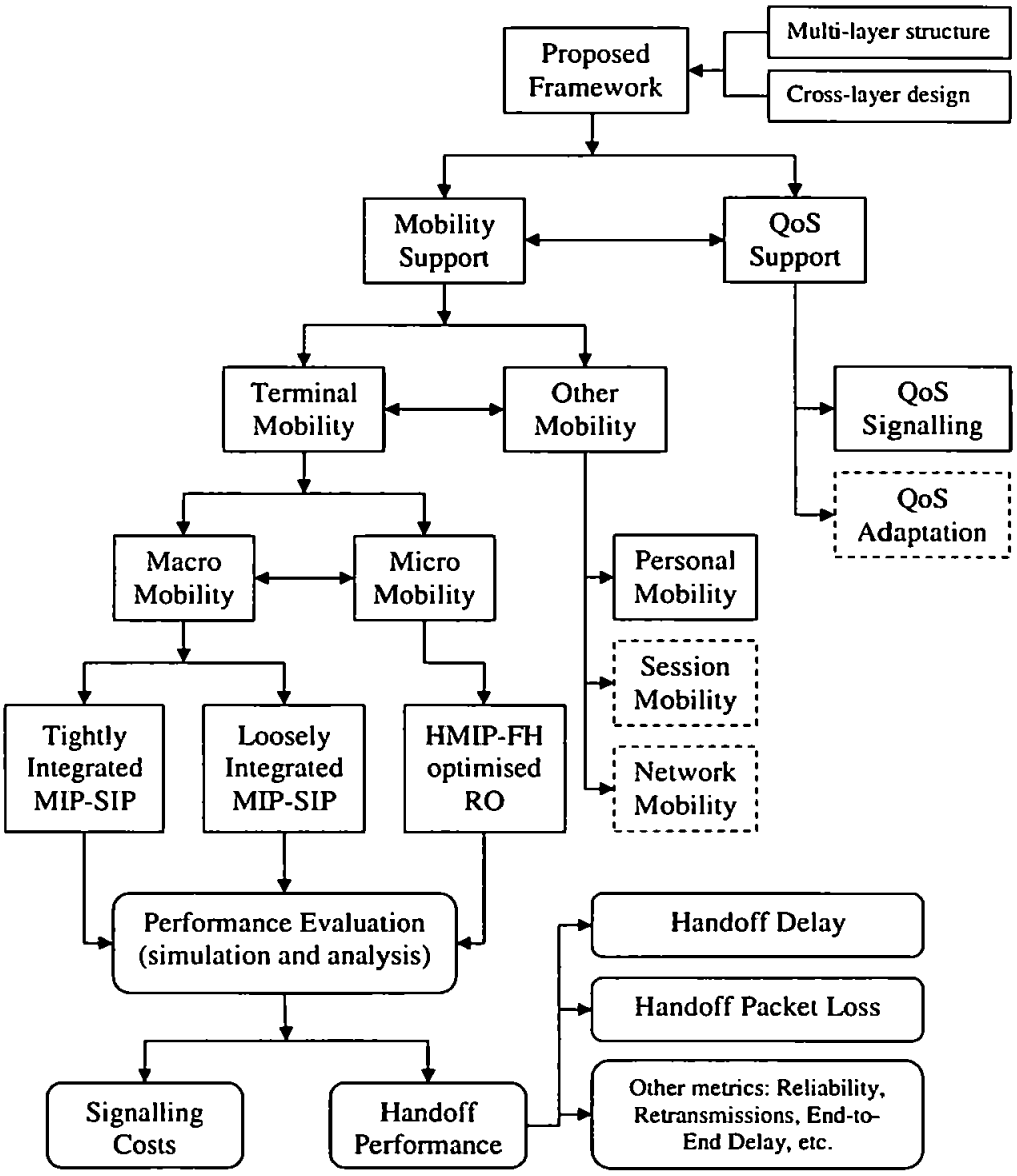


Figure 3.1 Outline of the Proposed Framework

Generally, the proposed framework adopts a cross-layer design approach and comprises a multi-layer structure exploiting extensive contributions to mobility support from multiple protocol layers. These two issues are addressed in Section 3.2 and Section

3.3, respectively. More specifically, various mobility scenarios are investigated from different angles including terminal and personal mobility, macro and micro mobility. Three architectures are designed and evaluated as the fundamental building blocks for the framework. Here, these architectures are referred to as Tightly Integrated MIP-SIP, Loosely Integrated MIP-SIP, and HMIP-FH optimised RO, respectively. They are all based on standardised protocols. This project also considers additional emerging mobility types such as session and network mobility, and QoS support including QoS signalling and adaptation. Performance evaluation is conducted through analysis and simulation under a set of metrics covering both cost-efficiency and handoff performance. Further explanations on these specific issues are provided in Section 3.4 (except QoS adaptation, which is discussed in Section 3.3). In Figure 3.1, the topics in the solid blocks are addressed in depth in this thesis, whilst the issues in the dotted ones are briefly discussed for completeness purposes.

3.2 The Proposed Cross-Layer Scheme CLASS: Cross-LAyer Signalling Shortcuts

3.2.1. Rationale for a Cross-Layer Design Approach

From the perspective of the protocol stack, the network layer is the most appropriate level to converge heterogeneous networks in an all IP vision. MIP has been standardised, targeting terminal mobility. MIP hides mobility using tunnelling/encapsulations from upper layers, which is especially useful for TCP-based applications because they must maintain unchanged IP addresses during their session lifetime. Though MIP can be extended for NEMO [RFC3963], it can hardly support the high-level mobility due to the

inborn network-layer constraints, i.e., the lack of advanced features specific to applications and user-level mobility requirements.

On the other hand, the application-layer SIP was designed for the management of multimedia sessions, especially UDP applications. Operating in the application layer, SIP can be extended [Schulzrinne and Wedlund 2000] to resemble MIP terminal mobility operations whereas more importantly it can further provide advanced and unique mobility features such as session renegotiation or update, and thus would improve the application QoS when handoffs take place. Unlike MIP, SIP inherently supports personal mobility with SIP infrastructure under worldwide deployment. Although MIP could be extended to achieve some (very limited) of these features such as a user-level identifier, it is not cost-effective to duplicate the standardisation efforts and it is difficult to extend MIP to fulfil many of SIP mobility functions conveniently operating at the application layer. SIP also has the potential capabilities to support other high-level mobility types by augmented signalling [Schulzrinne and Wedlund 2000]. However, currently SIP is not so much a mature mobility solution than an initial framework. When extended for terminal mobility, much complexity would be added to enable SIP support mid-session TCP mobility [Vakil etc 2001]. Furthermore, SIP messages incur additional processing delay in the application layer compared with MIP operating in the kernel of the operating system. In addition, in contrast to the binary-coded messages in MIP, SIP is much more generous in message size since SIP messages are text based, which indicates that SIP-only (Pure SIP) mobility will generate much higher signalling loads compared with MIP for terminal mobility. Therefore, the SIP-alone approach for a complete mobility support seems questionable.

Furthermore, some functions of the traditional link-layer mobility support could be utilised wherever available and appropriate. In addition, the link layer, together with the physical layer, could also help to tackle network-specific problems resulting from mobility,

such as adaptation to the next-generation heterogeneous communication environments. These tasks are beyond the network- and the application-layer mobility schemes including both SIP and MIP approaches. However, advanced mobility support architectures should consider all the above issues as well as support various mobility types.

In sum, the lessons we have learned are that a single-layer-specific mobility architecture can hardly meet the next-generation mobility support requirements. The intrinsic reason is that mobility brings about significant impacts on each layer, which in turn has its convenience to deal with different level mobility impacts. Thus, introducing a single collocated layer for various mobility tasks, if possible, would be too complex and heavy. Therefore, it would be simpler and more flexible to develop a co-ordinated multi-layer architecture that can make full use of each layer's contributions while still keeping the basic structure of the TCP/IP protocol suite.

In the meantime, cross-layer (or inter-layer) design [Haas 2001], especially via cross-layer signalling methods, has justified its introduction into wireless systems. In fact, this methodology has been successfully applied in several areas, such as error correction [Wu etc 1999], adaptation of wireless protocols [Sudame and Badrinath 2001], and optimisation of ad hoc networks [Chen etc 2002]. Obviously, there exists a good case to combine the multi-layer mobility support architecture and the cross-layer design methodology.

3.2.2. Design of CLASS

As aforementioned, next-generation mobility support entails a cross-layer design approach so that contributions to mobility support from multiple layers could be exploited in a coordinated and collaborative way for efficiency and effectiveness. However, the existing cross-layer signalling schemes seem neither efficient nor flexible enough as

discussed in Section 2.8 in Chapter 2. Therefore, we propose a method, named CLASS, as an efficient, flexible and comprehensive scheme with the following distinct features.

Firstly, flexible direct interactions between non-neighbouring layers are enabled. The basic idea is to break the layer ordering constrains while keeping the layering structure, i.e., let cross-layer messages propagate through local out-of-band signalling shortcuts. For instance, enable the direct communications between the application layer and the network layer without turning to the otherwise middleman, the transport layer. Although this approach is not unknown to the protocol stack designs, it only appeared as exceptions and was not designed for generic management functionality. Surely, this scheme also applies to signalling between neighbouring layers. The concept of this feature is demonstrated in Figure 3.2.

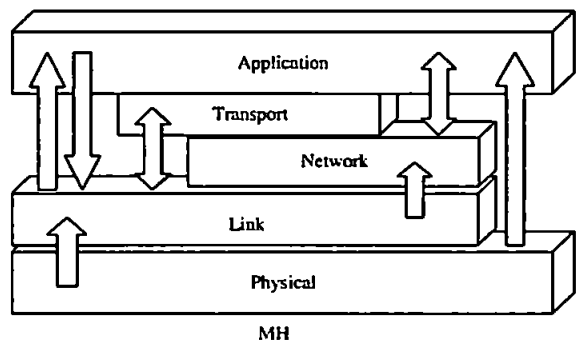


Figure 3.2 Concept model of CLASS

Secondly, light-weighted internal message format is designed. For internal signalling, it is not necessary to use standardised protocols, which are normally heavy-weighted for, e.g., transmission against errors in the network. For instance, Mehtod 2 [Sudame and Badrinath 2001] uses ICMP messages for internal signalling. In addition to the large IP header (20 or 40 bytes for IPv4 and IPv6 respectively without extension headers), a common ICMPv6 header itself is 8 bytes, where the required *checksum* field is 2 bytes, occupying 25%. Therefore, reducing additional headers and minimising the fields can simplify the internal message format. Although header compression techniques are

emerging, e.g., in the IETF Robust Header Compression or ROHC Working Group (WG), efficient message definition in the first place is still crucial. Generally, three essential fields are required in CLASS:

Destination Address, indicating the destination layer and destination protocol(s) or application(s).

Event Type, indicating an interested parameter, e.g., a new IP address or a specific L2 trigger.

Event Contents, the value of the parameter.

If we assign 2 bytes to the *Destination Address* and the *Event Type* respectively and let the *Event Contents* field takes 16 bytes, the whole message size is 20 bytes. Similarly, we examine an ICMPv6 message with 8-byte header and 16-byte contents, encapsulated in an IPv6 header (40 bytes). The whole message size is 64 bytes, more than twice bigger than that of CLASS. To improve the signalling efficiency even further, messages can also be propagated in an aggregate way by introducing an optional field, *Next Event*.

As to the external cross-layer signalling, standard protocol messages (not limited to ICMP) should be used. For complex cross-layer interaction scenarios, a message control protocol is expected to guarantee that dense simultaneous messages across layers can be exchanged in an optimised and organised way to achieve high efficiency and avoid possible conflicts. Regarding message generation and reading, the mechanisms in Method 2 [Sudame and Badrinath 2001] can be based on. In general, a message with a layer-specific parameter is generated from the specific layer whenever a significant change to the parameter happens (e.g., the parameter falls below or rises above a predefined threshold). Function calls are used to set and get the parameter, and system calls are used to read the message.

Notably, a specific implementation of the CLASS model may depend on the capabilities of the host OS. In an OS that does not facilitate signalling between non-neighbouring layers, CLASS may retreat to the layer-by-layer approach as in Method 2 though the efficient use of internal messages is still applicable. If preferred, the core CLASS concept may even be implemented in the user space so that modifications to the OS can be minimised. In that case, the user space module may act as a hub to convey the interactions in a simple task or as a coordinator and controller in more complex tasks. In addition, the actual interactions between layers and the corresponding external signalling (if involved) are task-dependent and protocol-specific.

In short, the design of CLASS is expected to serve as a generic, efficient and flexible model that could allow different implementation and application scenarios. The specific application of CLASS, together with other methods, in a proposed mobility support framework is described in Section 3.3.

3.2.3. Evaluations and Discussions

3.2.3.1. Evaluation Criteria

The following evaluation metrics are defined to reflect the major concerns when choosing or designing a cross-layer signalling method.

- Internal overhead (overhead within an MH): This metric is mainly determined by the complexity and average size of internal messages, the number of involved layers, and the signalling frequency. Reducing this overhead entails an optimised lightweight message format and signalling thresholds should be applied to avoid excess signalling. For comparison purpose, equal signalling frequencies are assumed.
- External overhead (overhead in the network): In some contexts, cross-layer signalling between an MH and a CH or a network node is desired. In these cases, it is

recommended that standard protocols be reused wherever appropriate. The incurred overhead is determined by the message size (or added length, e.g., of an extension header), the propagation distance in terms of IP-level hops and the signalling frequency. For comparison, equal propagation distances and signalling frequencies are assumed.

- Propagation latency: This refers to the time taken by the signalling transmission between the source layer and the target layer. The decisive factors include the propagation path and travel time between interfaces, and intermediate processing time (including queuing delay) in each layer along the path. For comparison purpose, more than one intermediate layers are assumed.
- Propagation direction: Cross-layer signalling messages can be propagated from lower layers to higher layers (upwards) or vice versa (downwards). Bi-directional propagation is required for cross-layer interactive tasks.
- Reachability: A generic cross-layer signalling method should enable signalling between any two arbitrary layers.
- Implementation complexity: This refers to the design requirements to implement a cross-layer signalling method, such as the different levels of OS modifications, and internal or external interface design.

3.2.3.2. Comparison of the Methods

Table 3.1 compares the existing methods with CLASS under the above criteria, and the main differences between CLASS and the other methods are explained as follows.

In contrast to the other methods as shown in the table, CLASS overcomes the two major drawbacks aforementioned in Section 2.8.3 in Chapter 2. Firstly, since CLASS uses the unique active and direct signalling between any two arbitrary layers in both directions, it has the lowest propagation latency with high efficiency and flexibility.

Table 3.1 Comparison of the cross-layer signalling methods

Criterion \ Method	Signalling Pipe (Method 1)	Selected Holes (Method 2)	Network Service (Method 3)	Local Profiles (Method 4)	Shortcuts (CLASS)
Internal message format	Extension header or variant (WEH or CLT)	ICMP	N/A	Author-defined	Lightweight, optimised
Internal overhead	Low to medium	Medium	N/A	Low to medium	Low
External message format	WEH or CLT	ICMP	Author-defined	N/A	Standard-based
External overhead	Low to medium	Medium	High	N/A	Low to medium
Propagation scheme	Layer-by-layer data or messages	Layer-by-layer messages	Messages over air	Read and write profiles	Direct messages between any 2 arbitrary layers
Propagation latency	High	Low	Highest	Medium (periodic)	Lowest
Propagation direction	WEH: downward CLT: bi-directional	Upward (basically)	N/A	Bi-directional	Bi-directional
Reachability	WEH: low CLT: medium	Medium	Low	High	High
Implementation complexity	WEH: low CLT: medium	Medium	High	Medium	High

The following presents a simple analysis of propagation latency across the protocol stack. For methods where a message travels layer by layer, the upward (or vice versa) propagation latency of a given message between any two layers, say layer 1 (the source layer, not necessarily the physical layer) and layer n (the destination layer, $1<n \leq 5$ in this case), can be formulated as:

$$T_{L1 \rightarrow Ln} = \sum_{i=1}^{n-1} (T_{ri} + T_{pi}), \tag{3.1}$$

where T_{ri} denotes the transport time between the interfaces of layer i and layer i+1, and T_{pi} denotes the processing time (including any queuing delay) at layer i+1.

Let

$$T_p = \sum_{i=1}^{n-1} T_{pi} \tag{3.2}$$

and assume

$$T_{ri} = T_r, \tag{3.3}$$

we obtain the value expression:

$$T_{L1 \to Ln} = (n-1) \times T_r + T_p . \tag{3.4}$$

For CLASS, the expression of the same metric is given by

$$T'_{L1 \to Ln} = T_r + T_{p(n-1)} . \tag{3.5}$$

Assuming the processing time is the same at each layer, then

$$T_{p(n-1)} = T_p / (n-1) . \tag{3.6}$$

Finally, summarising equations (3.4) to (3.6), we reach the conclusion:

$$T'_{L1 \to Ln} = T_{L1 \to Ln} / (n-1) . \tag{3.7}$$

In contrast to the layer-by-layer approach, the propagation latency in CLASS is only about $1/(n\text{-}1)$ as large. The more the layers, the more significance it makes. Only when n=2 (signalling between neighbouring layers), there is no difference. Note that bypassing the intermediate layers also lead to reduced processing costs as processing time is an indicator of processing costs.

Secondly, CLASS purposely distinguishes between the internal and external messages, and applies optimised or standardised formats for internal or external signalling respectively. Hence, it has the lowest overhead when applied within an MH and has a low overall overhead when implemented between an MH and its access network as well. Moreover, CLASS does not exclude the simultaneous use of other methods under some specific circumstances. Therefore, complex as it is, its efficiency, flexibility and scalability will justify its wide application perspective.

In addition, it is worthy noting that cross-layer designs would benefit those areas where a "global" system factor (GSF) is the target. A GSF can be defined and generated from one of the following three sources. First, the original layer separation and abstraction of a protocol stack had difficulties in clearly placing one service in a single layer, e.g., error correction exists in both link and transport layers to fight errors in different levels.

Second, the GSF itself is a system-level factor by nature, and can hardly be handled thoroughly in a specific layer. Examples include QoS, resource, energy (power), and security, whose better management would require a collaboration of multiple layers. Third, a GSF can also be a significant change to the original design basis of a protocol stack. Wireless and mobility are good examples, which challenge many design assumptions in the TCP/IP suites and affect all the layers' behaviours. Thus, mobility support and wireless adaptation would be another two application areas.

3.2.3.3. Discussions on Standardisation Work

As a promising design methodology, cross-layer design should take a cautionary approach because of the added complexity to the protocol stack [Kawadia and Kumar 2005]. Therefore, standardisation on cross-layer design is in need to ensure compatible and holistic designs.

It is worth noting that CLASS-style direct communications between non-adjacent layers have appeared in 2G and 3G standard wireless systems though such an approach has not been generalised as in CLASS. For instance, the L3 module RRM (Radio Resource Management) or RRC (Radio Resource Control) directly communicates with the physical layer in GSM [Walke 2002] and UMTS [Korhonen 2001], respectively. In fact, justified by the highly dynamic characteristics in wireless mobile systems, the telecommunication standardisation bodies like 3GPP do not tightly adhere to the layer-independence principle as the IETF usually does. Therefore, it is reasonable to expect more cross-layer signalling cases (including but not limited to the CLASS style) to emerge in the next-generation wireless systems considering the even more complex communication environments, e.g., heterogeneity in every aspect.

Meanwhile, recently in the IETF some strong indications have emerged on favouring cross-layer design under certain circumstances, driven by the booming convergence of

Internet and wireless systems. For example, a number of cross-layer issues are discussed and potential solutions through cross-layer designs can be inferred in the IETF advice given to network designers [RFC3819]. Firstly, the interactions between TCP and the link-layer Automatic Repeat Request (ARQ) protocol for error and delay control are examined, and coordination between these two competing protocols is suggested. In addition, the misinterpretation of packet loss due to wireless corruption as congestion is also acknowledged. Thus, an L2-L4 dialogue would be expected. Secondly, an L2-L3 interface is explicitly stated as an ideal solution to properly deal with packets during a temporary outage and thus prevent undesirable TCP back-offs. Similarly, an L2 to L3 signalling is desired for link layer to inform network layer of the offered delay and jitter so that IP QoS support protocols like IntServ (Integrated Services) [RFC1633] and RSVP (Resource ReSerVation Protocol) [RFC2205] can be facilitated. Thirdly, a real-time application such as a voice codec requires a mechanism to signal its tolerance of corrupted payload to use UDP-Lite [RFC3828] and to indicate the packet protection coverage to the link layer. Again, cross-layer signalling between L5 and L4/L2 is hinted here.

However, so far no explicit dedicated standardisation has been launched in the IETF for cross-layer signalling despite the fact that ad hoc designs do exist in its RFCs and Internet drafts. Currently, one may expect that the next-generation signalling protocol being standardised by the IETF NSIS (Next Steps in Signalling) WG could be exploited for external (and possibly internal) cross-layer signalling. The NSIS itself employs a two-layered paradigm, where the lower layer offers generic signalling transport whilst the upper layer provides application-specific signalling, e.g., QoS signalling. The proper interactions between these two layers and the underlying IP layer are also being defined. Conceptually, this paradigm is well suited for the nature of cross-layer signalling mentioned in the Introduction and thus can be a large step in this area.

In addition to the 3GPP and the IETF, the IEEE is another key player in standardising the related work, especially the 802 series. Notably, explicit standardisation work on cross-layer signalling has been underway in the IEEE 802.21 WG. A draft standard [IEEE802.21] has been proposed to optimize network-layer handoffs between heterogeneous 802 systems and between 802 systems and cellular systems by utilising link-layer indications (triggers). A set of primitives regarding cross-layer events and commands have been defined and will be introduced in Section 3.3.

In sum, there is a strong tendency, and actually a need, to accelerate cross-layer signalling standardisation for cross-layer design convenience and system compatibility considerations. With the joint efforts from the leading standardisation organisations and the general research community, one can be optimistic about the standardisation future of cross-layer signalling.

3.3 The Envisioned Multi-Layer Mobility Support Framework

As indicated in the previous sections, comprehensive mobility support entails a cross-layer design approach, which may take advantages of the contributions from multiple protocol layers. In this section, we prospect a generic multi-layer mobility support framework, which attempts to combine individual layer's contributions through a cross-layer design.

3.3.1. Contributions to Mobility Support from Each Layer

In an IP-based protocol stack, in fact each layer has more or less positive or negative effects on mobility support. In the following, we abstract their possible (positive) contributions to mobility support.

3.3.1.1. Physical and Link Layers

The physical and link layers can report current channel conditions and link properties, respectively, to upper layers, which can then adapt to the mobility. These reports are collectively known as L2 triggers though L1 is actually often involved. Typically, by detecting and reporting the imminent arrival of a forced handoff to the network layer in advance, the link layer expedites the IP-based handoff significantly with such L2 triggers [Fikouras etc 2001, Festag 2002, Aust etc 2003]. Furthermore, in the link layer, different MAC techniques enable different L2 handoff schemes, which are network-specific but could be utilised by the network layer in order to improve handoff performance. For example, a CDMA-based system can facilitate a soft handoff.

Typical system-specific L2 triggers include RSS (received signal strength), SIR (signal-to-interference ratio), BER (bit error rate), FER (frame error rate) etc. Their availability in popular systems including WLAN, GSM, UMTS (WCDMA), and Bluetooth etc. is discussed in [Festag 2002] and is summarised in Table 3.2.

Table 3.2 Selected system-specific L2 triggers and their availability

L2 Trigger	IEEE 802.11b	GSM	UMTS (WCDMA)	HiperLAN/2	Bluetooth
RSS downlink	√	√	√	√	√
RSS uplink		√	√	√	
SIR downlink	√		√	√	√
SIR uplink			√	√	
BER downlink		√			
BER uplink		√			
FER downlink	√		√	√	√
FER uplink			√	√	

Furthermore, as aforementioned the IEEE 802.21 WG is standardising cross-layer signalling, especially L2 trigger primitives, to enable handoffs between both 802 and non-802 networks [IEEE802.21]. Selected triggers are tabulated in Table 3.3. Similar investigations are also underway in the IETF, e.g., the IETF DNA (Detecting Network Attachment) WG.

Table 3.3 Selected L2 trigger primitives

Generic L2 Trigger	Description
Link_Up	This trigger is delivered when an L2 connection is established on the specified link interface and when upper layers can send packets.
Link_Down	This trigger is delivered when an L2 connection is broken and when no more packets can be sent on the specified link.
Link_Going_Down	This trigger is delivered when an L2 connection is expected to go down (Link_Down) within a certain time interval. It may be an indication to initiate handoff procedures.
Link_Event_Rollback	This trigger is fired if the link is no longer expected to go down in the specified time interval in case of Link_Going_Down.
Link_Detected	This trigger indicates that a new type of link has been detected for use so that the terminal can attempt to gain connectivity.
Link_Parameters_Change	This trigger indicates changes in link parameters have crossed specified threshold levels.
Link_Handover_Imminent	This trigger is generated before the L2 handoff occurs. It contains information about the new point of attachment and any application-specific data that might be useful for the running application(s).
Link_Handover_Complete	The Transport and Application layers can resume flows upon receiving this trigger.
No_Link	This trigger indicates that the MH is moving out of the current service area and no link will be available. Thus, the mobile user may choose to sacrifice mobility to finish the ongoing session(s).

The implementation of an L2 trigger primitive may depend on specific algorithms that make use of one or more available system-specific L2 triggers and possible other additional information. For instance, from the fast handoff perspective, one of the most important L2 primitives is Link_Going_Down, which is used to anticipate an imminent handoff. This primitive may correspond to a decay of the downlink RSS, which is widely available in all kinds of wireless systems as suggested in Table 3.2. In fact, an L2 handoff occurs when the RSS falls below a predefined threshold as specified in typical wireless systems. It is noted that an L2 handoff does not necessarily indicate an L3 handoff unless additional proof is given that the new access point (AP) is administrated by a new access router (AR). Such information should be carried in Link_Handover_Imminent, which can be the PrRtAdv (Proxy Router Advertisement) message in the FMIPv6 [RFC4068]

protocol context. More precise mobility prediction may be assisted by exploiting other information such as service range declaration from an AP/AR, map or other navigation tools etc [Curran and Parr 2002]. For Link_Parameters_Change, a change of any of the available system-specific L2 triggers can generate it. For Link_Detected, it usually indicated by an unsolicited beacon or a response to an MH's L2 probe (scanning) from a new point of attachment.

In addition to the acceleration of forced handoffs due to movement, with additional system context variables such as costs, L2 triggers can also be used to determine a policy-based handoff to make trade-offs among costs and performances [Wang 1999] or even richer contexts [Vidales etc 2004 a]. In an overlay-networking environment where more than one system coexists, the better or the best system could then be chosen by performing an inter-system handoff even when a user has not moved out of the service coverage of the current system. This concept is also known as "Always Best Connected" [Gustafsson and Jonsson 2003].

3.3.1.2. Network Layer

The major job of the network layer is to support basic terminal mobility, including IP-based handoff management and location management for both macro- and micro-mobility scenarios, as extensively discussed in Chapter 2. Additional low-level mobility types such as the network mobility [Lach etc 2003] may also be handled at this IP level by extending network-layer mobility protocols [RFC3963]. MIP and its variants, operating in this layer, are the dominating protocols for these IP mobility scenarios.

Moreover, similar to L2 triggers, the network layer (e.g., a MIP host) can report the IP mobility events (e.g., L3 handoff initiation or completion) to the upper layers to initiate upper-layer protocol or application adaptations, or facilitate some services that can benefit from mobility-awareness.

Furthermore, QoS support is desired for real-time applications in the mobile environments, especially a QoS mechanism for MIP is required [RFC3583]. QoS-aware handoffs could be achieved with well-designed interoperation of MIP and IP-based QoS protocols such as RSVP/IntServ [RFC2210] and DiffServ (Differentiated Services) [RFC2475] with proper mobility extensions [Moon and Aghvami 2001 and 2003, Taha etc 2005]. The use of other QoS-related mechanisms like MPLS (Multiprotocol Label Switching) [RFC3031], simultaneously [Alam etc 2001] or alternatively [Chiussi etc 2002], may also be justified.

In addition, network-level AAA schemes are needed to support roaming users. The application of IP-based AAA protocols in the IP mobility context is being investigated in several IETF WGs including AAA, PANA (Protocol for carrying Authentication for Network Access), Mobike (IKEv2 Mobility and Multihoming), MIP4 and MIP6 etc., among others. Standardisation work is underway to enable MIP to collaborate with IPSec [RFC3776], Diameter [RFC4004] and other AAA schemes.

3.3.1.3. Transport Layer

The transport layer is concerned with end-to-end data delivery. In particular, TCP is expected to deliver a reliable transmission service despite the error-prone wireless links and user mobility. When terminal mobility is handled by a network-layer protocol like MIP, TCP can keep ongoing sessions alive since the IP address change of the MH is hidden from it. However, mobility does have a harmful impact on TCP performances. On a typical handoff, packet loss occurs; the packet loss is interpreted by TCP as a sign of congestion so that the congestion avoidance procedures are triggered. As a result, TCP underutilises the system resources, and thus application throughput is dramatically reduced. Moreover, a handoff usually causes the connectivity to be temporally lost and a timeout may be required before TCP initiates the recovery. This long pause further aggregates the

end-to-end performances. The mentioned mobility notifications from the MIP host could facilitate solutions to these problems.

One solution is to exploit the notification of handoff completion so that fast retransmission is invoked immediately when the handoff completes other than wait for the timeout [Caceres and Lftode 1995]. Alternatively, the notification of handoff initiation at the MH can be reported to the CH, which then may omit the congestion avoidance and provoke an even faster recovery [Manzoni etc 1995]. A similar scheme is advocated in [Swami etc 2005], which proposes a new 3-byte TCP option that allows an MH to inform its CH of the initiation of an IP handoff. The CH can then adjust its congestion control behaviours accordingly for rapid recovery.

In short, the transport layer (especially TCP) can adapt its behaviours to IP mobility with the help of L3 notifications and alleviate the impacts of mobility by restoring to its normal transport status quickly.

3.3.1.4. Application Layer

The application layer is expected to take care of the high-level mobility types such as personal and session mobility, and their possible interactions with the network-layer terminal mobility support. In addition, the application layer can enrich the capabilities of terminal mobility by adding advanced application-specific mobility functionalities such as renegotiation of session parameters like the codec for the ongoing multimedia session on a handoff. SIP, together with its associated protocols like SDP (Session Description Protocol), has the potential to fulfil these expectations.

Moreover, getting aware of the timely information from the lower layers, many multimedia applications could become adaptive to the changing system environments such as available resources by transforming themselves automatically, e.g., adjusting the sending rate. Therefore, the live session dropping rate could be reduced during handoffs

from the current system or bearer to another one with fewer resources. In addition to adaptation, some applications e.g., certain location-based services, may also entail mobility awareness in the application layer.

Finally, we consider the users' contributions. Through the application layer, a user may provide useful input to the protocol stack to help mobility adaptation. In turn, the user may gain benefits in terms of improved user-perceived QoS, extended battery usage, etc. For instance, when a moving user sees that he/she is approaching a tunnel, he can reasonably predict a short outage. Then the user may indicate the terminal of this event so that appropriate adaptations can be initiated, e.g., the running applications can hold data delivery to lower layers, the transport layer can hold the states, and all the involved layers can buffer the outgoing packets to avoid packet loss. Another example of user-assisted mobility support is discussed in [Li etc 1997], which suggests that the user should participate in handoff support to reduce call-dropping rate and improve resource utilisation. The user is expected to declare the requirement of mobility support at call setup time. When a handoff cannot be supported, the user can be informed in advance so that he or she can decide whether to control movement since a user may sacrifice mobility for maintaining communication in progress.

3.3.1.5. Summary of Protocol Layers' Contributions to Mobility Support

To sum up, Table 3.4 lists the major contributions from each protocol layer to a comprehensive and advanced mobility support envisioned in the next-generation (B3G) wireless systems. Clearly, such a demanding task calls for the participation and coordination of multiple, if not all, the layers, which can only be enabled by a proper cross-layer design.

Table 3.4 Contributions to mobility support from protocol layers

Protocol Layer	Contributions
Physical and Link layers	L2 triggers
Network layer	Basic terminal mobility Additional low-level mobility such as network mobility IP mobility indication IP-based QoS support IP-based AAA support
Transport layer	Adjust transport behaviours to IP mobility
Application layer	Advanced features added to terminal mobility High-level mobility: personal mobility, session mobility etc. Applications' adaptation to mobility User's input to initiate or adapt to mobility

3.3.2. The Envisioned Multi-Layer Mobility Support Framework

The proposed mobility support framework is outlined in Figure 3.3 (not all the interactions are shown), enabled by a CLASS-based combination of the cross-layer signalling methods. Considering all the layers' contributions, we have identified the following interactions between layers.

3.3.2.1. CLASS-Based Interactions

In the proposed framework, CLASS can be used to achieve active bi-directional messaging across the protocol stack. The following inter-layer interactions are identified and illustrated in Figure 3.3.

Interactions between the network layer and the application layer for coordinated mobility management: CLASS is used for the direct coordination between these two layers without bothering the transport layer. There are intrinsic connections between network-layer and application-layer mobility, especially between MIP and SIP terminal mobility. Thus, the two layers should perform in a cooperative way to improve the mobility management efficiency. In the case of SIP and MIP mobility protocols running in the two layers, respectively, at the same time, reduced overheads over the wireless and wired links could be achieved by coordinating the two protocols. For instance, to obtain a new IP

address on a handoff, if without coordination, both SIP and MIP would turn to a certain network service, e.g., a DHCP ([RFC2131] for DHCPv4, [RFC3315] for DHCPv6) server or an FA/AR. Since it is more convenient for MIP to deal with this network-layer issue, we can configure MIP to communicate with the DHCP server (or FA/AR) only. Anyway, SIP has difficulties to detect the change of the IP address even if it is allowed to contact the DHCP server itself. SIP could use polling to detect an IP address change. However, polling is not optimal for this time-sensitive event since the polling interval is typically several seconds [Schulzrinne and Wedlund 2000], and polling at a higher frequency would invoke considerable internal overheads. Thus, an active notification is desired only when this event actually happened. CLASS is the right solution to this problem since it can send this event from the network layer (MIP host) to the application layer (SIP User Agent or UA) in a timely and efficient way. Other mobility-related events or interactions can be delivered or exchanged similarly whenever necessary to coordinate the two layers.

Interactions between the physical or link layer and the network layer for improved handoff performances: The link-layer handoff notifications and extra system-specific information to the network layer can accelerate the L3 handoffs in the case of MIP over 802.11b WLAN (e.g., [Fikouras etc 2001]). Similar mechanisms could be exploited for other 802 or non-802 access networks where such L2 triggers are attainable [Aust etc 2003]. In 3G and beyond systems, a rich set of radio parameters measured by the physical and link layers are available (e.g., [3GPP TS25.215]) and selected parameters from these measurements can be exploited for system-specific handoff optimisations. Moreover, particular L2 handoff mechanisms enabled by specific MAC techniques could also benefit L3 handoffs. For instance, the CDMA cellular systems support L2 soft handoffs that could lead to seamless handoffs in the network layer. In addition to improving the performance of intra-system handoffs, link-awareness can also help to smooth an inter-system handoff

[Bernaschi and Cacace 2004]. To handle all the generic L2 triggers and additional measurements in a more organised way, a unified module (e.g., called L2 trigger manager) may be needed to collect, update and sort all the L2 triggers from other L1 or L2 entities (protocols), report L2 triggers to upper layers (e.g., the MIP host), and delete out-of-date or invalid triggers. The upper layers can thus only need to work with this L2 trigger manager.

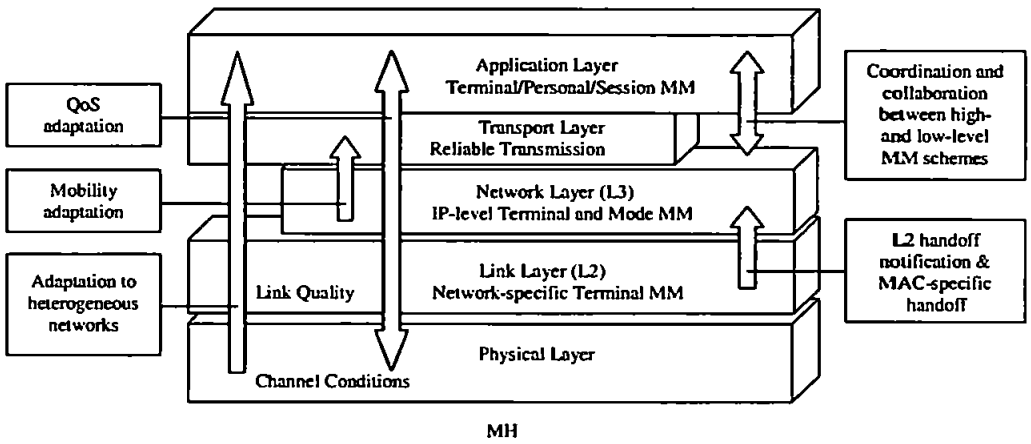


Figure 3.3 CLASS-based multi-layer mobility support architecture

Interactions between the low layers and the transport or application layer for mobility and QoS adaptations: The proposed multi-layer mobility framework also facilitates the QoS adaptation of applications and transport protocols to contexts such as mobility, heterogeneous networks and the time-varying radio channel conditions. The utilisation of the mentioned L2 triggers and additional 3G measurements can be utilised for mobility and QoS adaptations, possibly with other parameters abstracted from related layers. Meantime, different QoS requirements from different applications (or from the user) could be mapped into controllable or informational parameters of corresponding layers. All the parameters can be coded to CLASS messages for interactions across the protocol stack. Particularly, notifications of the start and the end of an L3 handoff are generated from the network layer and are sent to the upper layers. These messages are targeted to initiate the adaptations of applications and transport protocols especially TCP to mobility. The applications and

transport protocols can then adjust its behaviours and attempt to recover from the impacts of the mobility. Through these adaptations (depending on algorithms and transport protocol semantics), the end-to-end application performances during mobility could be significantly improved. More discussions on QoS adaptations are given in Section 3.3.2.3.

3.3.2.2. Potential Interactions Based on Other Methods

In some contexts, other methods can also be introduced. For example, if multimedia traffic is processed on a packet-by-packet basis, extended headers can be applied to carry extra traffic-specific information including QoS requirements from the top of stack to the bottom (Method 1). This "pipe" is acceptable since a data packet has to travel through all the layers anyway. A possible application of local profiles (Method 3) could be for information related to location updates. For necessary messaging between an MH and its access networks, new network services can be considered (Method 4). For example, the existing link-layer hints [Seshan 1995] are neither reliable nor widely available, but can be enhanced and enriched through a link-layer agent in the network side. This agent is not necessarily as complex as the dedicated WCI server is. It can be collocated in border base stations or dedicated mobility servers such as the 3G Gateway Location Registers [3GPP TS 23.119] located between a WLAN and a cellular network. It can monitor and provide overall physical- and link-layer information. An approaching MH can then be informed of the new system characteristics and capabilities in a heterogeneous environment, and thus it can prepare for an inter-system handoff. Within a system, the MH itself can observe and adapt to its contexts via the CLASS-based internal cross-layer interactions. The incurred overheads in the network can thus be minimised.

3.3.2.3. Considerations on QoS Adaptation

In this subsection, we present a generic QoS-adaptive protocol stack, where cross-layer signalling is intensively applied to enable application-centric adaptations. The QoS

adaptation design is mainly motivated by the analysis of [RFC3819] presented in Section 3.2.4.3. All direct cross-layer information exchanges between non-adjacent layers are enabled by CLASS, though a hybrid use of the discussed methods is designed to demonstrate their application scenarios.

In the initialisation, an application signals its QoS tolerance to the underlying layers. The parameters can include bit error rate, packet loss ratio, delay and jitter bounds, mobility preference (mobility-aware or mobility-transparent), etc. Notably, these parameters can be either quantitative or qualitative, though qualitative parameters may facilitate the interpretation and mapping at the lower layers. Per-packet-level application adaptation requirements can be carried in headers (Method 1).

In response to this signalling, the involved lower layers map selected parameters to controllable metrics wherever appropriate, and takes predefined actions to fulfil the QoS requirements. In this case, the link-layer ARQ and the transport layer TCP collaborate to achieve trade-offs between packet error/loss control and delay/jitter control, e.g., by adjusting the number of retransmissions. For UDP (or UDP-Lite [RFC3828]) applications, the interactions between the L2 ARQ and the L4 Datagram Congestion Control Protocol (DCCP) [Kohler etc 2005] may also lead to a similar trade-off between reliability and timeliness. For external IP QoS signalling, e.g., using RSVP (or NSIS in the future), the link layer reports the offered QoS (trade-off results) to the network layer. Preferably, the reports are directly understandable to the IP QoS protocol, e.g., in terms of the TSpec/RSpec model if the IntServ model is implemented [RFC3819].

Furthermore, context-aware proactive applications can benefit from selected L1 and L2 measurement reports on channel state information, such as SIR and RSS, widely available in wireless systems. Based on such information, e.g., a video codec can dynamically change the compression degree and thus modify the transmission rate to

maximise the picture quality [Haratcherev etc 2005]. Whilst CLASS (Method 5) or ICMP selected holes (Method 2) can be used for real-time reports, history records can be stored in local profiles (Method 4) and be made available to any interested layers. Moreover, additional information may be obtained from third-party servers, like the WCI server (Method 3).

Finally, it is noted that this protocol stack could be extended to incorporate the management of radio resource, energy, security and even more in an integrated or coordinated way. For one thing, the QoS adaptation should be achieved under the resource and energy constraints. For another, the security mechanisms may impose constrains on the implementation of a chosen cross-layer signalling method. For complex tasks like this, a coordinator module is needed to utilise the cross-layer contributions fully whilst avoiding any potential conflicts. A generic policy-based coordination framework is presented in [Gao etc 2004].

3.3.3. The Design Emphasis

In the previous section, we presented a comprehensive and generic mobility support framework from the cross-layer perspective of the IP-based protocol stack. The proposed framework is comprehensive since it covers the fundamental mobility management issues such as the basic terminal mobility and its enhancements, and advanced mobility support features such as personal mobility, QoS support etc. The framework is also generic as it has the potential to allow progressive development of the involved functionality, e.g., QoS adaptation could be coupled in a future stage. In the remainder of this thesis, we concentrate on the fundamental mobility management with selected advanced mobility support features, provided jointly by the network and the application layers. Specifically,

we centre our designs around integrated MIP-SIP macro-mobility architectures with optimised IP micro-mobility support.

3.4 Essential Building Blocks of the Proposed Framework

3.4.1. Introduction

In this section, referring to Figure 3.1 we further explain the essential building blocks to achieve the envisioned mobility support framework under our design emphasis considerations.

Note that for presentation and evaluation conveniences, in the subsequent chapters, terminal and personal mobility (together with other additional mobility types) are addressed in the context of macro-mobility architectures, whilst the interactions of mobility protocols with QoS signalling protocols are emphasised in the micro-mobility architecture though the macro-mobility and QoS interaction case is also discussed there. Thus, the following subsections in this section provide an overview of these designs from the perspectives of the macro- and micro-mobility architectures.

3.4.2. Macro-Mobility Support Architectures

3.4.2.1. Support for Macro Mobility

Macro mobility management is mainly concerned with mobility operations between an MH and its mobility servers at the home domain, and those between an MH and its CH(s), i.e., end-to-end mobility behaviours. We propose two architectures, where macro-mobility is jointly supported by MIP and SIP. In the tightly integrated MIP-SIP architecture (TL-MIP-SIP), the mobility-related home network entities of MIP and SIP are fully converged into a uniform mobility server with redundancies removed. This architecture is recommended for maximised cost-efficiency and performance improvements in the long term. Another alternative architecture is the loosely integrated

MIP-SIP architecture (LI-MIP-SIP), where interactions between MIP and SIP entities for common mobility procedures are introduced with the entities themselves almost intact (except minor enhancements). Despite a bit interior to the TI-MIP-SIP architecture in terms of costs and performances, the LI-MIP-SIP architecture also clearly outperforms existing hybrid MIP-SIP architectures yet at the same time offers a prompt deployment advantage. Both integrated architectures utilise standard-based MIP and SIP messages for signalling mobility-related operations among an MH, its home mobility server and its CH(s). The IPv6 networking environment is focused on though the IPv4 context is also discussed.

3.4.2.2. Support for Terminal Mobility

From the terminal mobility point of view, the architecture supports efficient and effective inter-domain location management and handoff management. Protocol signalling operations are designed for the TI-MIP-SIP and the LI-MIP-SIP architectures, respectively. Location management are proposed by dynamic use of selected MIP and SIP messages. In principle, end-to-end handoff management for TCP-based non-real-time applications and UDP-based real-time applications are supported by MIP and SIP, respectively, though both protocols are optimised and enhanced for improved performances.

3.4.2.3. Support for Personal Mobility

We consider two major capabilities in personal mobility: one is the capability for the network to locate a user through a user-level ID for setting up a session regardless the user's locations or terminal(s) being used; the other is the capability for a user to register more than one terminal whenever preferred. In the proposed architectures, the first capability of the personal mobility is embedded as an integral part of location management, i.e., the session setup procedure; and the second capability is achieved in both location

management and handoff management operations through either dynamic SIP-MIP signalling or extended MIP signalling only.

3.4.2.4. Support for Additional Mobility

Though the crucial terminal and personal mobility management is emphasised in the project, additional mobility types are taken into account in the integrated MIP-SIP architectures, which can actually support these additional mobility types thanks to its incorporation of the powerfulness of both MIP and SIP. Specifically, the support for session mobility and network mobility are discussed.

3.4.3. Micro-Mobility Support Architecture

3.4.3.1. Support for Micro Mobility

As a complementary component to the macro-mobility architectures, micro-mobility management is designed based on an optimised combination of HMIPv6 and FMIPv6 with influential enhancements. The proposed architecture combines the merits of both HMIPv6 and FMIPv6 whilst circumventing their shortcomings. It provokes the least expected total costs during a session's lifetime compared with two other combination variants, and achieves faster handoffs than the standard FMIPv6 does. The proposed micro-mobility architecture can harmonise either of the proposed macro-mobility architectures, TI-MIP-SIP or LI-MIP-SIP, and their interactions are addressed.

3.4.3.2. Support for QoS Management

IP QoS signalling protocols with mobility extensions are incorporated into the mobility support architectures for real-time applications. The RSVP over DiffServ model is based on for an end-to-end QoS management. The interworking of QoS management protocols with macro- and micro-mobility schemes is designed though the latter case is emphasised. In particularly, a trade-off of between two QoS signalling approaches is achieved simultaneously in the combined FMIPv6 and HMIPv6 architecture.

3.4.4. Evaluation Methodology

The proposals are evaluated numerically in terms of well-defined metrics, and compared with existing and/or alternative approaches wherever appropriate. The evaluation methodology is a combination of theoretical analyses and software simulations to validate and/or complement each other. The analyses are built upon well-established analytical models with necessary enhancements to carter to our evaluations. The simulations are developed and performed with OPNET® Modeller® 11.0 [OPNET] or Microsoft® Visual C++ 7.0. C++ is used to obtain signalling costs (loads) in Chapter 4 and Chapter 6, whilst in Chapter 5 OPNET is used to evaluate delay-sensitive metrics such as handoff delays, which require a more accurate network setting. The operating system is Microsoft® Windows® XP Professional and the computer running the simulations is equipped with a Pentium IV 2.80-GHz CPU and 496-MB RAM. The major evaluation metrics including signalling costs and handoff performance in terms of handoff delay, handoff packet loss etc. More details on analysis and simulation configuration, and metric definitions are provided in the performance evaluation sections of the subsequent chapters.

3.5 Summary

In this chapter, we envisioned a distinct multi-layer framework for comprehensive and advanced mobility support through the cross-layer design approach. Firstly, we identified the next-generation (B3G) mobility support requirements and the design challenges, which motivated us to switch from the conventional single-layer design approach to a more powerful cross-layer design methodology. Next, we analysed the pros and cons of the existing cross-layer signalling methods and proposed a new generic, efficient and flexible scheme called CLASS, which appears to be the most promising candidate supporting scheme for the envisioned multi-layer mobility support framework though other methods

can be used alternatively or jointly. Subsequently, the contributions from each protocol layer to mobility support are investigated and the potential cross-layer interactions are specified in the framework. Finally, we narrowed down the framework to specific architectures and protocols, which are the design focuses in the remaining of the thesis, as crucial building blocks to achieve the framework. The details of the proposed architectures and protocols are expounded from the next chapter.

Chapter 4

The Tightly Integrated MIP-SIP Architecture for Macro-Mobility Support

In this chapter, we propose and evaluate a tightly integrated MIP-SIP architecture, referred to as TI-MIP-SIP, for macro-mobility support. This chapter is partially based on three publications [Wang and Abu-Rgheff IJCS, 3G2003, ICC04].

4.1 Introduction

As discussed in Section 2.6.4 in Chapter 2, due to the complementary functionality in mobility support, the joint MIP-SIP approach has gained growing importance. In typical hybrid MIP-SIP architectures [Politis etc 2004, Wong etc 2003], MIP (or its variant) and SIP are exploited for TCP and UDP mobility to achieve effective non-real-time and real-time application support, respectively. Nevertheless, these hybrid architectures tend to incur excessive overheads that may seriously degrade the performance of the system mainly because MIP and SIP operate in a rather independent way and little joint optimisation has been applied. Therefore, a better solution is entailed for improved system efficiency.

The remaining of this chapter is organised as follows. In Section 4.2, we describe the building blocks in the proposed TI-MIP-SIP architecture. Then in Section 4.3, we present

the protocol signalling design in the contexts of SIP integration with MIPv4 and MIPv6 (the resultant protocols are referred to as TI-MIPv4-SIP and TI-MIPv6-SIP, respectively). Section 4.4 reflects our considerations on the support of various mobility types. Subsequently, we evaluate the performances of the proposed protocols by theoretical analyses and simulations in Section 4.5. Finally, concluding remarks are provided in Section 4.6.

4.2 Architectural Design of the Tightly Integrated MIP-SIP Architecture

In contrast to the hybrid MIP-SIP architectures, we propose the integration approach for efficient macro-mobility management, applicable to both IPv4 (MIPv4) and IPv6 (MIPv6). The underlying principle is to introduce coordination into the hybrid MIP-SIP context for optimised system performances. Depending on the degree of the coordination an operator may prefer, two approaches can be adopted to integrate MIP and SIP. In the rest of this chapter, we focus on the first approach and the proposed Tightly Integrated MIP-SIP Architecture (TI-MIP-SIP), whilst the other approach and the corresponding Loosely Integrated MIP-SIP Architecture (LI-MIP-SIP) are addressed in Chapter 5.

In this section, we expound the design of the building blocks in the proposed TI-MIP-SIP architecture. Section 4.2.1 presents an architecture overview. The functional elements of the integrated mobility servers are identified and their interactions for proper operation are described in Sections 4.2.2 and 4.2.3, respectively. Finally, Section 4.2.4 provides the design of a uniform address management.

4.2.1. Architecture Overview

Considering the overlapping functionalities of MIP and SIP mobility management, we propose unified network architecture, on which our integrated mobility support is based, as shown in Figure 4.1.

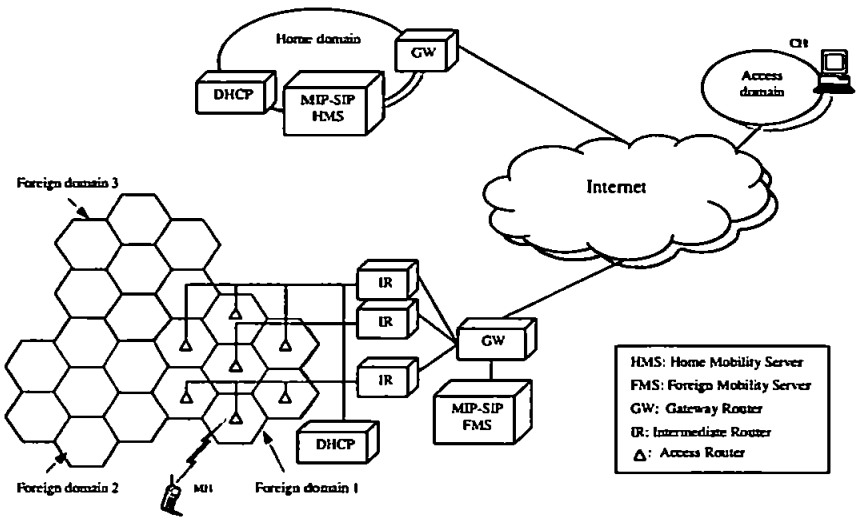


Figure 4.1 TI-MIP-SIP: network model

In the home domain of an MH, the MIP HA and the SIP HS are optimally merged to form a new MIP-SIP home mobility server (HMS). Similarly, in a foreign domain the SIP FS is integrated with a MIP-based domain micro-mobility server, whose specific type depends on the micro-mobility protocol (HMIPv6 or MIPv4-RR) selected. Consequently, a MIP-SIP foreign mobility server (FMS) is produced and preferably collocated with the GW. (More design details of the mobility servers are presented in Section 4.2.2.) Only when it moves between access routers belonging to two different domains, an MH needs to perform home registrations at the HMS. When moving between subnets within a foreign domain, it merely reports its new locations to the FMS and a proposed micro-mobility protocol (addressed in Chapter 6) is in charge. To simplify new address distribution, we adopt a unified mechanism like the MMM. In an IPv4 networking environment, DHCPv4

[Droms 1997] is utilised; in IPv6, an MH turns to either a DHCPv6 [Droms etc 2003] server or a MIPv6 access router for stateful or stateless auto-configuration, respectively. In this architecture, both TCP and UDP mobility is supported and their data flows are separated at an MH.

In this unified network platform, we reuse selected MIP and SIP messages for mobility signalling, which are handled efficiently through the integrated mobility servers. Overall, the architecture is designed to minimise the functionality redundancy, the signalling duplication and the corresponding processing repetition. Therefore, we expect that the overall costs can be substantially reduced compared with the hybrid approach.

4.2.2. Mobility Server Integration

4.2.2.1. Methodology of Integrating Mobility Servers

Surely, to deal with both MIP and SIP signalling and data, all the functionalities of both architectures should be included whilst optimisation entails that similar entities are integrated, rather than simply collocated. Note that a simple collocation of the MIP and SIP mobility servers does not solve the undesirable redundancy problems found in the typical hybrid MIP-SIP architectures. For instance, Jung etc proposed to collocate (rather than optimally integrate as in TI-MIP-SIP) MIP HA and SIP HR in [Jung etc 2003]. However, the superfluous MIP and SIP mobility signalling and processing costs are similarly provoked as those in the EVOLUTE architecture.The methodology for our optimisation and integration is as follows. Firstly, we decompose similar MIP and SIP entities to independent functional elements; secondly, we integrate the similar elements, and retain the distinguished ones intact or with necessary enhancements; and finally we establish interactions among these elements. Applying this methodology to home or foreign MIP and SIP entities, we can create the desired HMS and FMS, respectively.

4.2.2.2. Integration of Home Mobility Servers

First, we consider a MIP HA and a SIP HS to obtain an integrated MIP-SIP HMS. According to its dominant functions, a MIP HA is decoupled to a MIP home registrar (HR) and a Tunnelling Agent. Roughly speaking, a MIP HR deals with location-related MIP signalling and serves as the entry point for MIP (TCP) data routing, and a Tunnelling Agent encapsulates and forwards incoming data to the CoA of an MH in a foreign domain. (The interactions of the functional elements are provided in Section 4.2.3.) Consequently, the SIP home registrar (HR) is merged with the MIP HR to handle both SIP and MIP registrations and other location services, and we call the new entity MIP-SIP Home Registrar. This new entity is featured by a unified binding list with the merger of the MIP built-in location database and the SIP associated location database for uniform address management (details in Section 4.2.4). So far, we have produced two new functional elements, a MIP-SIP Home Registrar and a Tunnelling Agent, which are indeed the core parts of an HMS and should be tightly integrated as a whole. The remaining composite of the SIP HS is the home SIP proxy (or redirect) sever. Since its functionality is unique and specific to SIP sessions, we keep it intact. In addition, as far as efficient AAA is concerned, we also propose to incorporate a home AAA server (AAAH) into the HMS. This AAAH is expected to provide both MIP and SIP AAA services, though its design is beyond the scope of the thesis. It is worth noting that the SIP home proxy (or redirect) server and the AAAH are both logical entities and thus can be physically collocated with the MIP-SIP Registrar and the Tunnelling Agent to yield an HMS. Alternatively, the MIP-SIP Registrar and the Tunnelling Agent themselves can be tightly coupled to comprise an HMS, and the other two servers can exist as stand-alone servers and interact with this kind of HMS. We assume the former case for presentation purpose.

4.2.2.3. Integration of Foreign Mobility Servers

The construction of an FMS follows the same methodology and results in a similar structure with the following differences. First, the local AAA server (AAAL) replaces the role of AAAH. Second, the SIP FS is integrated with a MIP-based domain micro-mobility server, depending on the specific micro-mobility protocol. For IPv4 (MIPv4) networking we recommend MIPv4-RR (MIPv4 Regional Registration [Gustafsson etc 2004]) whilst for IPv6 (MIPv6) we propose to use HMIPv6 (Hierarchical MIPv6 [Soliman etc 2004]). Accordingly, the micro-mobility server can be a MIPv4-RR Gateway FA (GFA) or a HMIPv6 domain Mobility Anchor Point (MAP). Both protocols are surveyed in Chapter 2, though we propose an enhanced and optimised design in Chapter 6.

4.2.3. Mobility Server Operation

Now that the functional elements comprising a mobility server are identified, we define the interfaces among them by describing the mobility server operation to fulfil the desired mobility management tasks. We focus on the operation of an HMS as illustrated in Figure 4.2 whereas leave that of an FMS to Chapter 6 since the major role of an FMS is micro-mobility support. In Figure 4.2, the letters A, C, C1, D and E indicate pairs of request-reply messages whilst B and B1 to B3 designate the flows of MIP (TCP) data.

Among the functional elements in an HMS, the MIP-SIP Home Registrar (with the built-in uniform address binding list) is the focal point to process the location-related MIP and SIP signalling, basically in a client-server way. The involved operations include home registrations (or refreshes) from an MH (A), location queries from the SIP home proxy or redirect server (C1), and possibly binding requests from a CH with MIPv4-RO adopted or MIPv6 extended to enable such operations for MIP session setup (D, referred to as the SS option and discussed in Section 4.3.1). Moreover, it delivers incoming MIP data packets (B)

to proper destinations (B1, B2), together with the MIP Tunnelling Agent (B3). For all these operations, the MIP-SIP Home Registrar may interact with the AAAH for AAA purposes (the conception is shown by E). Since AAA procedures are strongly dependent on specific implementations, we omit their operations for clarity in the following discussions. Note that the MIP-related signalling messages involved depend on the versions of MIP, and are specified in the protocol design (Section 3). Figure 4.2 demonstrates the MIPv6 context, where BU, BA, and BRR stand for Binding Update, Binding Acknowledgement, and Binding Refresh Request, respectively. In addition, the interfacing between the Home Registrar and the Tunnelling Agent (B2) is logical and does not need explicit messaging.

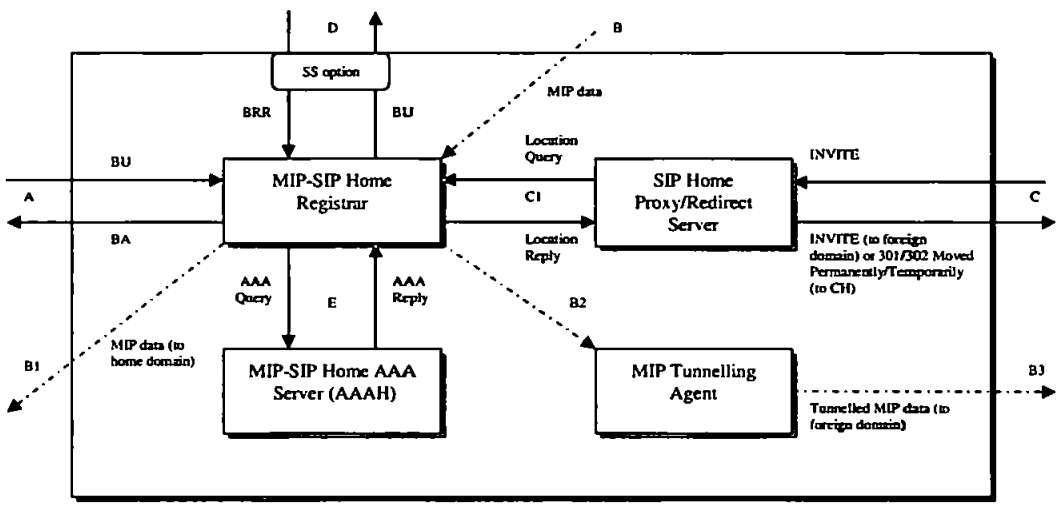


Figure 4.2 TI-MIP-SIP: home mobility server operation

4.2.4. Uniform Address Management

The address management functionality of MIP and SIP is also integrated to reflect the complete addresses related to a user. A user is globally identified with a SIP URI called AOR (Address of Record) and can register with one or more terminals (end hosts). Each terminal obtains its own MIP HoA and a MIP CoA (or SIP contact IP address) in the home and a foreign domain, respectively. As in the MMM architecture, Co-located CoAs are applied since a CoA is also used as a SIP contact IP address. Since HoAs and AOR are

semi-permanent, only CoA changes are reported to the MIP-SIP HMS through home registration. In addition, AAA servers request that a MIP user include an NAI (Network Access Identifier) [RFC2486] as a unique ID for registrations ([RFC2794] for MIPv4, [Patel etc 2004] for MIPv6). Furthermore, we propose to utilise the NAI as an alternative user ID so that the MIP and SIP location update can be unified through a mapping of the SIP AOR and the MIP NAI of the same user. Since the NAI and SIP AOR share a similar format, either a mapping between them or a merger of them may be applied. Table 4.1 exemplifies a generic record of an address-binding list for a mobile user who registers with two terminals in an HMS (or more precisely, in a MIP-SIP Home Registrar). The binding list managed in an FMS can be similarly constructed though the local address of an MH in the visiting domain should be added for micro-mobility support (discussed in Chapter 6). This uniform address management reduces the system costs for managing two separate address databases found in MIP and SIP, respectively.

Table 4.1 TI-MIP-SIP: a record of the binding list in an HMS

NAI	*AOR*	*Terminal ID*	*Terminal Current Address*	*Remaining Registration Lifetime*
MIP NAI	SIP AOR	MIP HoA1	IP address 1	T1
		MIP HoA2	IP address 2	T2

Furthermore, the MIP binding list in an end host is also enhanced with SIP AOR so that the diverse addresses can be managed efficiently and effectively on a uniform platform. In addition to regular refreshes, this enhanced binding list is updated whenever the end host gets aware of a location change to its correspondents through an operation involved in either MIP or SIP sessions. A binding update from an MH (or the HMS) is surely an example of such an operation. Moreover, we propose that the CH updates the binding list after a SIP session setup where an MH is involved. Consequently, the uniform address management benefits both MIP and SIP from the end host point of view. To setup a SIP

session, a CH can now enquire its binding list before sending a session invitation to the HMS of the targeted MH, and thus the triangular SIP session-setup signalling between the CH and the HMS can be avoided as long as the binding of the MH is still valid. For MIP sessions, a CH can make use of the location enquiry results from the SIP session setup last time and thus the probability that the CH has a valid binding list increases before it sends TCP packets to an MH so that the probability of triangular routing is reduced. In either scenario, the system overheads for signalling or routing can be decreased.

4.3 Protocol Signalling Design of the Tightly Integrated MIP-SIP Architecture

This section specifies the protocol signalling in the proposed TI-MIP-SIP architecture. Both location and handoff management procedures are proposed in the contexts of IPv4 (MIPv4) and IPv6 (MIPv6) by reusing standard-based MIP and SIP messages. Correspondingly, wherever appropriate the protocols are referred to as TI-MIPv4-SIP and TI-MIPv6-SIP, respectively.

In contrast to the redundant mobility routines in Hybrid MIP-SIP architectures as shown in Figure 2.19, mobility procedures in the proposed architecture are integrated to minimise the signalling and processing loads. Figure 4.3 illustrates this design concept whilst detailed signalling design is presented in the following subsections. Notably, we do not introduce new messages to achieve these integrated mobility procedures. Instead, MIP and SIP messages are reused to utilise the standard protocols fully, and extensions are minimum and well justified. This design approach should facilitate implementations. For MIP and SIP messages of similar functionality, MIP messages are reused since the MIP message sizes are much smaller than their SIP counterparts are. For simplicity, we assume the CHs are static in the subsequent signalling diagrams, though they can be mobile. We

focus on macro IP mobility in the rest of this chapter and leave the micro IP mobility to Chapter 6. It is important to note that only the HMS is indispensable to the basic operation of macro-mobility management. When no FMS is deployed in a foreign domain, the protocol would work as a stand-alone macro-mobility scheme, analogous to MIP operation without a micro-mobility scheme. Unless stated otherwise, in the protocol design, an HMS is discussed as a whole entity as the internal interfaces and operations of an HMS have been defined in Section 4.2.3; in addition, the FMS is omitted for brevity.

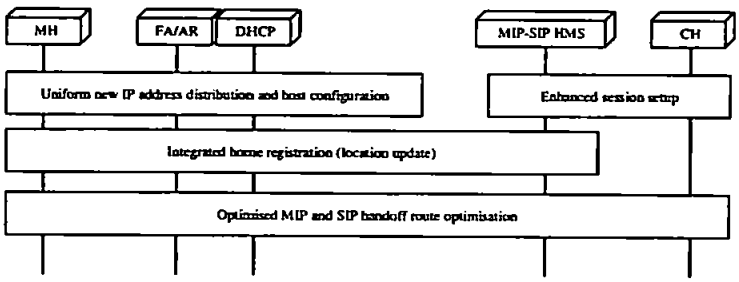


Figure 4.3 TI-MIP-SIP: mobility signalling block diagram

4.3.1. Location Management

The location management discussed here includes the location update procedure at the HMS (i.e., home registration) and the session setup procedure.

4.3.1.1. Home Registration

The home registration is further subdivided into initial home registration and home re-registration from a foreign domain, as illustrated in Figure 4.4 and Figure 4.5, respectively.

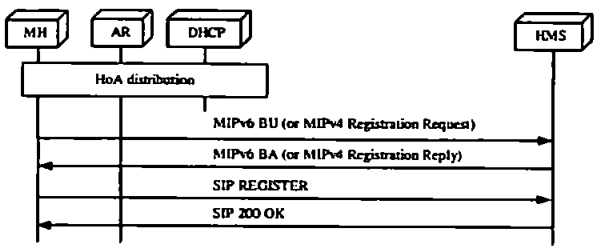


Figure 4.4 TI-MIP-SIP: initial home registration

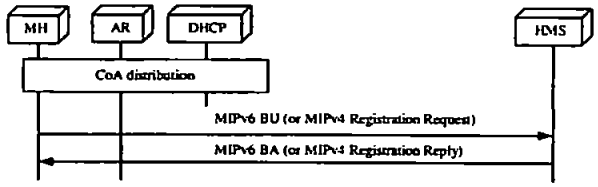


Figure 4.5 TI-MIP-SIP: home re-registration from a foreign domain (basic mode)

To obtain future mobility support, after acquiring an HoA an MH performs an initial home registration, usually taking place in its home domain. We propose to apply both MIP and SIP registration messages to create a unified MIP-SIP record for the MH in the uniform binding list maintained by the HMS. In the new record, both the HoA and the current terminal address are set to be the HoA and the mapping between MIP NAI and SIP AOR is also established. Alternatively, either MIP or SIP registration messages could be extended to fulfill this initial registration. However, modifications such as new fields then have to be introduced to the standard messages, and thus this approach is not recommended. Note that this redundancy only happens for the initial home registration, which happens rarely because of the semi-permanency of an HoA or AOR. Subsequent home registrations or refreshes just use MIP registration messages, in contrast to the parallel use of both MIP and SIP registration messages in the hybrid MIP-SIP architectures all the time.

On the other hand, advanced registrations may entail the dynamic use of SIP REGISTER sometimes. Notably, although most users in cellular networks only carry and register with one cell phone through the above basic mode, a user may occasionally register with more than one terminal, e.g., an additional local terminal in the visited domain for communication convenience. Though MIPv6 BU or MIPv4 Registration Request could be extended to accommodate such information, we recommend using the SIP REGISTER message, which has been designed to carry a list of contact addresses in its Contact header field with priorities set in the 'q' parameter. Thus, in our architecture the

default registration messages are MIP registration messages. In case of multiple-address registration, SIP REGISTER and its 200 OK are issued. In addition, MIP registration messages can be utilised for the optional explicit de-registration with the previous FMS (PFMS) via the new FMS (NFMS), when FMSs are present in the involved domains. Figure 4.6 illustrates such an advanced mode for registration operations including new IP address distribution (and other host configuration), advanced home registration (the dotted lines indicate the dynamic use of SIP REGISTER), and optional explicit de-registration with the previous foreign domain. For brevity, only the IPv6 scenario is shown. The DHCPv6 Rapid Commit mode is demonstrated for host configuration including the new CoA distribution.

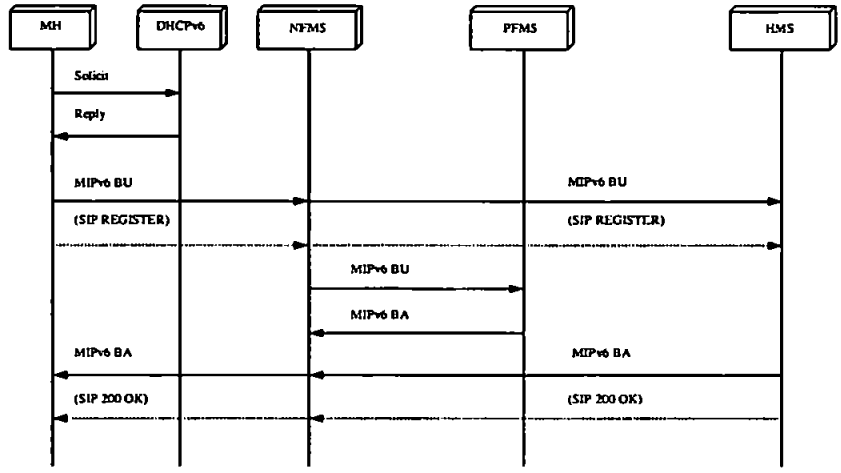


Figure 4.6 TI-MIP-SIP: home re-registration from a foreign domain (advanced mode)

4.3.1.2. Session Setup

Next, we look at the session setup procedure illustrated in Figure 4.7. Adhering to the MIP and SIP standards, the basic signalling is similar to that in the hybrid architecture as shown in except for the merger of the MIP HA and the SIP HS to a MIP-SIP HMS. Additionally, to reduce MIP triangular routing the architecture also supports a session setup option (SS option) for MIP sessions. When the SS option is adopted, a CH can

enquire for the up-to-date binding of the targeted MH at the HMS before sending any TCP data to the MH's HoA if it does not have a valid binding. Although this optional process is not defined in the base MIPv4 or MIPv6, it can be achieved by a pair of MIPv4-RO or MIPv6 messages that are well defined. As far as privacy is considered, on receiving such an enquiry the HMS may choose not to disseminate the MH's current binding to the CH, based on a pre-defined privacy rule. An example simple rule is proposed in MIPv4-RO: the MH may set the proposed private bit in the Registration Request message to indicate that it would like the HMS to keep the binding private. In base MIPv6, an MH itself flexibly determines whether to reveal its current binding to a CH in the route optimisation process. By combining both rules, a flexible trade-off between privacy and routing efficiency could be achieved.

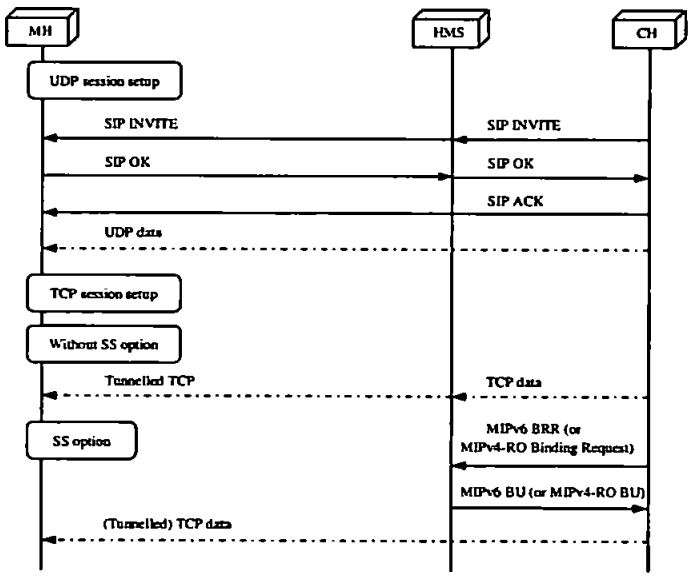


Figure 4.7 TI-MIP-SIP: session setup

4.3.2. Handoff Management

Unlike the location management, the handoff management procedures are significantly different in MIPv6 and MIPv4 as described in Chapter 2, and so are the

resultant integrated handoff procedures in the IPv6 and IPv4 version of TI-MIP-SIP. Thus, we propose different integrated handoff signalling for TI-MIPv6-SIP and TI-MIPv4-SIP.

4.3.2.1. Handoff in TI-MIPv6-SIP

In TI-MIPv6-SIP, for TCP mobility, an MH performs the MIPv6 end-to-end route optimisation by sending a BU directly to the CH after completing the Return Routability (RR) process for authorisation purpose as defined in the standard [RFC3775]. Before the RR process, the HMS should have received and authorised the new binding through the home registration process initiated by the MH so that the home test involved in the RR process can be carried out. In addition, an enhanced RO (ERO) option, inspired by the RO option defined in MIPv4-RO, is proposed here. With this ERO option, the MH indicates the HMS to inform the CH of its new CoA by enclosing the address of the CH in the BU message sent to the HMS. Note that only when the HMS and the CH has established a security association (SA) before, the CH can accept this binding update from the HMS on behalf of the MH. When this process is successful, the CH should send a BA to the MH directly, and the MH can then skip the remaining RR test and the subsequent CH binding.

For UDP mobility, an MH applies SIP messages and reuses the MIPv6 home registration, similar to the operations in the TI-MIPv4-SIP context. Whether to reuse the RR process depends on the AAA implementation. In the TI-MIPv4-SIP and the hybrid MIPv4-SIP architectures, the end-to-end SIP binding update at a CH is authorised implicitly by an assuming that an AAA mechanism is in place. For instance, the authorisation keys may be pre-established in the session setup stage that is mandatory for a SIP session. Thus, for comparison purpose we have followed this assumption in the design of the TI-MIPv4-SIP handoff signalling. However, since this authorisation is explicitly defined in the MIPv6 standard without pre-configuration for authorisation assumed, we recommend that the RR process be reused in supporting the UDP mobility. Particularly,

when both TCP and UDP applications are running with a same CH upon a handoff, the RR process is naturally shared by both MIPv6 and SIP mobility. In this simultaneous TCP and UDP mobility case, the CH binding processes in MIPv6 and SIP may be unified to the SIP messages. When the SIP re-INVITE message arrives at the CH, the binding list cached in the CH is updated so that the ongoing TCP sessions can be redirected to the MH's new location. The SIP messages are chosen since they normally enclose session-specific renegotiations in addition to the binding update. However, for implementation simplicity, both MIPv6 and SIP CH binding messages are recommended in parallel use in this rather rare scenario. In addition, this redundancy is unlikely to cause significant overheads thanks to the compactness of MIP messages. Figure 4.8 depicts the above handoff operations, referred to as the basic mode.

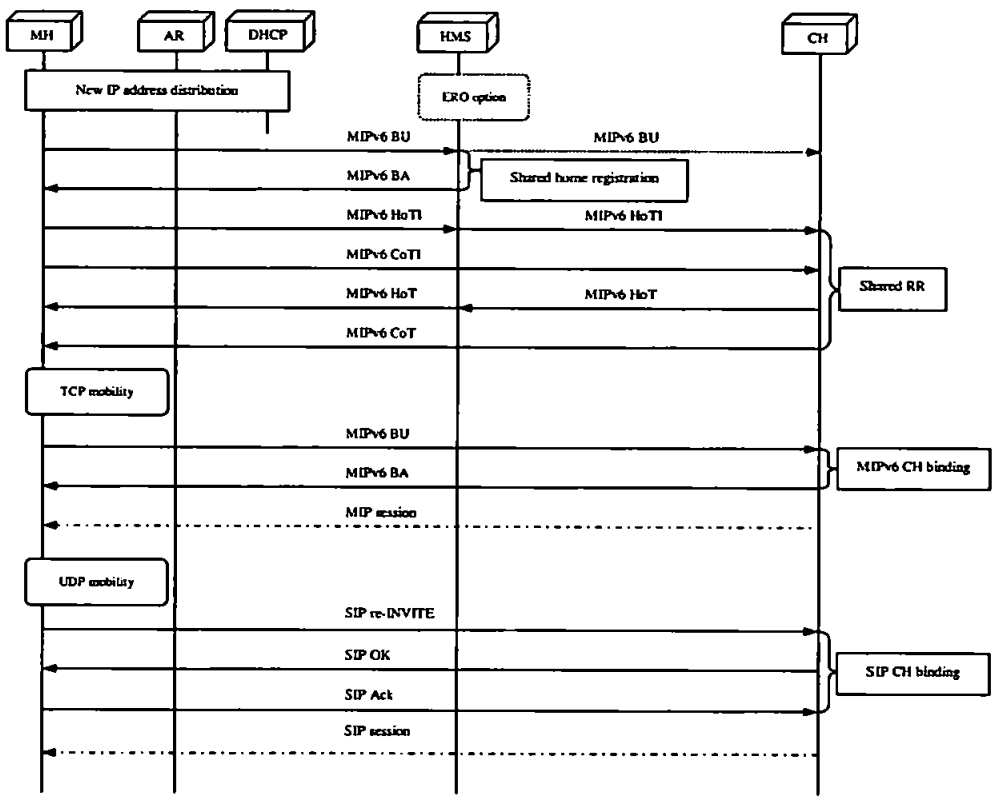


Figure 4.8 TI-MIPv6-SIP: handoff (basic mode)

Furthermore, we consider the advanced mode for macro handoff as illustrated in Figure 4.9.

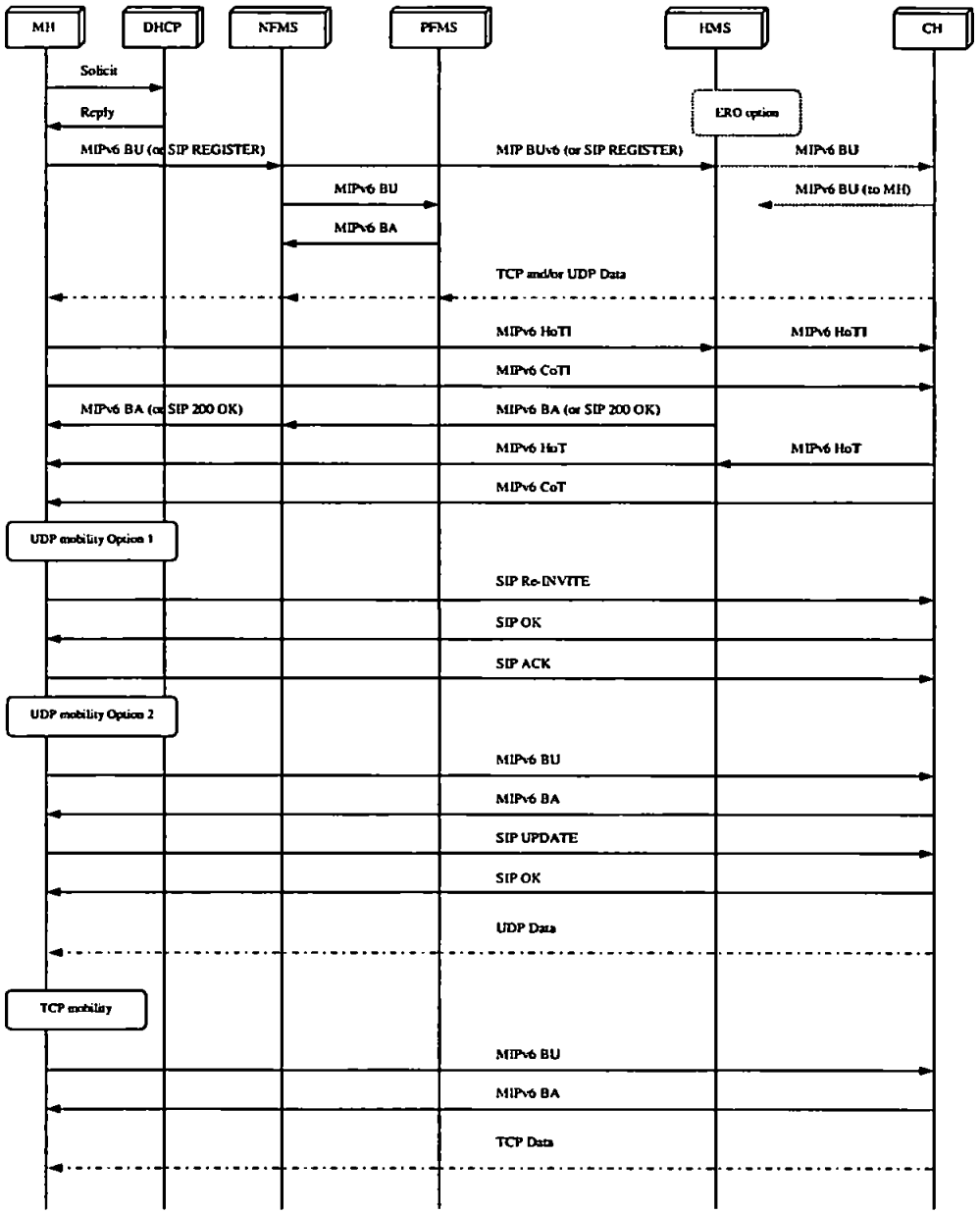


Figure 4.9 TI-MIPv6-SIP: handoff (advanced mode)

This advanced mode of IPv6 handoff procedure comprises host configuration, home registration, smooth handoff, RR tests, and CH binding update. The host configuration and home registration processes are the same as those in the advanced mode of the location

update procedure, and the RR tests are the same as aforementioned. The CH binding update for UDP sessions is enhanced. A SIP Re-INVITE message can be used to renegotiate SIP sessions (UDP Option 1) as in the basic mode. Alternatively, a MIP Binding Update can be applied if the renegotiation is unnecessary, and in case of a change in session parameters such as the codec a SIP UPDATE message is sent to update the change (UDP Option 2). Meanwhile, the NFMS may have updated the bindings in the PFMS for smooth handoffs initiated during the home registration process. The PFMS then tunnels the in-flight packets to the NFMS, which again de-tunnels the packets first and re-tunnels them to avoid dual encapsulations. Finally, these packets arrive at the MH. The MIPv6 BU and BA messages can be employed to enable such a smooth handoff.

At last, it is worth noting that the operations for the TCP mobility can be omitted in either the basic or the advanced mode when a handoff involves both TCP and UDP sessions from the same CH. The reasons for this recommendation are as follows. First, the CH can learn the new CoA of the MH from the UDP mobility signalling. Second, the UDP mobility signalling also encloses application-specific information that is not available in the TCP mobility messages.

4.3.2.2. Handoff in TI-MIPv4-SIP

In TI-MIPv4-SIP, for TCP mobility, an MH just performs a MIPv4 home registration and the subsequent TCP data would then be tunnelled to its new location by the HMS. To handle the triangular routing, a similar ERO option as that proposed in the TI-MIPv6-SIP is proposed here. In addition to the RO function (triggered by the arrival of a TCP packet at the HMS) defined in MIPv4-RO, this ERO option further enables an MH to indicate the HMS to conduct a binding update at the CH by incorporating the address of the CH in the Registration Request message. This is especially useful when the CH is tunnelling the TCP packets directly towards the MH (enabled by either the SS option or the RO option)

whereas no FMS or FA is available in the previous foreign domain, out of which the MH has just moved for this inter-domain handoff. In MIPv4-RO, the previous FA sends a Binding Warning message to the HA on receiving a packet towards an MH that has moved, and the HA then sends a BU to the CH. The previous FA learns the current address of the MH by binding update from the new FA of the MH. This strategy is applicable to the scenarios where an FMS is deployed in both domains involved in the handoff.

For UDP mobility, the MH initiates the binding update for end-to-end route optimisation and session renegotiation at the CH using the SIP re-INVITE message. Meanwhile, it conducts the MIPv4 home registration. Similar to the IPv6 case, if a handoff involves both TCP and UDP sessions from the same CH, only UDP mobility signalling is carried out. In that case, after processing the SIP re-INVITE, the CH may tunnel the following TCP data directly towards the new location of the MH, bypassing the HMS, and the ERO option may not be needed. Figure 4.8 shows the above handoff operations. For brevity, only the basic mode is demonstrated.

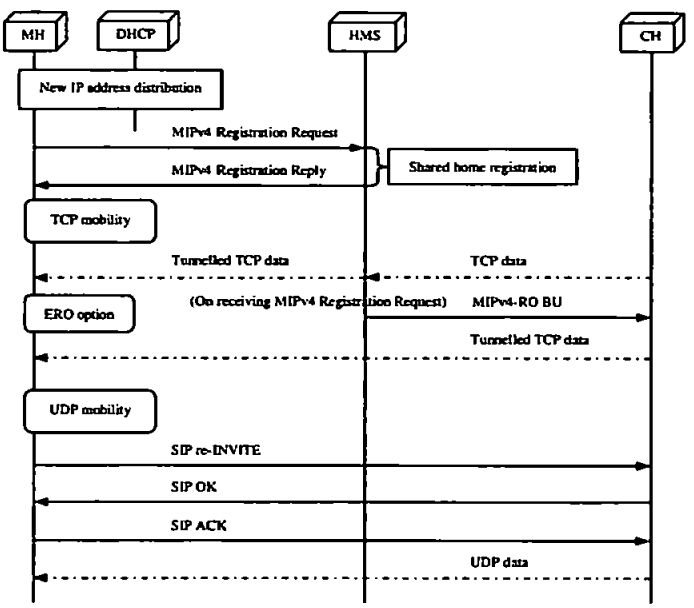


Figure 4.10 TI-MIPv4-SIP: handoff

4.3.2.3. Handoff in HY-MIPv6-SIP

Finally, in contrast to the TI-MIPv6-SIP handoff management, we present the handoff signalling in a reference hybrid MIPv6-SIP model as illustrated in Figure 4.11. This hybrid architecture, referred to as HY-MIPv6-SIP, is constructed by applying the design philosophies (illustrated in Figure 2.19) found in the hybrid MIPv4-SIP architectures. For simplicity, only the basic mode is shown here. In HY-MIPv6-SIP, both MIPv6 and SIP messages are applied independently to support TCP and UDP mobility in the IPv6 (MIPv6) context. To authorise the binding update at the CH, a SIP RR process, analogous to the MIPv6 RR process though running in the application layer, is assumed based on SIP INFO messages [RFC2976] instead of reusing MIPv6 RR. In addition, for a location update regardless of the MH's mode (active or idle), both MIPv6 and SIP home registration procedures would be triggered simultaneously. For session setup, the HY-MIPv6-SIP follows that (Figure 2.18) in the hybrid MIPv4-SIP architecture. Clearly, this architecture also suffers from similar serious redundancy found in its IPv4 counterpart.

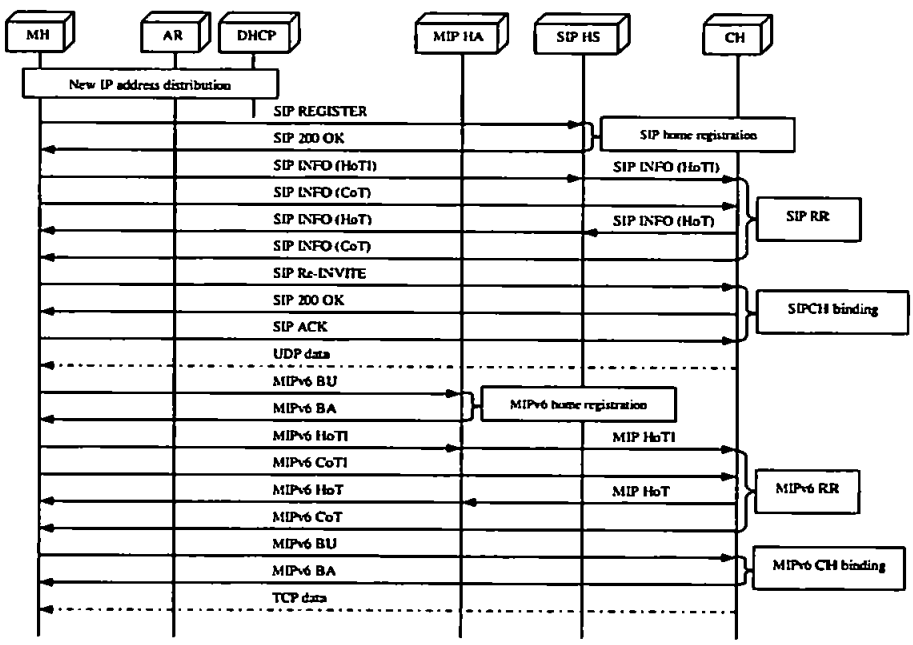


Figure 4.11 Hybrid MIPv6-SIP: handoff (basic mode)

4.4 Support for Various Mobility Types

As pointed out in Chapter 3, only terminal mobility is emphasised in conventional wireless systems. Nevertheless, various emerging mobility types are expected in the next-generation IP-based networks. In this section, we present our considerations on supporting selected mobility types in the proposed integrated MIP-SIP architecture.

4.4.1. Mobility Support Policy

In the proposed architecture, mobility decisions for different mobility types are made according to a series of pre-defined handoff policies residing in a mobility policy table installed in the MH. The design of the policy table is based on the observations that different mobility types are detectable and differentiable with the help of user input, L2 triggers and the System Profile that accommodates retrievable system-specific information such as network type, service provider and QoS parameters, downloadable from a network entity such as an AR. After the handoff detection, the MH decides the respective type of mobility by consulting the policy table and the User Profile that contains user and application preferences on mobility support. The mobility is then executed by enhanced MIP and SIP mobility schemes, referred to as MIP+ and SIP+, respectively, in the integrated MIP-SIP architecture. Figure 4.12 illustrates the process of detection, decision and execution of selected mobility types. Generally, low- and high-level mobility types are handled by MIP+ and SIP+, respectively. Among the high-level mobility types, we consider session and personal mobility; whilst in the low-level mobility category, terminal and mode mobility will be discussed. In fact, more mobility types and their detection, decision and execution could be added to the mobility policy table.

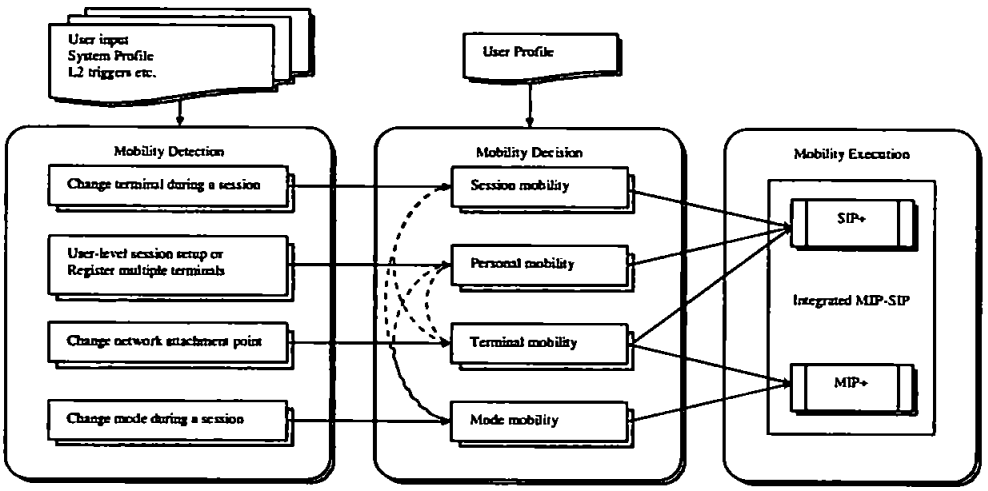


Figure 4.12 TI-MIP-SIP: support for various mobility types

4.4.2. Support for Terminal and Personal Mobility

The support for terminal and personal mobility has been achieved in the design of the proposed integrated MIP-SIP architecture. As indicated in the design, terminal mobility are supported by either SIP+ or MIP+ when the differentiation is considered between (UDP) real-time applications and (TCP) non-real-time applications. Regarding personal mobility, the basic mode in the design supports user-level session setup whilst the advanced mode further supports multiple-terminal registrations.

Notably, more than one type of mobility can happen simultaneously. The dotted lines in Figure 4.12 indicate the possible combinations between different mobility types, though we only discuss the simultaneous terminal and personal mobility case. For instance, when a user changes network attachment point, he or she may register a new terminal administrated by the new network attachment point in addition to the one being used. In this scenario, both terminal mobility (idle or active mode) and a kind of personal mobility (registrations for multiple terminals) are detected and supported by the advanced mode of TI-MIP-SIP as depicted before.

4.4.3. Support for Session Mobility

Session mobility occurs when a session needs to be transferred from one terminal (MH1) to another trusted terminal (MH2), e.g., in a PAN belonging to a user. This procedure is usually initiated for cost-effective communications, e.g. a PAN user may switch a multimedia session from his/her PDA to his/her PC when he/she enters his/her office from outdoors. It can also be triggered by pre-defined user or application preferences so that the handoff is transparent to the user. Figure 4.13 illustrates the signalling and data flows in the presence of the optional FMS, based on [Schulzrinne and Wedlund 2000].

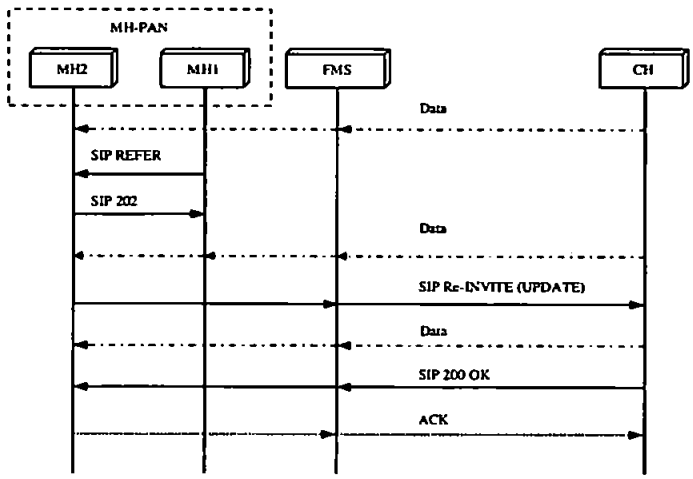


Figure 4.13 TI-MIP-SIP: session mobility

The SIP REFER method [RFC3515] plays a central role to facilitate such an operation. Firstly, MH1 sends a SIP REFER message to MH2, indicating the ongoing session with a SDP description. Necessary AAA information related to the session is also transferred via the REFER. We assume that authentication between members of a PAN has been established. Thus, MH2 replies with a SIP 202 to MH1 if this reference is accepted. Based on the session description from the MH1, MH2 may send Re-INVITE (or UPDATE) to the CH if it determines that it needs to renegotiate (or update) the session with the CH. The AAA information transferred from MH1 needs to be enclosed for authorisation etc.

purpose. On completion of this process, the incoming data flows are transferred from MH1 to MH2. Notably, on receipt of the SIP 202, MH1 can start to relay incoming packets to MH2 to reduce packet loss during the handoff.

Note that compared with [Schulzrinne and Wedlund 2000] two enhancements are proposed here: one is the context transfer of AAA information, among other session-related information; the other is the relay of in-flight session data during the handoff. These two enhancements could enable a more secure and smooth session handoff.

4.4.4. Support for Mode Mobility

Mode mobility happens when an MH changes from the ad hoc mode to the infrastructure mode or vice versa during an ongoing session. We consider a special case of mode mobility, where an ad hoc network interworks with an infrastructure-based network. This mobility scenario is referred to as network mobility (NEMO) in the IETF NEMO Working group. Under network mobility, an MH within a moving network communicates a CH connected to the infrastructure network via a Mobile Router (MR), which serves as the gateway of the moving network. So far, the NEMO basic support protocol [RFC3963] has been standardised based on MIPv6 without route optimisation. In this protocol, when the MR, together with the moving network, enters a foreign domain, it sends a BU to its HA to register its new CoA on behalf of the moving network. In the BU, a new Mobility Header Option is defined to carry the moving network's prefix information so that the HA can forward the MR the packets meant for hosts in the moving network. On successful home registration, a bi-directional tunnel is established between the HA and the MR, and all the traffic between the MH and the CH passes through the HA (and the MR). Surely, this protocol is also applicable in the proposed integrated MIP-SIP architecture, where the

HMS can takes the role of the HA. Figure 4.14 illustrates the network mobility support in the proposed integrated MIP-SIP architecture.

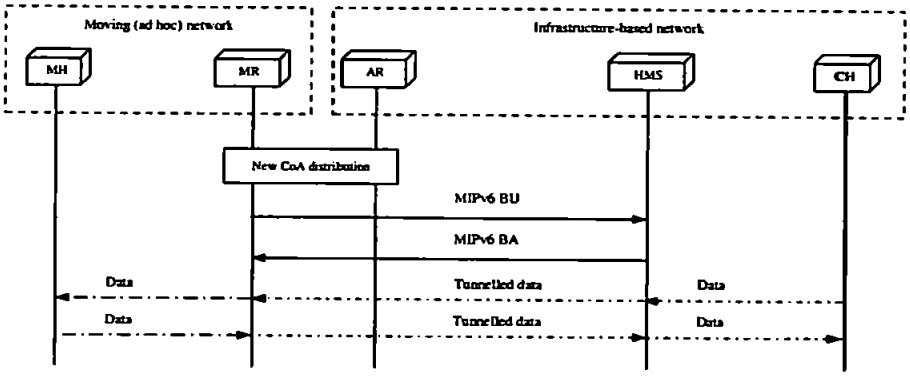


Figure 4.14 TI-MIP-SIP: network mobility

The routing between an MH and the MR (or other host in the moving network) within the moving network may be based on a proper ad hoc mobility routing protocol [Royer and Toh 1999, Abolhasan etc 2004], similar to the conception in the MIPMANET architecture [Jönsson etc 2000]. Furthermore, research is underway to add route optimisation to the basic support protocol, among other advanced requirements [Lach etc 2003]. Nevertheless, almost all the optimisations and enhancements are proposed in the MIP (especially MIPv6) platform. Therefore, in principle the applicability of these proposals in the proposed integrated architecture should be no problem.

To sum up, various mobility types can be supported in the proposed integrated MIP-SIP architecture, thanks to its integration of the powerfulness of both MIP and SIP protocols.

4.4.5. Support for Emergency Services

In addition to the above support for diverse mobility scenarios, mobile users also expect to summon IP-based emergency services, comparable to the existing services reachable at a well-known number such as 999 in UK, 911 in North America, and 112 in many other countries. In the IETF, the Internet Emergency Preparedness (IEPREP), the

Emergency Context Resolution with Internet Technologies (ECRIT) and related WGs such as the Session Initiation Proposal Investigation (SIPPING) are tackling this problem. SIP-based solutions are a natural choice as SIP has been chosen as the VoIP signalling protocol by both IETF and 3GPP. Technically, SIP has an existing "priority" field in the Request-URI that distinguishes sessions of different importance levels. The five values currently defined are "emergency", "urgent", "normal", "non-urgent" and "other-priority". However, to distinguish an emergency call (session) for public emergency service from one for private urgent communication (e.g., between colleagues), it may be desirable to define a universal emergency SIP URI such as sip(s):sos@domain [Schulzrinne 2006], analogous to 999. Once an emergency call is identified, the SIP infrastructure can deliver the call to an appropriate Public Safety Answering Point (PSAP).

In light of SIP's capability to specify and route an emergency call, the proposed TI-MIP-SIP architecture can be easily extended to support IP-based emergency services by using SIP. Figure 4.15 illustrates the signalling and data flows, based on [Schulzrinne and Arabshian 2002]. To initiate an emergency service request, an MH sends an INVITE to its MIP-SIP HMS or FMS. In the INVITE, the location of the caller is included so help can be dispatched to the right place. For this purpose, location information provided by positioning services such as GPS should be enclosed. To indicate the call is for public emergency service, a predefined emergency URI is used as the Request-URI. On receiving an INVITE of this kind, the HMS or FMS consults an Emergency Provider Access Directory (EPAD) to retrieve the contact information of a (nearby) PSAP. The messages exchanged between the HMS/FMS and the EPAD may be non-SIP-based. After obtaining the PSAP's address, the HMS/FMS can either deliver the INVITE to the PSAP as a SIP proxy or replies the MH with a Moved Temporarily message enclosing the PSAP's address

so that the MH can contact the PSAP directly by sending a new INVITE. Figure 4.15 demonstrates the latter case.

Since security is vital for an emergency service, IPsec [RFC4301] and/or SIP security functionality must be exploited to protect the integrity of the signalling information and to verify the authorisation of a request before allowing it to use the emergency service [RFC4190].

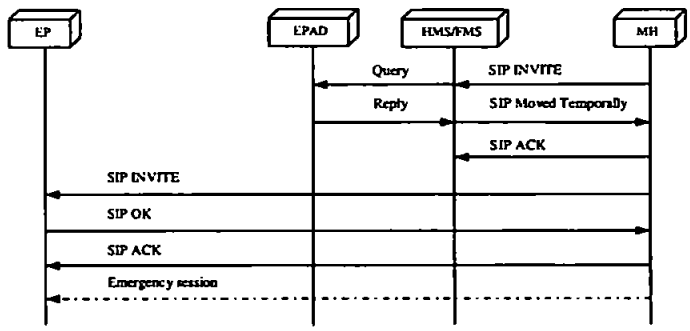


Figure 4.15 TI-MIP-SIP: emergency services

4.5 Performance Evaluation

In this section, we evaluate the performance of the proposed tightly integrated MIP-SIP architecture, focusing on the IPv6 version (i.e., TI-MIPv6-SIP), and compare TI-MIPv6-SIP with its hybrid counterpart HY-MIPv6-SIP, the Pure SIP approach and the Pure MIPv6 approach. We focus our evaluation on the support of terminal and personal mobility in the advanced mode for comparison.

In the rest of this section, we first justify and define the evaluation metric in Sections 4.5.1. Subsequently, we present the analytical model and configuration parameters for the evaluation in Section 4.5.2. The analysis and corresponding analytical results are provided in Sections 4.5.3 and 4.5.4, respectively. To validate the analytical results, simulation results are presented and discussed in Section 4.5.5.

4.5.1. Evaluation Metric

In the evaluation of mobility-management architectures for wide-area wireless networks, cost-efficiency (or cost-effectiveness) has always been among the top design considerations [Bafutto etc 1994, Pollini etc 1995, La Porta etc 1996, Akyildiz and Wang 2002, Wu etc 2002, Pack and Choi 2004, Lo etc 2004], and thus it is the focus of this evaluation. In the following, we define and justify signalling costs as the major metric for cost-efficiency assessment. (Additional related evaluations under more metrics are performed in Chapter 5.)

Mobility signalling traffic accounts for a great fraction of the overall signalling load in a wireless mobile system, and the signalling load generated by wireless mobile users is significantly larger than that generated by their wired counterparts [Pollini etc 1995]. Plethoric signalling loads tend to over-consume the valuable bandwidth of the links, and the processing capacity of the routers and the involved servers, and thus may lead to system performance degradation and affect the committed QoS of the services [Bafutto etc 1994]. In particular, signalling load generated by a macro-mobility protocol exerts a global burden on the system and thus is the major concern in the protocol design. Therefore, signalling costs is widely used as the top or even solo metric in evaluating a mobility management architecture in the literature (e.g., [Bafutto etc 1994, Pollini etc 1995, La Porta etc 1996]).

The contribution of an individual message to the network load depends on the message length (or size) and the sequence of visited network nodes on the path between its origin and destination [Bafutto etc 1994]. Therefore, the signalling costs generated by a message ($C_{signalling}$) can be calculated as the product of the message length ($L_{message}$) and the distance it traverses between the origin node A and the destination node B (or B to A) in the network [Lo etc 2004, Wu etc 2002]. The value of a distance parameter can be assigned

with an absolute value of hops (H_{A-B}) or a weighted value (ω_{A-B}). The aggregate signalling costs generated by a mobility protocol are the summation of the costs contributed by all the involved individual messages.

4.5.2. Analytical Model and Configuration Parameters

4.5.2.1. Domain Model

In the analysis, the reference foreign domain consists of K rings of regular hexagonal cells (subnets), centred on cell '0' with increasing label numbers to K ($K \geq 0$), as illustrated in Figure 4.16.

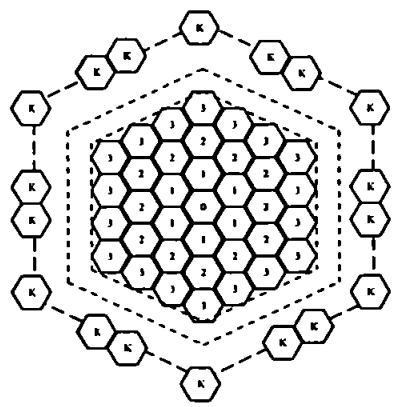


Figure 4.16 Domain model

Table 4.2 lists the configuration parameters and their typical values or formulae related to the domain model [La Porta etc 1996, Akyildiz and Wang 2002, Pack and Choi 2004]. Note that with the default parameter values the domain area is considerably large: $A_D = \sqrt{3} \cdot L_C^2 \cdot N(4)/24 = 1761 \text{ mile}^2$.

Table 4.2 Configuration parameters in the domain model

Symbol	*Parameter*	*Typical (Default) Value or Formula*
L_C	Perimeter of a cell (subnet)	20 mile
K	Number of rings in a domain	4
$N(K)$	Number of cells in a domain	$3K(K+1)+1$
L_D	Perimeter of a domain	$(2K+1) \cdot L_{cell}$
A_D	Area of a domain	$\sqrt{3} \cdot L_C^2 \cdot N(K)/24$

4.5.2.2. Mobility Model

For macro mobility, we utilise the fluid-flow model [Pollini etc 1995, La Porta etc 1996, Akyildiz and Wang 2002], which is suitable for modelling MHs with high mobility yet infrequent speed and direction changes. The model assumes that the direction of an MH's movement is uniformly distributed over [0, 2π] in a domain of arbitrary shape and the MHs in the domain are uniformly populated. The mean inter-domain crossing rate per MH (i.e., macro mobility rate) is given by

$$\lambda_M = v \cdot L_D / (\pi \cdot A_D), \tag{4.1}$$

where v is the average movement velocity, and L_d and A_d is the perimeter and the area of the domain, respectively. When an MH crosses the domain boundary, either the handoff or the location update procedure is invoked, depending on the current mode (active or idle). Furthermore, as widely accepted [Akyildiz and Wang 2002, Sen etc 1999] we assume that the session arrival to an MH is a Poisson process with the mean rate λ_s and the session holding time obeys an exponential distribution with the mean value $1/\mu$. λ_s / λ_m is known as call-to-mobility rate (CMR). To simplify the evaluation, a CH is assumed a static (wireless or wired) host who initiates sessions in a different domain and sends packets to the MH visiting in a foreign domain. When a handoff occurs, the MH is supposed to be receiving packets from a CH, involving one live session based on TCP or UDP, or both with one TCP and one UDP (with different probabilities). The configurations of all the parameters in the mobility model are listed in Table 4.3.

4.5.2.3. Message Lengths

Next, we identify the typical lengths of the messages. The MIP message lengths are estimated from the involved IETF RFCs. SIP messages are text-based and session-specific, and thus we approximate their typical values based on empirical implementations [Dutta etc 2001B, 3GPP-GP508]. Moreover, the length of a message may vary slightly along its

path from the source to the destination as intermediate nodes may modify some headers, though we disregard this effect for simplicity. In this analysis, DHCP is used as the common mechanism for host configuration including new IP address distribution, and thus the DHCP signalling is not included in our analysis. Table 4.4 lists the lengths of the involved MIP and SIP messages in the IPv6 context (with IPv6 and UDP headers).

Table 4.3 Configuration parameters in the mobility model

Symbol	*Parameter*	*Typical (Default) Value or Formula*
v	Mean speed of MHs	50 mile/hr
λ_M	Mean inter-domain movement rate per MH (i.e., macro mobility rate)	$v \cdot L_D / (\pi \cdot A_D)$
λ_S	Mean session (call) arrival rate	2 /hr/MH
$1/\mu$	Mean session (call) holding time	1/20 hr (3 min)
$P_{SS\text{-}TCP}$, $P_{SS\text{-}UDP}$	Probability that an arrival session is TCP or UDP based	0.5, 0.5
P_{TCP}, P_{UDP}, P_{TCPUDP}	Probability that a macro handoff involves TCP, UDP or both kinds of traffic with a CH	0.45, 0.45, 0.1
$P_{multi\text{-}register}$	Probability that a location update involves multiple-terminal registration	0.2
$P_{explicit\text{-}deregister}$	Probability that an explicit de-registration is applied	0.2
P_{UPDATE}	Probability that a UDP session needs renegotiation or update upon a macro handoff	0.5
$\omega_{MH\text{-}HMS}$	Distance weight between an MH and its home mobility server	0.75
$\omega_{MH\text{-}CH}$	Distance weight between an MH and its CH	1.00
$\omega_{HMS\text{-}CH}$	Distance weight between the CH and the MH's home mobility server	0.40
$\omega_{FMSn\text{-}FMSo}$	Distance weight between the old and the new foreign mobility servers	0.10

Table 4.4 Typical lengths (bytes) of MIPv6 and SIP messages (with IPv6 UDP headers): (a) MIPv6 message length, (b) SIP message length

(a)

MIPv6 Message		Symbol	Message Length
Home Test Init (HoTI)	MH->HA	L_{HoTI}	116
	HA->CH		64
Care-of Test Init (CoTI)	MH->CH	L_{CoTI}	64
Home Test (HoT)	CH->HA	L_{HoT}	72
	HA->MH		124
Care-of Test (CoT)	CH->MH	L_{CoT}	72
BU	MH->HA	L_{BU}	108
	MH->CH		80
BA	HA->MH	L_{BA}	76
	CH->MH		80

(b)

SIP Message	Symbol	Message Length
Re-INVITE	$L_{Re\text{-}INVITE}$	490
200 OK (Re-INVITE)	L_{OK}	420
ACK (200 OK)	L_{ACK}	256
UPDATE	L_{UPDATE}	490
200 OK (UPDATE)	L_{OK}	420
Re-REGISTER	$L_{Re\text{-}REGISTER}$	465
200 OK (Re-REGISTER)	L_{OK}	450
De-REGISTER	$L_{De\text{-}REGISTER}$	412
200 OK (De-REGISTER)	L_{OK}	550
INFO	L_{INFO}	400
200 OK (INFO)	L_{OK}	400

4.5.3. Cost Analysis and Analytical Results

In this section, we analyse the efficiency of the proposed Integrated MIP-SIP architecture, compared with Hybrid MIP-SIP, Pure SIP and Pure MIP, by computing the signalling costs based on the analytical models provided in Section 4.5.2. For brevity, we only demonstrate the advanced mode in the IPv6 (MIPv6) context though the basic mode and the IPv4 (MIPv4) context can be similarly analysed.

4.5.3.1. Computation Methodology of Signalling Costs

The average IP-level signalling costs generated per unit time by a mobility procedure i consisting of n processes are calculated as

$$C_{procedure-i} = R_{procedure-i} \cdot \sum_{j=1}^{n} \xi_{process-j}$$
$$= R_{procedure-i} \cdot \sum_{j=1}^{n} \sum_{k=1}^{m_j} (P_{message-k} \cdot L_{message-k} \cdot \omega_{A-B})_j , \quad (4.2)$$

where $R_{procedure-i}$ is the rate at which procedure i is invoked, $\xi_{process-j}$ is the costs generated by process j, $P_{message-k}$ is the probability that message k is involved, $L_{message-k}$ is the IP-level length of message k, m_j is the number of messages involved in process j, and ω_{A-B} is the non-directional weighted distance for message k crossing between the source A and the destination B and vice versa.

In the following, we derive these involved parameters except those that have been identified in the previous subsections. To simplify the analysis, as commonly adopted a roaming MH communicates with a CH that is assumed a wired host in a remote domain. When a macro handoff occurs, the MH is communicating with the CH involving one ongoing session based on TCP or UDP, or both with one TCP and one UDP connection. Moreover, DHCP is supposed as the common host-configuration mechanism and thus is not included in our analysis, and messaging for periodical refreshes is not counted either.

We focus on the two major mobility procedures: location update (LU) and handoff (HO) and their corresponding and aggregate signalling costs generated per roaming MH. The HO rate R_{HO}^{Macro} and LU rate R_{LU}^{Macro} per MH due to inter-domain (macro) movements are given by

$$R_{HO}^{Macro} = \lambda_M \cdot P_{busy}, \text{ and} \tag{4.3}$$

$$R_{LU}^{Macro} = \lambda_M \cdot (1 - P_{busy}), \text{ respectively,} \tag{4.4}$$

where P_{busy} is the probability that an MH is in the active mode. Assume that an MH is an M/M/1/1 system, P_{busy} is given by [Sen etc 1999]

$$\begin{aligned} P_{busy} &= \frac{\lambda_S}{\lambda_S + \mu} \\ &= \frac{\lambda_S / \mu}{\lambda_S / \mu + 1} \\ &= \frac{\varepsilon}{\varepsilon + 1}, \end{aligned} \tag{4.5}$$

where ε is the product of session arrival rate and session holding time (i.e., $\varepsilon = \lambda_S / \mu$), known as Erlang(s).

4.5.3.2. Signalling Costs of Involved Processes

Based on Figure 4.6 and Figure 4.9, Table 4.5 presents the equations for calculating the signalling costs incurred by the mobility processes in TI-MIPv6-SIP. Equations are also derived for Pure MIP, Pure SIP and Hybrid MIP-SIP, and listed in Table 4.6, Table 4.7, and Table 4.8, respectively. Notably, Pure MIP only supports basic terminal mobility and its equations are listed for completeness and reference purpose only.

Table 4.5 Mobility process equations for Integrated MIP-SIP

Process	Equation	
Home registration	$\xi^{LU\,HO-HR}_{Int-MIP-SIP} = ((1-P_{multi-register})\cdot(L_{BU}+L_{BA}) + P_{multi-register}\cdot(L_{Re-REGISTER}+L_{OK}))\cdot\omega_{MH-HMS}$	(4.6)
Explicit de-registration	$\xi^{LU-Derreg}_{Int-MIP-SIP} = P_{explicit-derregister}\cdot(L_{BU}+L_{BA})\cdot\omega_{FMSo-FMSo}$	(4.7)
Smooth handoff	$\xi^{HO-SH}_{Int-MIP-SIP} = (L_{BU}+L_{BA})\cdot\omega_{FMSo-FMSo}$	(4.8)
Return routability	$\xi^{HO-RR}_{Int-MIP-SIP} = (L_{HoTI}+L_{HoT})\cdot\omega_{MH-HMS} + (L_{HoTI}+L_{HoT})\cdot\omega_{HMS-CH} + (L_{CoTI}+L_{CoT})\cdot\omega_{MH-CH}$	(4.9)
CH binding update Option 1	$\xi^{HO-BU-1}_{Int-MIP-SIP} = (P_{UDP}+P_{UPD+TCP})\cdot(L_{Re-INVITE}+L_{OK}+L_{ACK})\cdot\omega_{MH-CH} + P_{TCP}\cdot(L_{BU}+L_{BA})\cdot\omega_{MH-CH}$	(4.10)
CH binding update Option 2	$\xi^{HO-BU-2}_{Int-MIP-SIP} = (P_{UDP}+P_{UPD+TCP})\cdot((L_{BU}+L_{BA})\cdot\omega_{MH-CH} + P_{UPDATE}\cdot(L_{UPDATE}+L_{OK})\cdot\omega_{MH-CH}) + P_{TCP}\cdot(L_{BU}+L_{BA})\cdot\omega_{MH-CH}$	(4.11)

Table 4.6 Mobility process equations for Pure MIP

Process	Equation	
Home registration	$\xi^{LU\,HO-HR}_{Pure-MIP} = (L_{BU}+L_{BA})\cdot\omega_{MH-HMS}$	(4.12)
Explicit de-registration	$\xi^{LU-Derreg}_{Pure-MIP} = P_{explicit-derregister}\cdot(L_{BU}+L_{BA})\cdot\omega_{FMSo-FMSo}$	(4.13)
Smooth handoff	$\xi^{HO-SH}_{Pure-MIP} = (L_{BU}+L_{BA})\cdot\omega_{FMSo-FMSo}$	(4.14)
Return routability	$\xi^{HO-RR}_{Pure-MIP} = (L_{HoTI}+L_{HoT})\cdot\omega_{MH-HMS} + (L_{HoTI}+L_{HoT})\cdot\omega_{HMS-CH} + (L_{CoTI}+L_{CoT})\cdot\omega_{MH-CH}$	(4.15)
CH binding update	$\xi^{HO-BU}_{Pure-MIP} = (L_{BU}+L_{BA})\cdot\omega_{MH-CH}$	(4.16)

Table 4.7 Mobility process equations for Pure SIP

Process	Equation	
Home registration	$\xi^{LU\,HO-HR}_{Pure-SIP} = (L_{Re-REGISTER}+L_{OK})\cdot\omega_{MH-HMS}$	(4.17)
Explicit de-registration	$\xi^{LU-Derreg}_{Pure-SIP} = P_{explicit-derregister}\cdot(L_{De-REGISTER}+L_{OK})\cdot\omega_{FMSo-FMSo}$	(4.18)
Smooth handoff	$\xi^{HO-SH}_{Pure-SIP} = (L_{INFO}+L_{OK})\cdot\omega_{FMSo-FMSo}$	(4.19)
Return routability	$\xi^{HO-RR}_{Pure-SIP} = (L_{INFO}+L_{OK})\cdot\omega_{MH-HMS} + (L_{INFO}+L_{OK})\cdot\omega_{HMS-CH} + (L_{INFO}+L_{OK})\cdot\omega_{MH-CH}$	(4.20)
CH binding update	$\xi^{HO-BU}_{Pure-SIP} = P_{UDP}\cdot(L_{Re-INVITE}+L_{OK}+L_{ACK})\cdot\omega_{MH-CH} + P_{TCP}\cdot(L_{INFO}+L_{OK})\cdot\omega_{MH-CH} + P_{UDP+TCP}\cdot(L_{Re-INVITE}+L_{OK}+L_{ACK}+L_{INFO}+L_{OK})\cdot\omega_{MH-CH}$	(4.21)

Table 4.8 Mobility process equations for Hybrid MIP-SIP

Process	Equation	
Home registration	$\xi^{LU\,HO-HR}_{Hyb-MIP-SIP} = \xi^{LU-HR}_{Pure-MIP} + \xi^{LU-HR}_{Pure-SIP}$	(4.22)
Explicit de-registration	$\xi^{LU-Derreg}_{Hyb-MIP-SIP} = \xi^{LU-Derreg}_{Pure-MIP} + \xi^{LU-Derreg}_{Pure-SIP}$	(4.23)
Smooth handoff	$\xi^{HO-SH}_{Hyb-MIP-SIP} = P_{TCP}\cdot\xi^{HO-SH}_{Pure-MIP} + P_{UDP}\cdot\xi^{HO-SH}_{Pure-SIP} + P_{UDP+TCP}\cdot(\xi^{HO-SH}_{Pure-MIP}+\xi^{HO-SH}_{Pure-SIP})$	(4.24)
Return routability	$\xi^{HO-RR}_{Hyb-MIP-SIP} = P_{TCP}\cdot\xi^{HO-RR}_{Pure-MIP} + P_{UDP}\cdot\xi^{HO-RR}_{Pure-SIP} + P_{UDP+TCP}\cdot(\xi^{HO-RR}_{Pure-MIP}+\xi^{HO-RR}_{Pure-SIP})$	(4.25)
CH binding update	$\xi^{HO-BU}_{Hyb-MIP-SIP} = P_{TCP}\cdot\xi^{HO-BU}_{Pure-MIP} + P_{UDP}\cdot(L_{Re-INVITE}+L_{OK}+L_{ACK})\cdot\omega_{MH-CH} + P_{UDP+TCP}\cdot(L_{Re-INVITE}+L_{OK}+L_{ACK}+L_{BU}+L_{BA})\cdot\omega_{MH-CH}$	(4.26)

4.5.3.3. Unit Signalling Costs of Each Process

By substituting the typical values listed in Tables 4.3 and 4.4 into the equations in Tables 4.5-4.8, we obtain the unit signalling costs generated by each process as shown in Figure 4.17. In most processes, compared with Pure SIP and Hybrid MIP-SIP, Integrated MIP-SIP dramatically reduces the costs by more than half thanks to our systematic integration and optimisation. Notably, Integrated MIP-SIP with Option 2 (referred to as Integrated MIP-SIP 2) performs even better than Integrated MIP-SIP with Option 1 (referred to as Integrated MIP-SIP 1) in terms of fewer CH binding update costs generated. This exemplifies a typical improvement from our dynamical combination of MIP and SIP messages.

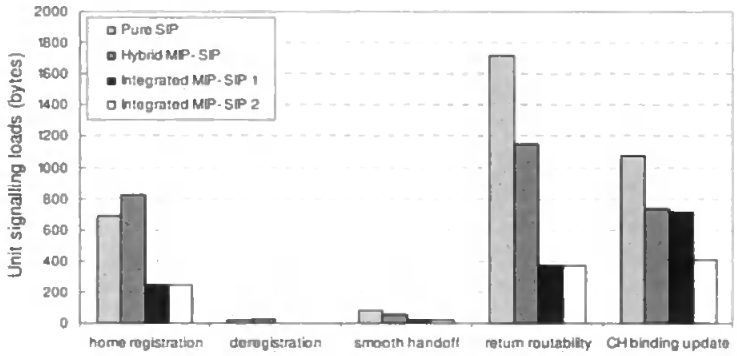


Figure 4.17 Unit signalling costs of each mobility process

4.5.3.4. Dynamic Signalling Costs

The total signalling costs per MH per hour of location updates, handoffs, and both (collectively referred to as mobility management) in scheme X are given by

$$C_X^{LU} = (\xi_X^{LU/HO-HR} + \xi_X^{LU-Dereg}) \cdot R_{LU}^{Macro}, \tag{4.27}$$

$$C_X^{HO} = (\xi_X^{LU/HO-HR} + \xi_X^{HO-SH} + \xi_X^{HO-RR} + \xi_X^{HO-BU}) \cdot R_{HO}^{Macro}, \text{ and} \tag{4.28}$$

$$C_X^{Mobility-Management} = S_X^{LU+HO} = S_X^{LU} + S_X^{HO}, \text{ respectively.} \tag{4.29}$$

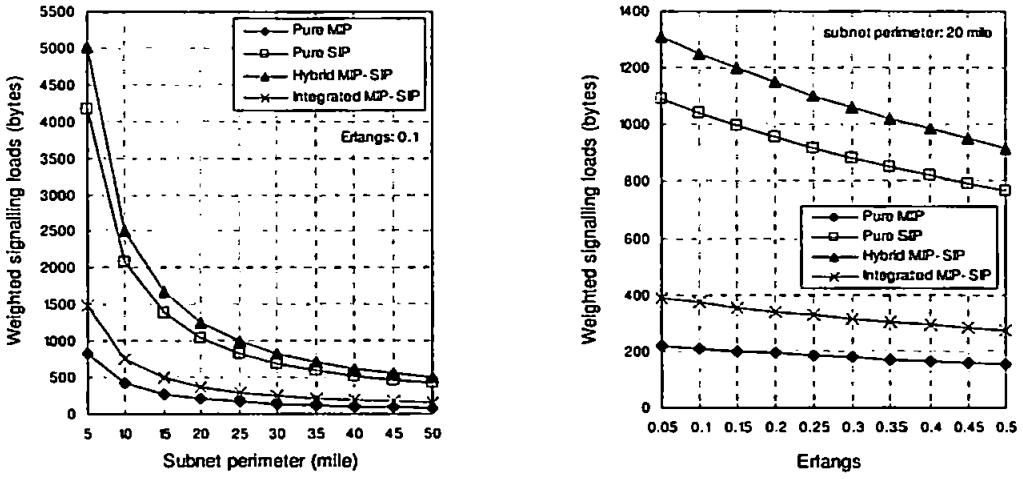
The above signalling costs depend on the parameters of the mobility model.

Generally, in each scheme as shown in Figure 4.18 (a) and Figure 4.19 (a) with the increase of the subnet perimeter, the rates of domain crossing decrease fast and so do the location updates and handoffs. Consequently, the signalling costs generated by either location updates or handoffs, or the aggregate of both shown in Figure 4.20 (a) decrease sharply. In contrast, with the increase of Erlangs the probability that an MH is in active mode on a domain crossing also increases whilst the idle-mode probability decreases. Thus, more handoffs occur and more handoff signalling costs are yielded as shown in Figure 4.19 (b). The total mobility management costs also increase in Figure 4.20 (b) in spite of the decrease in idle-mode location updates observed in Figure 4.18 (b).

Another overall observation is that the signalling costs generated by handoffs accounts for a non-trivial proportion of the mobility management even when the Erlangs are small. For instance, when Erlangs = 0.05, the ratio of handoff costs to the total mobility management costs (referred to as HMR) ranges between 14% and 21% when all the architectures are concerned. This observation results from the fact that many more messages are involved in handoffs than those in location updates due to the complexity of IP handoff procedures. Furthermore, in 3G and beyond cellular networks, it is expected that more applications are emerging in addition to traditional voice calls, and thus users are likely to be occupied by various active sessions for more time, leading to an increase of Erlangs and the HMR. For example, when Erlangs = 0.35, the HMR of all architectures ranges from 53% to 65%. This also explains why the total mobility management costs increase despite the fact that the location-update costs decrease as the Erlangs increase, shown in Figure 4.20 (b).

We now examine the signalling performance of individual mobility architecture. Of all the architectures, Pure MIP serves as a reference benchmark as it only supports basic

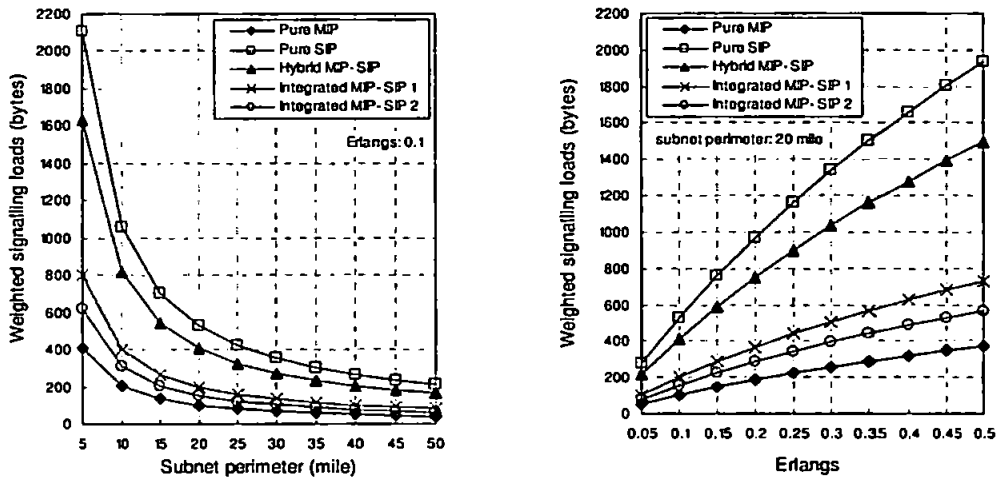
terminal mobility. Among the other three schemes that support advanced terminal mobility and personal mobility, Hybrid MIP-SIP and Pure SIP invoke the largest costs for location updates and handoffs, respectively, and comparable large aggregate costs for mobility management. Hybrid MIP-SIP tends to double the costs generated in Integrated MIP-SIP since redundant MIP and SIP messages have to be triggered simultaneously from time to time because MIP and SIP are unaware of each other's protocol syntax and their entities are independently deployed. The fact that Pure SIP incurs similar huge costs is also predictable because of the large SIP message sizes. In contrast, Integrated MIP-SIP generates the lowest signalling costs, thanks to its flexible use of MIP and SIP messages between integrated entities, and particularly the maximised selective use of MIP messages to take the advantage that MIP messages are much smaller than SIP ones for some common routines. The load reductions compared with Hybrid MIP-SIP and Pure SIP in location update, handoff or in total are almost all over 62%. In addition, Integrated MIP-SIP 2 outperforms Integrated MIP-SIP 1 thanks to its more efficient CH binding update process. Notably, all these schemes except Pure MIP can all benefit from the emerging SIP message compression [RFC3486] for fewer costs as the harmful generosity of SIP messages sizes have been noticed. However, even if the compression could be applied to Pure SIP on every hop of the signalling path, Pure SIP cannot achieve an average compression ratio of anywhere near 60%. Let alone that practically this compression is only applied to the wireless hop. For instance, a recent implementation of SIP compression [Pous etc 2004] shows that the compression ratio achieved for the REGISTER and 200 OK messages are just 20% and 49%, respectively. In addition, these compression and decompression operations introduce more system complexity, processing cost and delays. Clearly, Integrated MIP-SIP appears more efficient in supporting advanced mobility management.



(a) (b)

Figure 4.18 Signalling costs of macro location updates:

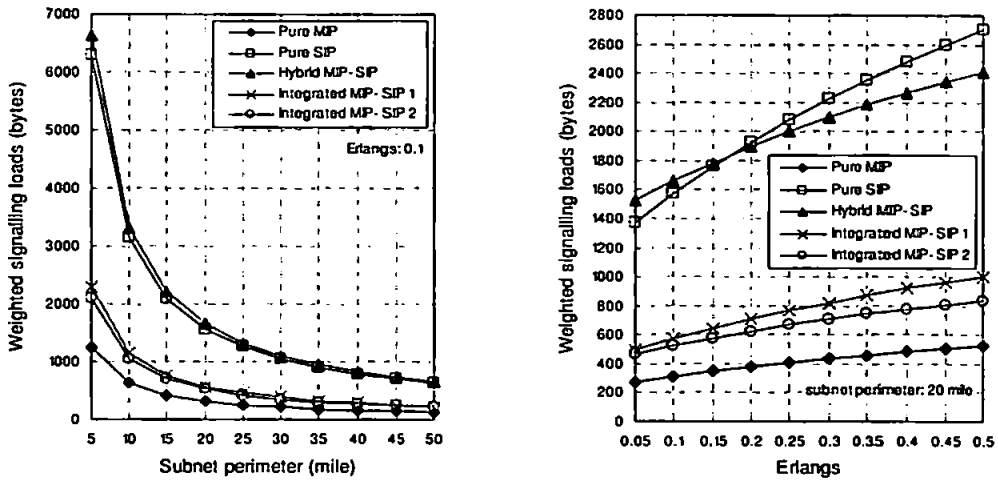
(a) when the subnet perimeter varies, (b) when the Erlangs vary



(a) (b)

Figure 4.19 Signalling costs of macro handoffs:

(a) when the subnet perimeter varies, (b) when the Erlangs vary



(a) (b)

Figure 4.20 Signalling costs of macro mobility management:

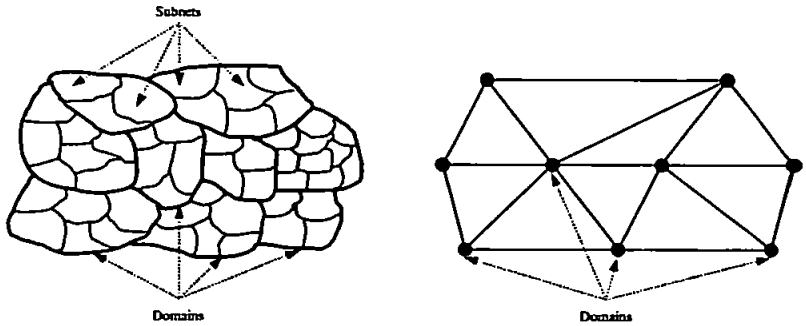
(a) when the subnet perimeter varies, (b) when the Erlangs vary

4.5.4. Simulation Results

So far, we have derived the analytical results of macro-mobility signalling costs based on a fluid-flow mobility model and a hexagonal network layout model. In this section, we attempt to capture the realistic results and validate our major conclusions drawn from the analytical results by simulating mobility and networking scenarios that are more practical. The simulations are developed with Microsoft Visual C++ 7.0.

4.5.4.1. Simulation Configuration

Similar to [Sen etc 1999], the network layout is modelled as a bounded-degree, connected graph $G = (V, E)$, where the node-set V represents the domains and the edge-set E represents the access paths between pairs of domains. Each domain consists of a number of subnets, and each subnet is featured by an arbitrary shape and has an arbitrary number of neighbouring subnets. The network model for the simulation is depicted in Figure 4.21 (b), which corresponds to an actual domain layout [Sen etc 1999] as shown in Figure 4.21 (a).



(a) (b)

Figure 4.21 Simulation network model:

(a) realistic network layout, (b) network layout graph

Regarding the mobility model, we consider a directional inter-domain movement model since macro mobility deals with large-scale movements where an MH usually moves on purpose rather than completely randomly. In this directional mobility model, initially an MH resides in one of the nine domains as shown in Figure 4.21 (b) and selects a destination domain from the other eight domains randomly at the beginning of the simulation. Then the MH finds out the shortest path in terms of inter-domain hops to the destination domain. If more than one shortest path is discovered, a random one is taken. Afterwards, the MH heads to the destination domain by passing zero or more intermediate domains. After the MH reaches the destination subnet of the destination domain, it reselects the next destination domain randomly and repeats the above steps. This model is close to real-world macro mobility, which is a mixture of deterministic and random movements [Sen etc 1999]. For simulation purpose, we have the following assumptions. The destination domain selection is uniformly distributed. The number of subnets passed by the MH to exit a domain or reach the destination subnet of a destination domain obeys a uniform distribution [Akyildiz and Wang 2002] on 1 to 5 inclusive. The subnet resident time follows a Gamma distribution [Akyildiz and Wang 2002] with the mean value 3 min and the variance 3 min. A remote fixed CH keeps on initiating TCP and UDP sessions

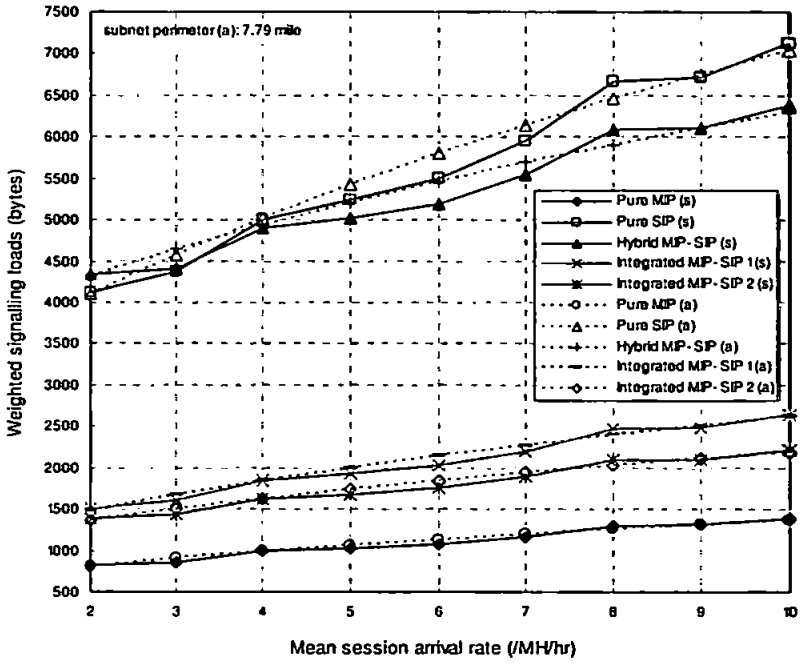
alternately towards the MH. The session arrival is assumed a Poisson process with mean rate λ_s. The session holding time is exponentially distributed with the mean 3 min. In the simulations, λ_s is the variable.

4.5.4.2. Simulation Results

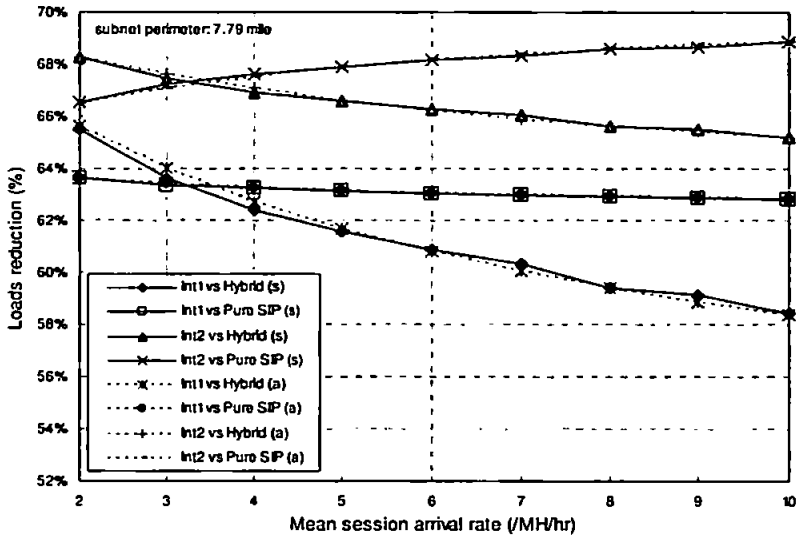
In each simulated scenario for a given mean session arrival rate λ_s, a simulation stops when 30 inter-domain movements have taken place, which corresponds to 5 ~ 10 simulated working hours for a day. Each scenario is repeatedly simulated and the average results are then obtained. During all the simulations, the mean inter-domain crossing rate in each scenario turns out to range from 3.95 to 4.37 (times/MH/hr) with the average value 4.18, which is equivalent to the domain-boundary crossing rate using the hexagonal network model where the perimeter of a subnet is 7.79 mile. Hence, in the following we compare the simulation results (denoted by *s*) with the corresponding analytical results (denoted by *a*).

Figure 4.22 (a) and (b) compares the proposed architecture with the other concerned ones in terms of hourly mobility-management signalling costs in absolute values and reduction percentages, respectively. In both cases, we note that the simulation results (solid lines) roughly resemble the analytical ones (dotted lines) though reasonable fluctuations exist. Thus, the major conclusions based on the analytical results can be largely validated. Specifically, as shown in Figure 4.22 (a) with the increase of the mean session arrival rate the signalling costs of all the concerned architectures tend to increase, whereas the costs in Hybrid MIP-SIP and Pure SIP are consistently much larger than those in Integrated MIP-SIP are. These differences are clearly expressed in reduction ratios in Figure 4.22 (b), where most of the reductions are more than 60% when both Integrated MIP-SIP 1 and 2 (simply denoted by Int1 and Int2 in the figure) are compared with the others. Particularly, Integrated MIP-SIP 2 reduces the costs by over 65% in all the scenarios. In summary,

again we can conclude that the proposed integrated architecture significantly outperforms the traditional ones in terms of signalling cost-efficiency.



(a)



(b)

Figure 4.22. Comparison of simulation and analytical results:

(a) weighted signalling costs, (b) costs reduction

4.6 Concluding Remarks

In this chapter, we have proposed a tightly integrated MIP-SIP architecture called TI-MIP-SIP for cost-effective macro-mobility management. The design motivation and methodology is to make full use of standardised work from both protocols, select composite processes that are more efficient for common functions, integrate similar entities and procedures to reduce redundancies, and avoid further duplicate standardisation. Both IPv4 (MIPv4) and IPv6 (MIPv6) contexts are investigated and appropriate protocols are designed.

The proposed architecture combines the complementary powerfulness of MIP and SIP architectures and protocols, which are the two dominant approaches to macro-mobility support. As desired in the next-generation wireless systems, the architecture supports advanced IP-based terminal and personal mobility. It is featured by the capability to locate a roaming user globally regardless of his or her current location or the terminal being used, the effective support for both TCP and UDP applications, the choice to register with multiple terminals, and the adaptation to macro handoffs by session renegotiation or update. By integrating MIP and SIP entities and operations of similar functionality, redundant processing and signalling as found in the traditional hybrid MIP and SIP approach are minimised and thus the system efficiency is significantly improved. Notably, standard MIP and SIP messages are reused with minimum extensions instead of introducing new messages, and hence the deployment is facilitated. The efficiency improvements are evaluated in terms of signalling costs. Both the analytical and the simulation results reveal that the Integrated MIP-SIP architecture consistently outperforms its hybrid counterpart and the Pure SIP scheme greatly. In most cases, the reduction in signalling costs is over

60%. Therefore, the proposed architecture can be far more cost-efficient in the use of the MH's battery, radio spectrum and network resources.

In addition to the focused terminal and personal mobility support, the proposed TI-MIP-SIP architecture allows the compatibility of both MIP and SIP and thus facilitates the support for various additional mobility types such as session mobility and network mobility. Furthermore, a number of options are proposed so that the system functionalities such as session setup and handoff management are enhanced.

Chapter 5

The Loosely Integrated MIP-SIP Architecture for Macro-Mobility Support

In this chapter, we present an alternative approach to the Tightly Integrated MIP-SIP (TI-MIP-SIP) for macro-mobility support and the proposed architecture is called Loosely Integrated MIP-SIP (LI-MIP-SIP). This chapter is partially based on two publications [Wang and Abu-Rgheff 3G2004, ICC04].

5.1 Introduction

By integrating MIP and SIP mobility entities and procedures in the Tightly Integrated MIP-SIP architecture (TI-MIP-SIP, Chapter 4), one can expect to minimise the serious redundancy found in the traditional hybrid schemes and thus maximise the system cost-effectiveness. However, though such a tight integration of both protocols would prove to be most cost-effective in a long run, a more prompt or flexible deployment may necessitate efficient joint MIP-SIP architecture where MIP and SIP physical entities are located separately, rather than merged or collocated. Therefore, for short- to mid-term deployment considerations we propose an alternative architecture called Loosely Integrated MIP-SIP architecture (LI-MIP-SIP). The main idea is to establish necessary interactions between

selected MIP and SIP entities to enable efficient joint mobility support without physically merging the MIP and the SIP infrastructure.

The remainder of this chapter is structured as follows. Section 5.2 and 5.3 describe the architectural and signalling design of the proposed LI-MIP-SIP architecture. The performances of LI-MIP-SIP, together with TI-MIP-SIP wherever appropriate, are analysed in Section 5.4, and the analytical and simulation results are presented in Section 5.5 and Section 5.6, respectively. Concluding remarks are provided in Section 5.7.

5.2 Architectural Design of the Loosely Integrated MIP-SIP Architecture

In this section, we present the architectural design of the LI-MIP-SIP architecture. We start with an overview of the architecture in Section 5.2.1, and then describe the operation of the enhanced mobility servers and the management of the diverse addresses in Sections 5.2.2 and 5.2.3, respectively.

5.2.1. Architecture Overview

The network structure of LI-MIP-SIP is shown in Figure 5.1. It looks like the one in TI-MIP-SIP (Figure 4.1) though the MIP and SIP mobility servers in a domain are not merged as a whole. In fact, it more resembles the deployment in the typical hybrid architecture (Figure 2.14), where MIP and SIP home or foreign servers coexist independently. In LI-MIP-SIP, the MIP HA and the SIP HS are located individually whereas connected to each other in the logical home domain of an MH. Notably, these two mobility servers are not necessarily placed in a same physical domain (though it is the common scenario especially in corporate network environments), and thus we actually do not assume any constrains on their locations. The inter-connection (directly or indirectly as

indicated by the clouds) between the home servers physically facilitates their interactions to be discussed. In a foreign domain, the MIP and SIP foreign servers (FSs) are collocated with the domain gateway (GW), like the typical hybrid architecture scenario. As explained for the TI-MIP-SIP FMS in Chapter 4, the specific format of a MIP FS depends on the choice of the micro-mobility protocol, and the presence of the MIP FS and the SIP FS is not mandatory for macro-mobility management. Therefore, similarly we leave the discussions regarding these foreign servers to Chapter 6. In addition, DHCP is again assumed for uniform IP address distribution.

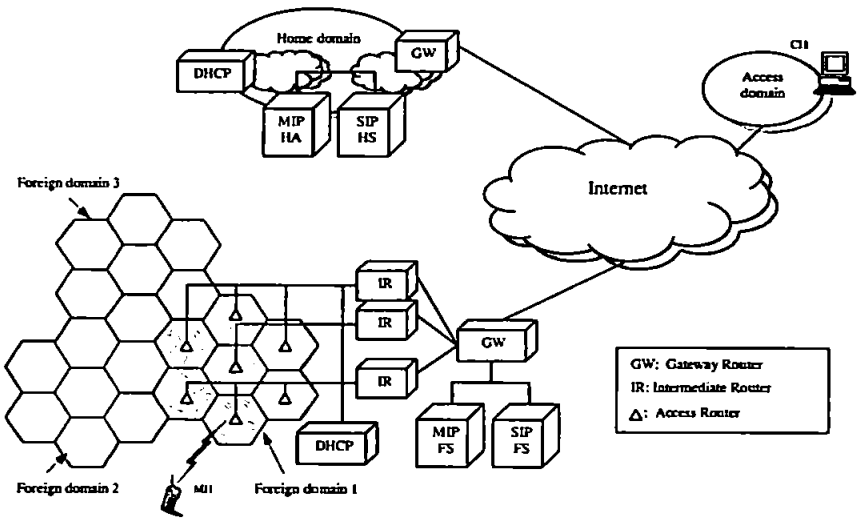


Figure 5.1 LI-MIP-SIP: network model

According to the above description, the design focus for this macro-mobility management is to introduce necessary interactions between the MIP HA and the SIP HS to share mobility information to reduce the otherwise redundant system costs, as identified in the hybrid architectures. Particularly, the duplicate home registrations and refreshes between an MH and its home servers can thus be replaced with sole messaging like TI-MIP-SIP. Such efficient location-update signalling also extends the battery life of the MH and reduces the traffic over the wireless link. With some choices for the design considered (discussed in the next section), two schemes are proposed to establish the desired

interactions. Scheme I is called on-demand location enquiry (ODLE), whilst Scheme II is named synchronised location update (SYLU). In ODLE, the SIP HS enquires for the current location of the targeted MH from the MIP HA when it needs a location service, e.g., for setting up a SIP session. In SYLU, the MIP HA updates the SIP location services at the SIP HS on behalf of an MH after receiving a home registration or refresh from the MH. Through these two schemes, the LI-MIP-SIP architecture is expected to achieve improved cost-effectiveness compared with the hybrid architectures.

5.2.2. Mobility Server Enhancements

5.2.2.1. Design Choices for Interactions

There exist various design choices to establish interactions between the MIP HA and the SIP HS, and each choice may lead to different enhancement requirements on the mobility servers. For instance, it is possible for the MIP HA and the SIP HS to have a full-scale mutual sharing of information available at each entity if they are enabled to understand the protocol syntax of one another. However, such a design would entail significant modifications to both entities (e.g., MIP may have to be modified to understand the SIP URI). We argue that any enhancements should be well justified for deployment purpose. In other words, only necessary information sharing should be enabled by a more careful design. Therefore, we take an alternative that only enables the SIP HS to be aware of selected MIP messages whilst the MIP HA does not need to understand the SIP syntax. One may consider an opposite design, which only enables the MIP HA to understand some SIP syntax. Nevertheless, the corresponding MIP enhancements require heavy modifications in the operating system, compared to the more handy enhancements in the user space where SIP is implemented. Thus, we prefer the design choice where the enhancements mainly take place in the SIP HS. This indicates that the interactions

preferably utilize MIP messages (thus, SIP HS should be enabled to understand selected MIP messages). Another reason to choose MIP messages lies in the fact that MIP messages are far more compact than their SIP counterparts are and thus would generate significantly less overheads, particularly when the MIP HA and the SIP HS are located far away from each other. In sum, the above considerations direct us to the design of the two mentioned schemes, whose interactions are based on MIP messages.

5.2.2.2. Address Mapping

The next design step is to establish proper address links to connect the two location databases available in the MIP HA and the SIP HS. As aforementioned, SIP and MIP use different location bindings: SIP (AOR, contact URI) and MIP (HoA, CoA) together with an NAI. A SIP contact URI can be created by appending the @ symbol and the SIP contact IP address to the SIP AOR whose @ symbol is replaced with the # symbol. For example, if the AOR is SIP: Jane@home.com, the contact URI can be Jane#home.com@contact IPv4 address or Jane#home.com@[contact IPv6 address]. Note that an IPv6 address should be placed in square parentheses. As a uniform IP address distributor such as a DHCP server is preferred, a new IP address serves as the MIP co-located CoA and the new SIP contact IP address. We thus need to use HoA or NAI (or both) to act as the index identifier in the interaction messages between the two location databases. Since NAI and AOR are both user-level identifiers that can be further mapped to various terminals, we propose to use the NAI as the primary index identifier for location enquiry or update. Though the HoA index can also be supported, we do not assume that the SIP HS keeps a record of an MH's HoA. Therefore, only the NAI needs to be mapped with the AOR in the SIP HS, and this mapping can be established when an MH initially registers with its SIP HS using both SIP and MIP registration messages. Sharing a similar format with a SIP URI, the NAI (in the form of username@domain's name) may be formulated from the AOR according to a pre-

defined rule so that the mapping between them can be simplified, though a generic mapping is assumed here for generality.

5.2.2.3. Enhancements to Mobility Servers

Consequently, the following enhancements are identified to the two schemes, respectively. For the ODLE scheme, the SIP HS uses the MIP HA as its principal location server, instead of its default associated location service. In fact, SIP does not mandate a particular mechanism for implementing the location service as long as the SIP HS is able to access to the service. For enquiry purpose, two primitives, Query and Response, are mentioned in the SIP standard between a SIP HS and its location service. The two primitives can be embodied by a pair of MIP messages: MIPv6 BRR (Binding Refresh Request) and BU (Binding Update), or MIPv4-RO Binding Request (BR) and BU. Thus, the SIP HS and the MIP HA are enabled to apply these messages to fulfil the query and response in a location enquiry initiated by the SIP HS.

Regarding the SYLU scheme, the MIP HA initiates the location update at the location service associated with the SIP HS on behalf of an MH, and the SIP HS is enabled to understand the MIP message for location update and acknowledge with another MIP message. The applicable pair of messages is MIPv6 BU and BA (Binding Acknowledgement), or MIPv4 Registration Request and Registration Reply.

5.2.3. Mobility Server Operation

In this section, we present the detailed operation performed by the MIP and SIP mobility servers in the two interaction schemes, ODLE and SYLU, respectively.

5.2.3.1. Mobility Server Operation in ODLE

In the ODLE scheme, the SIP HS initiates the interaction by sending a MIP query message to the MIP HA. This operation is normally triggered by a SIP INVITE message

from a CH (or another SIP proxy server), e.g., for setting up a session between an MH and itself. On receiving an INVITE, the SIP HS checks the targeted URI (in the To header), which is the SIP AOR of the called MH in a standard SIP session setup. Note that the ODLE scheme further supports the URI based on an NAI (in the form of SIP: NAI) and the URI based on MIP HoA (in the form of SIP: username@IPv4 address or SIP: username@[IPv6 address]) thanks to the address mapping and server interactions pre-established between MIP and SIP. Therefore, an MH can be identified by its SIP AOR, its NAI, or its MIP HoA in the invitation. This would facilitate the success of a session invitation. In the AOR case, the SIP HS maps the AOR to an NAI according to the pre-established address mapping record, and then sends the MIP query message with the NAI as the query word. In the NAI-based URI case, the mapping process is omitted. In the IP-address-like AOR case, the SIP HS simply queries the MIP HA using the IP address. If the URI is invalid (neither a registered URI or NAI, nor a URI based on an IP address), the SIP HS sends the CH a SIP error response such as a 404 (Not Found) message. Figure 5.2 illustrate the operation flows at the SIP HS.

On receiving the MIP query message from the SIP HS, the MIP HA looks up its location database for record(s) matching the query identifier. If the identifier is a registered NAI, the MIP HA returns the SIP HS with one or more matched records, indicating the current IP address(es) of the matched terminal(s). If the identifier is an IP address, the MIP HA attempts to match it to an HoA. If a match is found, the CoA bound to the HoA is then returned to the SIP HS. If the above two attempts both fail, the MIP HA sends a MIP error response such as an ICMP error message with an error code to the SIP HS. These MIP HA operations are shown in Figure 5.3.

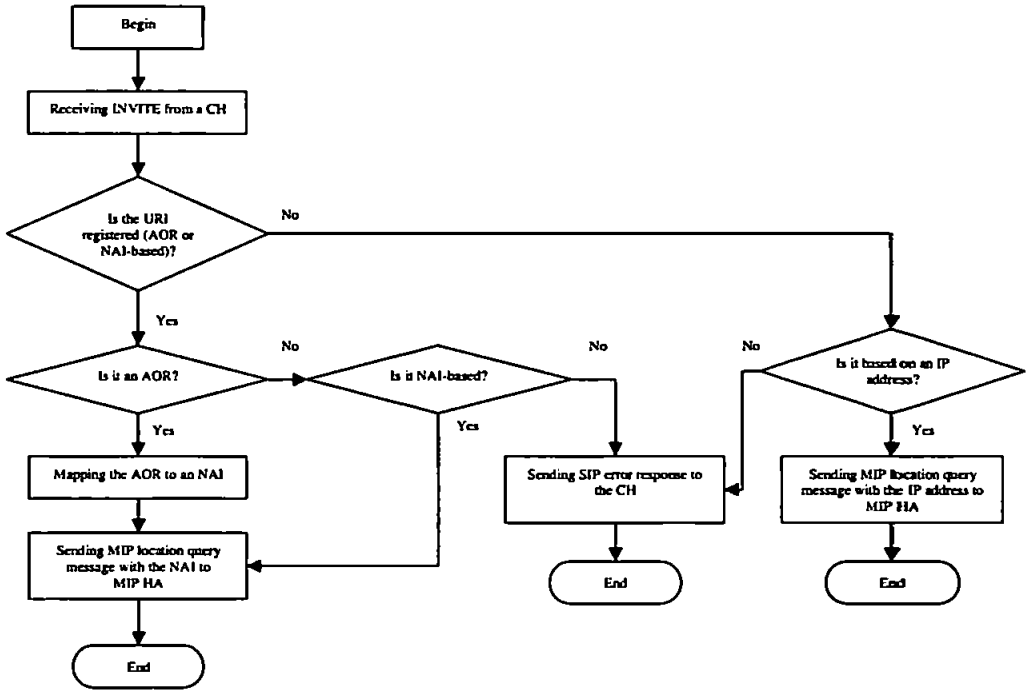


Figure 5.2 LI-MIP-SIP ODLE: SIP HS operation

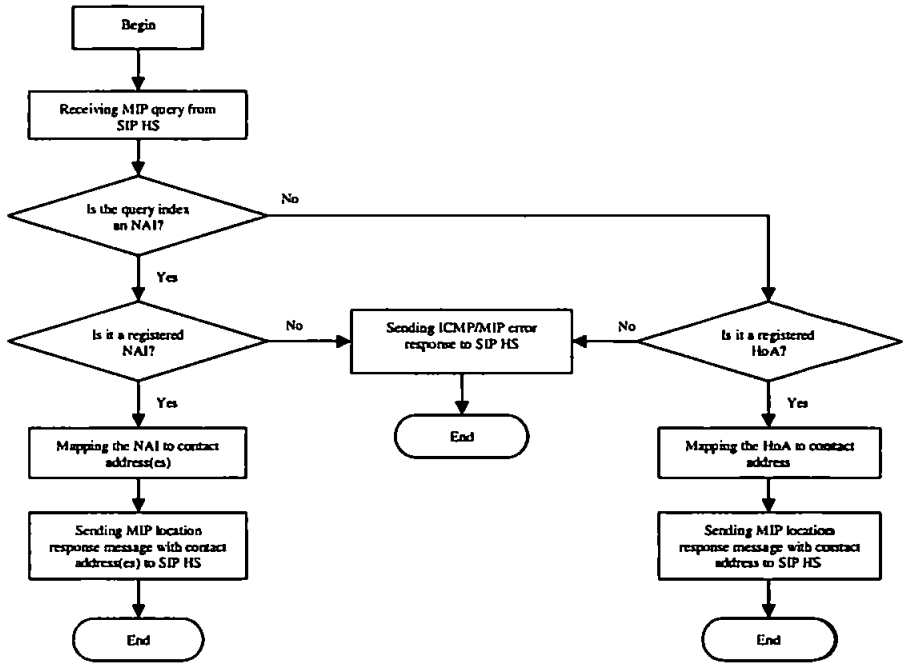


Figure 5.3 LI-MIP-SIP ODLE: MIP HA operation

Upon the receipt of the query results from the MIP HA, the SIP HS works as if it received the results from its associated location service, and act as either a proxy or a

redirect server. In addition, it may cache the results for a predefined short period, as the macro movements of an MH's are normally infrequent and so is the resultant home location update. This caching helps to reduce the signalling costs for location query and the session-setup delays.

5.2.3.2. Mobility Server Operation in SYLU

Regarding the alternative SYLU scheme, Figure 5.4 and Figure 5.5 demonstrate the operation of the MIP HA and SIP HS, respectively.

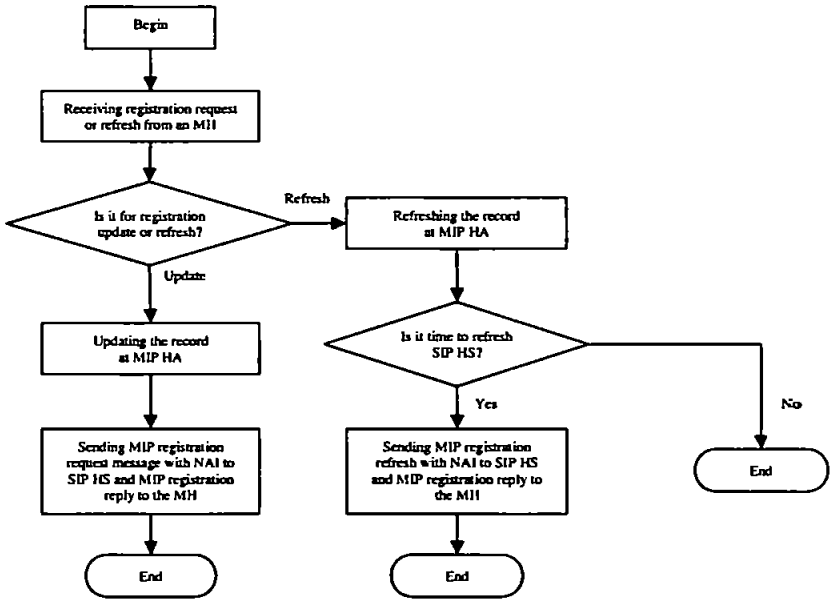


Figure 5.4 LI-MIP-SIP SYLU: MIP HA operation

In contrast to the ODLE scheme, it is the MIP HA that starts the interaction by sending a MIP registration request to the SIP HS in the SYLU scheme. This operation is normally triggered after the MIP HA processes a registration request or refresh from an MH. Note that the registration refresh at the SIP HS is not necessarily triggered immediately after a refresh at the HA; it is actually performed according to the pre-configured SIP refresh timer. On the other side, the SIP HS conducts the location update or refresh accordingly on the receipt of the MIP registration request from the MIP HA. The

validity check of the registration request and the corresponding error responses are not shown in Figure 5.4 or Figure 5.5 for brevity.

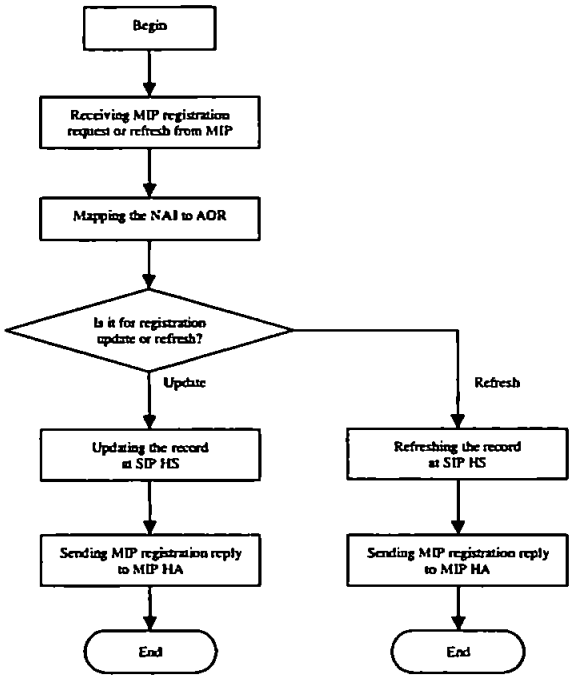


Figure 5.5 LI-MIP-SIP SYLU: SIP HS operation

5.2.4. Mobility Server Interactions

With the internal operation of the mobility servers described, we present the full picture of their external interactions, as shown in Figure 5.6.

Compared to the home mobility server (HMS) operation (Figure 4.2) in the TI-MIP-SIP architecture, the MIP HA and the SIP HS are not tightly integrated as a whole. Instead, they are located separately, possibly in different domains, though we do not preclude that they can be physically collocated. All the interactions between the MIP HA and the SIP HS are external signalling, which should be based on standard messages. Depending on the adopted interaction scheme, the two home servers exchange MIP-based messages for location enquiry (C1) or location update (A1) alternatively, normally triggered by C and A, respectively. An AAA association should be pre-established between the MIP HA and the

SIP HS for prompt interactions. They also collaborate with the home AAA server (AAAH) separately (E, F), especially when processing location-sensitive requests. Note that the MIP HA and the SIP HS may share a same AAAH or cooperate with different ones wherever appropriate, e.g., when they are deployed in different domains. The other signalling (A, C) or MIP data flows (B, B1, and B2) are similar to those of TI-MIP-SIP; in particular, the MIP SS (session setup) option (D) and the MIP ERO (enhanced route optimisation) option (not shown here) proposed in Chapter 4 are also applicable to the LI-MIP-SIP architecture, though they are not shown in the illustrations in Section 5.3 for clarity.

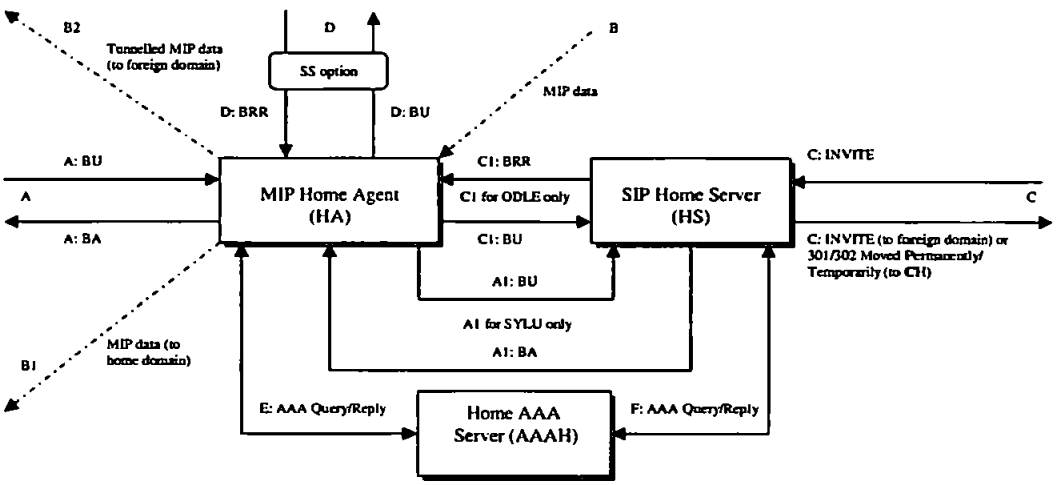


Figure 5.6 LI-MIP-SIP: home mobility servers interaction

5.3 Protocol Signalling Design of the Loosely Integrated MIP-SIP Architecture

This section specifies the signalling in the LI-MIP-SIP architecture. Similar to the TI-MIP-SIP architecture, the integration and interaction of SIP with both MIPv4 and MIPv6 are considered, and the corresponding location and handoff management procedures are proposed. The resultant protocols are referred to as LI-MIPv4-SIP and LI-MIPv6-SIP, respectively.

5.3.1. Location Management

Figure 5.7 and Figure 5.8 illustrate the initial home registration and home re-registration or refresh from a foreign domain, respectively.

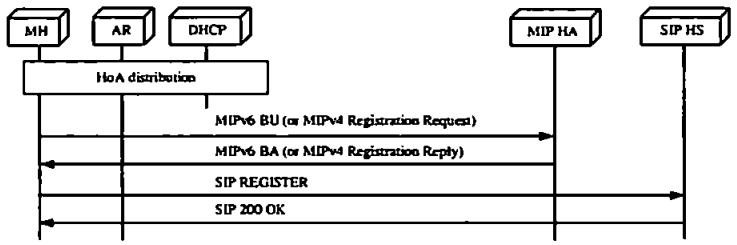


Figure 5.7 LI-MIP-SIP: initial home registration

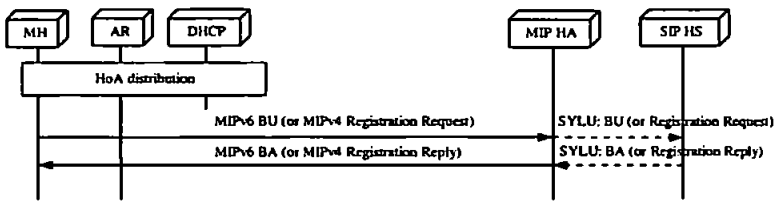


Figure 5.8 LI-MIP-SIP: home registration or refresh from a foreign domain

In the initial home registration, an MH applies both MIP and SIP registrations to create a record in the MIP HA and the SIP HS, respectively. The SIP REGISTER message is extended to carry the NAI of the MH for establishing a mapping between the NAI and the SIP AOR at the SIP HS. Subsequent home registrations or refreshes just use MIP registration messages. In the case of multiple HAs available for an MH, for ODLE the MH should inform the SIP HR of the address of the HA it has registered with as a MIP terminal through the initial SIP registration, though the SIP HR may be able to find out the correct HA itself by other means. In ODLE, an MH only updates the MIP HA; whereas in SYLU the MIP HA, updated by an MH, will then in turn updates the SIP HR using MIP home registration messages wherever appropriate. To register more than one terminal from a foreign domain, the MIPv4 Registration Request or the MIPv6 BU is extended to bear the additional information. Other optional operations such as the explicit deregistration with the previous FMS in an idle-mode location update and smooth handoff between the

previous and the new FMSs in a handoff can be enabled by the MIP registration messages if the FMSs are present. For simplicity, these optional operations are not shown.

Because of the different approaches in location updates, upon a SIP session setup, SYLU works as the hybrid architecture since the SIP HS has the up-to-date location of the MH. On the other hand, in ODLE the SIP HS turns to the MIP HA to enquire about the latest location of a targeted MH (or user) since no location updates have been done. Figure 5.9 shows the default session-setup procedure in LI-MIP-SIP.

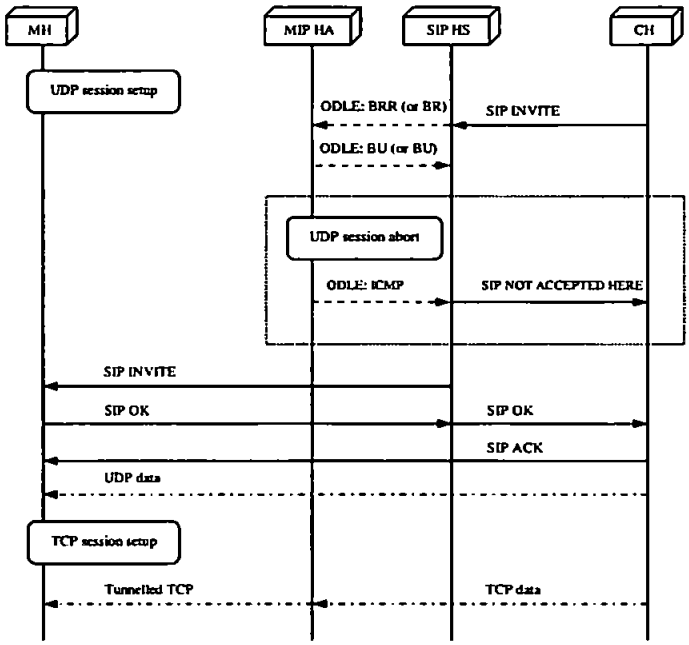


Figure 5.9 LI-MIP-SIP: session setup

After the session setup, a SIP session (UDP data) is established between the CH and the MH directly. For a MIP session (TCP data), a CH is able to send the data towards the MH directly if it has a valid binding entry or has acquired the up-to-date location information from the MH's MIP HA via the MIP SS option proposed in Chapter 4 (not shown in Figure 5.9). Otherwise, the CH simply sends the data to the MIP HA, which then tunnels the data to the MH by default or optionally sends a BU to the CH in MIPv4-RO and MIPv6.

Note that a UDP session setup may be aborted if the SIP HS cannot obtain the location information of the MH from either its own location service (SYLU) or from the MIP HA (ODLE). In either case, the SIP HS rejects the INVITE by returning a SIP 488 (Not Acceptable Here) response to the CH. In the response, a Warning header field value explains that the session setup has to be aborted due the callee is unreachable. In the case of ODLE, since the MIP HA is unable to locate the MH, it sends an ICMP host unreachable (or a MIP Binding Error) message to the SIP HS.

5.3.2. Handoff Management

The proposed macro-handoff signalling for LI-MIPv6-SIP and LI-MIPv4-SIP are shown in Figure 5.10 and Figure 5.11, respectively. In both ODLE and SYLU regardless of IPv6 and IPv4, MIP and SIP shares the MIP home registration at the MIP HA initiated by an MH, though in SYLU the location updates at the SIP HS are subsequently performed. The other signalling flows are similar to those in the TI-MIP-SIP architecture, and thus only a brief description is provided as follows.

5.3.2.1. Handoff in LI-MIPv6-SIP

In LI-MIPv6-SIP, for TCP mobility, the MIPv6 end-to-end route optimisation is triggered following the Return Routability (RR) tests. Note that the ERO option proposed in the context of TI-MIPv6-SIP in Chapter 4 is also applicable (though not shown in Figure 5.10). For UDP mobility, the SIP UDP-session handoff is employed (only the three-way handshake case is shown in this chapter though the other option proposed in Chapter 4 is applicable), and the MIPv6 home registration is reused. As discussed in TI-MIPv6-SIP, whether to reuse the RR process depends on the AAA implementation. If both TCP and UDP sessions from the same CH are involved in a handoff, the TCP mobility operations

are omitted as discussed in Chapter 4. In addition, the optional smooth handoff signalling between previous and new FMSs if present is not shown for brevity.

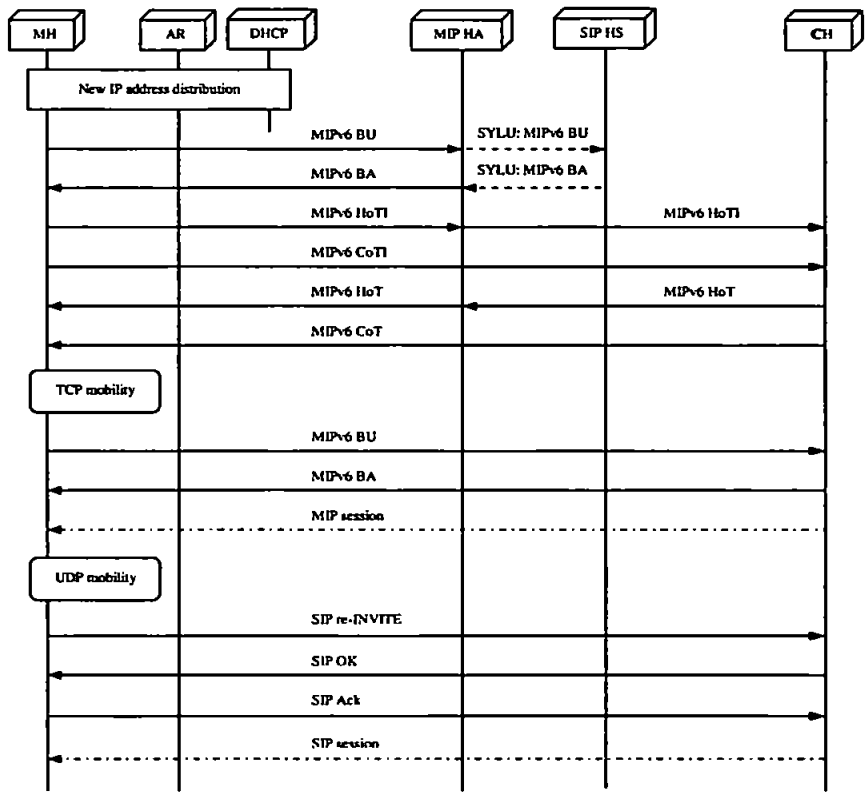


Figure 5.10 LI-MIPv6-SIP: handoff

Finally, it is worth noting that in any case the location update or refresh at the SIP HS performed in SYLU does not account for the handoff delays since the SIP HS is not involved for rerouting any data traffic on a handoff.

5.3.2.2. Handoff in LI-MIPv4-SIP

In LI-MIPv4-SIP, for TCP mobility, an MH performs a MIPv4 home registration at the MIP HA, which tunnels the subsequent TCP data to the new location of the MH. The ERO option proposed for TI-MIPv4-SIP in Chapter 4 is also applicable. For UDP mobility, end-to-end route optimisation and session renegotiation between the MH and the CH are conducted using SIP messages, and the home registration is accomplished by the MIPv4 registration messages. If a handoff involves both TCP and UDP sessions from a same CH,

only one MIPv4 home registration is performed in addition to the SIP messages between the MH and the CH for UDP-session handoff.

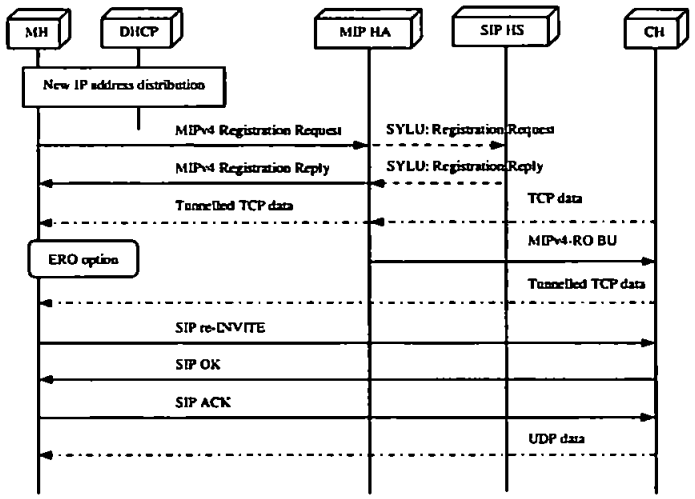


Figure 5.11 LI-MIPv4-SIP: handoff

5.4 Performance Analyses

In this section, we evaluate the performances of the proposed ODLE and SYLU protocols in the LI-MIP-SIP architecture and compare them with the TI-MIP-SIP and the Hybrid MIP-SIP (HY-MIP-SIP) architectures wherever appropriate. The evaluation metrics include costs, delays (handoff delay and session-setup delay), and handoff packet loss. The numerical results are either derived from theoretical analyses or collected from simulations, and sometimes obtained from the combinations of both methodologies. Both IPv6 and IPv4 scenarios are discussed wherever appropriate though we emphasise the IPv6 context.

For the analysis, we define the distances between the interested entities, illustrated in Figure 5.12. Let *a*, *b* and *c* represent the triangular distances between the MH, the CH and the HA (or HS or HMS), and *d* denote the distance between the MIP HA and the SIP HS.

Note that in TI-MIP-SIP, the HA and the HS is integrated into the HMS, and thus d is equal to zero.

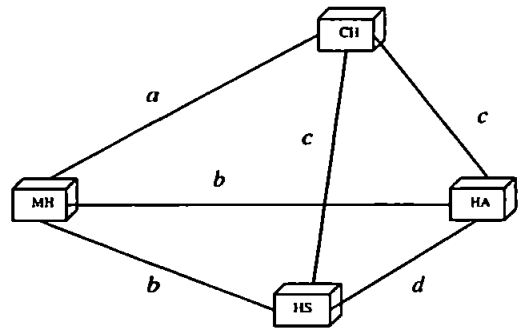


Figure 5.12 Distances (in hops) between entities

5.4.1. Signalling Cost Analysis

In this subsection, we analyse the signalling costs (as defined in Chapter 4) generated by the ODLE and the SYLU schemes in the LI-MIP-SIP architecture, referred to as LI-ODLE and LI-SYLU, respectively. The signalling costs incurred in TI-MIP-SIP and HY-MIP-SIP for location updates and handoffs have been analysed in Chapter 4, though we further include the signalling costs for session setups since LI-ODLE uses a different session setup procedure. In addition, the signalling costs for handoffs with the ERO option and for home location refreshments (at a frequency of λ_R in all protocols) are also included. Other differences from the signalling costs analyses in Chapter 4 are that the signalling costs generated here are invoked by the total roaming users (the number is denoted by N_V) rather than per user averagely in a foreign domain, and the absolute values for hops are used instead of weighted values to align with the subsequent delay evaluations. The involved parameters and assumptions have been defined in Chapter 4 unless stated otherwise.

The concerned signalling costs are provoked by location updates (LU), session setups (SS), and handoffs (HO). Thus, we calculate the aggregate signalling costs, which are a

sum of the location-update (LU) costs, session-setup (SS) costs and handoff (HO) costs, i.e., $C_{Sig}^{X} = C_{LU}^{X} + C_{SS}^{X} + C_{HO}^{X}$, where X stands for a protocol name. In the following, we analyse these costs generated in the LI-MIPv6-SIP architecture, and the corresponding costs in LI-MIPv4-SIP can be similarly derived and thus are omitted for brevity. For presentation clarity, we use the name of message to denote its length.

First, based on Figure 5.8, signalling costs for LU from a foreign domain in LI-ODLE are given by

$$C_{LU}^{LI-ODLE} = (BU + BA) \cdot b \cdot (R_{LU}^{Macro} + \lambda_R) \cdot N_V . \tag{5.1}$$

In contrast, the LU signalling costs in LI-SYLU include additional costs for home location updates and refreshes at the SIP HS, and thus they are given by

$$C_{LU}^{LI-SYLU} = C_{LU}^{LI-ODLE} + (BU + BA) \cdot d \cdot (R_{LU}^{Macro} + \lambda_R^{SIP}) \cdot N_{MHs} . \tag{5.2}$$

Next, assuming a CH initiates the invitation towards an MH in a foreign domain and the home SIP server of the MH acts as a proxy server (as shown in Figure 5.9), we calculate the SIP session-setup signalling costs in LI-SYLU by

$$C_{SS}^{LI-SYLU} = ((Invite + Ok) \cdot (c + b) + Ack \cdot a) \cdot (\lambda_S \cdot P_{SS-UDP}) \cdot N_V . \tag{5.3}$$

Regarding LI-ODLE, the SS signalling costs may involve additional costs for location enquires and replies between the SIP HS and the MIP HA, and thus the costs are given by

$$\begin{aligned} C_{SS}^{LI-ODLE} = {} & ((Invite + Ok) \cdot (c + b) + Ack \cdot a + P_{expire} \cdot (BRR + BU) \cdot d) \\ & \cdot (\lambda_s \cdot P_{SS-UDP}) \cdot N_{MHs} , \end{aligned} \tag{5.4}$$

where P_{expire} is the probability that the cached location information of an MH at the SIP HS has expired when the HS receives an INIVITE towards that MH. When no caching is implemented at the HS, a location query is always triggered, i.e., $P_{expire} = 1$.

Last, based on Figure 5.10, the handoff signalling costs in LI-ODLE are given by

$$C_{HO}^{LI-ODLE} = C_{HO-TCP}^{LI-ODLE} \cdot P_{HO-TCP} + C_{HO-UDP}^{LI-ODLE} \cdot P_{HO-UDP} + C_{HO-TCPUDP}^{LI-ODLE} \cdot P_{HO-TCPUDP} , \tag{5.5}$$

where $C_{HO-TCP}^{LI-ODLE}$, $C_{HO-UDP}^{LI-ODLE}$ and $C_{HO-TCPUDP}^{LI-ODLE}$ are the costs for TCP, UDP, and simultaneous TCP and UDP mobility (from the same CH), respectively. These costs are calculated as follows.

$$C_{HO-TCP}^{LI-ODLE} = [(BU + BA)\cdot(a+b) + \xi_{RR}^{MIP}]\cdot R_{HO}^{Macro}\cdot N_V, \tag{5.6}$$

where

$$\xi_{RR}^{MIP} = (HoTI + HoT)\cdot(b+c) + (CoTI + CoT)\cdot a. \tag{5.7}$$

$$C_{HO-UDP}^{LI-ODLE} = [(BU + BA)\cdot b + (reInvite + Ok + Ack)\cdot a + P_{RR}\cdot\xi_{RR}^{MIP}] \tag{5.8}$$

$$\cdot R_{HO}^{Macro}\cdot N_V, \tag{5.9}$$

where $P_{RR} = 1$ when the RR is needed, otherwise, $P_{RR} = 0$.

$$C_{HO-TCPUDP}^{LI-ODLE} = C_{HO-UDP}^{LI-ODLE}. \tag{5.10}$$

Since LI-SYLU incurs additional costs for home location updates at the SIP HS during handoffs, the HO signalling costs are given by

$$C_{HO}^{LI-SYLU} = C_{HO}^{LI-ODLE} + (BU + BA)\cdot d\cdot R_{HO}^{Macro}\cdot N_{MHs}. \tag{5.11}$$

Similarly, we can calculate the signalling costs generated in the other concerned architectures, and those incurred in LI-MIPv6-SIP with the ERO option (proposed in Chapter 4). For brevity, these expressions are omitted here.

5.4.2. Delay Analysis

In this subsection, we evaluate the handoff delays and session-setup delays in the concerned protocols. For the delay analysis, the following parameters are defined as listed in Table 5.1.

Table 5.1 Parameters for delay analysis

Symbol	*Parameter*
$B_{wireless-n}$	Bandwidth of the wireless hop n
$B_{wired-n}$	Bandwidth of the wired hop n
$T_{wireless-n}$	Latency of the wireless hop n
$T_{wired-n}$	Latency of the wired hop n
e	An entity of L3 or above, i.e., a router, a server, or an end host
h	Hops between the source entity and the destination entity of a message
S_i	Size of message (or data) of type i
$T_i^{(e)}$	Average sojourn latency for a message of type i at an entity e
$E^{(h)}$	Set of entities along the h hops
$\mu_i^{(e)}$	Mean service rate of message of type i at entity e
$\rho^{(e)}$	Utilisation (or load) at entity e
T_i^h	End-to-end one-way delay of message of type i across h hops
T_{IP}^X	Mean delay for an MH to acquire a new valid IP address on a handoff in mobility management protocol X

Based on [Lo etc 2004, Kim and Kim 2003, Choi etc 2004], the end-to-end one-way delay of a message (or a data) along a path of h hops is estimated by

$$T_i^h = \sum_{n=1}^{j}\left(\frac{S_i}{B_{wireless-n}} + T_{wireless-n}\right) + \sum_{n=1}^{h-j}\left(\frac{S_i}{B_{wired-n}} + T_{wired-n}\right) + \sum_{e \in E^{(h)}} T_i^{(e)}, \qquad (5.12)$$

where the first and second terms are transmission and propagation delays over the j wireless hops and the $(h - j)$ wired hops along the end-to-end path, respectively; and the third term is the accumulative sojourn delays incurred at the $(h + 1)$ entities along the path. Note that if there are no wireless or wired hops involved, the first or second term of the right hand in (5.13) becomes zero and j or $(h - j)$ is replaced with h, respectively.

To determine $T_i^{(e)}$, we model each entity as an M/M/1 queuing system [Willmann and Kuhn 1990, Murakami etc 2004]. Applying the queuing theory [Kleinrock 1976], we obtain

$$T_i^{(e)} = \frac{1}{\mu_i^{(e)}(1-\rho^{(e)})}. \qquad (5.13)$$

5.4.2.1. Handoff Delay

The handoff delay here is defined as the time elapsed between the instance when an MH requests for a new IP address on movement detection and the instance when the following incoming session traffic can be rerouted correctly to the new location of the MH. Therefore, the handoff delay is the accumulative delays incurred by the distribution of a new IP address and the operations to enable the traffic rerouting, e.g., by home registration in base MIPv4-based protocols or by CH registration (end-to-end route optimisation) in MIPv6-based protocols. We focus on analysing the handoff delays in IPv6 contexts. The proposed integrated protocols, TI-MIPv6-SIP and LI-MIPv6-SIP, are collectively referred to as Int6. In the following analysis, we use T^{X}_{HO-MIP} and T^{X}_{HO-SIP} to denote the handoff (HO) delay in protocol X for MIP sessions (TCP mobility) and SIP sessions (UDP mobility), respectively.

A. Handoff delay for MIP sessions (TCP mobility)

In MIPv6-based handoff protocols, the involved RR process imposes additional delays. According to the MIPv6 specification, an MH should initiate the RR (by sending a HoTI and a CoTI simultaneously) after sending a BU to its HA (or HMS) for home registration. Assuming that the RR is initiated immediately after a home registration BU is sent (the time difference between message transmissions is negligible and disregarded), the handoff delay for a MIPv6 session in Int6 is given by

$$T^{Int6}_{HO-MIP} = T^{Int6}_{IP} + \max((T^{b+c}_{HoTI} + T^{b+c}_{HoT}), (T^{a}_{CoTI} + T^{a}_{CoT})) + T^{a}_{BU}, \qquad (5.14)$$

where the second term of (5.16) is the RR delay and the third term is the subsequent CH binding-update delay. Since an RR process is completed only when both home and care-of tests are fulfilled (i.e., both HoT and CoT are received by the MH), the RR delay is decided by the larger delay for home or care-of test. Similarly, for HY-MIPv6-SIP (hyb6), the MIPv6 session handoff delay is given by

$$T_{HO-MIP}^{Hyb6} = T_{IP}^{MIP6} + \max((T_{HoTI}^{b+c} + T_{HoT}^{b+c}), (T_{CoTI}^{a} + T_{CoT}^{a})) + T_{BU}^{a}. \quad (5.15)$$

Next, we consider the handoff delays in TI-MIPv6-SIP and LI-MIPv6-SIP with the ERO option. It is noted that the RR process assume no pre-established security between an MH and its CH(s), regardless of the security relationship between the MH's HA and the CH(s). When the MH's HA and the CH(s) shares a pre-configured mobility security association, the ERO option can be initiated by the HA on receiving a BU for registration update from the MH. In addition to MIPv4, the ERO option is also applicable to MIPv6. Therefore, for Int6 with the ERO option, the handoff delay of a MIPv6 session is determined by the CH binding update involved in either the ERO option following the home registration or the default MIPv6 handoff procedure with the RR, whichever is finished more quickly, i.e.,

$$T_{HO-MIP}^{Int6-ERO} = \min((T_{IP}^{Int6} + T_{BU}^{b} + T_{BU}^{c}), T_{MIP}^{Int6}). \quad (5.16)$$

Furthermore, when an MH and the CH shares a pre-configured mobility association (PMA), the MIP RR can be skipped and the handoff delay is given by

$$T_{HO-MIP}^{Int6-PMA} = T_{IP}^{Int6} + T_{BU}^{a}. \quad (5.17)$$

If the ERO is available, the handoff delay is given by

$$T_{HO-MIP}^{Int6-PMA-ERO} = \min((T_{IP}^{Int6} + T_{BU}^{b} + T_{BU}^{c}), T_{HO-MIP}^{Int6-PMA}). \quad (5.18)$$

B. Handoff delay for SIP sessions (UDP mobility)

For SIP sessions, there are three scenarios. In the first scenario, the MH shares a pre-configured SIP mobility security association (e.g., established in the session-setup stage) with the CH so that the re-INVITE (with proper AAA headers) can be authenticated and authorised by the CH. Therefore, the handoff delay is equal to the sum of new IP address acquisition delay and the one-way delay for the re-INVITE, i.e.,

$$T_{HO-SIP}^{Int6} = T_{IP}^{Int6} + T_{re-INVITE}^{a}. \quad (5.19)$$

Since this pre-configuration can exist in both Int6 and the Hyb6, the SIP session handoff delay in the Hyb6 is similarly given by

$$T_{HO-SIP}^{Hyb6} = T_{IP}^{SIP} + T_{re-INVITE}^{a}. \tag{5.20}$$

In the second scenario, there is no such pre-configured security relationship between the MH and the CH, and the MIPv6 RR process is reused as proposed in Int6. Similar with (5.16), the handoff delay in this case is given by

$$T_{HO-SIP}^{Int6-RR} = T_{IP}^{Int6} + \max((T_{HoTI}^{b+c} + T_{HoT}^{b+c}), (T_{CoTI}^{a} + T_{CoT}^{a})) + T_{re-Invite}^{a}. \tag{5.21}$$

In the third scenario, a SIP-based return routability process analogous to the MIPv6 RR is used as inferred for the Hybrid MIPv6-SIP when no pre-configured security relationship is available. The corresponding handoff delay is expressed by

$$T_{HO-SIP}^{Hyb6-RR} = T_{IP}^{SIP} + \max((T_{Info}^{b+c} + T_{Info}^{b+c}), (T_{Info}^{a} + T_{Info}^{a})) + T_{re-Invite}^{a}. \tag{5.22}$$

In addition, the Int6-ERO, the Int6-PMA and the Int6-PMA-ERO schemes are also applicable to UDP mobility. The equations are similar to (5.17) to (5.19) and are omitted here. Note that an RR process is assumed in the Int6, the Int6-ERO and the Hyb6 for TCP mobility, as shown in (5.15) to (5.17). In contrast, for UDP mobility, protocols with an RR process are explicitly marked as *X-RR*, as shown in (5.22) and (5.23).

Finally, note that the delays to acquire a new valid IP address vary from protocols. In SIP mobility, the application-layer SIP User Agent (UA) has to poll the operating system (OS) to detect an IP address change. The polling interval is usually set to be a few seconds [Schulzrinne and Wedlund 2000]. The maximum value of the notification delay can thus be one polling interval and the mean value is half of the interval as the delay is uniformly distributed. On the other hand, in MIP, the MIP host part in the network layer (normally implemented in the OS) can detect the new IP address immediately after the address is

configured. Thus, the average delay to acquire a new IP address in SIP is equal to the delay in MIP plus the average notification delay $\overline{T_{polling}}$, i.e.,

$$T_{IP}^{SIP} = T_{IP}^{MIP6} + \overline{T_{polling}}. \tag{5.23}$$

Though this delay can be reduced by increasing the polling frequency, the internal signalling costs will increase accordingly and is very inefficient for low mobility where such IP address changes are infrequent. In the Int6, we apply an interruption-style active cross-layer signalling (e.g., using the CLASS scheme proposed in Chapter 3) as the enabler for such notifications from the network layer to the application-layer SIP UA. Consequently, the notification delay becomes negligible, and thus in the Int6 the delay to acquire a new IP address in SIP is same to that in MIPv6 (T_{IP}^{MIP6}). T_{IP}^{MIP6} includes the L2 handoff delay T_{L2} for switching from the old L2 access attachment to the new one, the L3 handoff detection delay T_{L3} for detecting the new access router and the new IP subnet, and the Duplicate Address Detection (DAD) delay T_{DAD} for validating the new IP address, i.e.,

$$T_{IP}^{Int6} = T_{IP}^{MIP6} = T_{L2} + T_{L3} + T_{DAD}. \tag{5.24}$$

In IPv6, either the stateless host auto-configuration [RFC2462] or the DHCPv6 [RFC3315] can serve as the IP address distribution mechanism. In both mechanisms, the involved DAD process is the dominating time-consumer.

5.4.2.2. Session-Setup Delay

The session-setup delay refers to the time elapsed between the instant when a session initiator signals its initial invitation towards a session invitee and the instant when the session setup is completed and thus the session data traffic can be transmitted. We focus on the involved SIP session setup procedures in IPv6 contexts.

To setup a SIP session, a three-way handshake signalling is needed between the session initiator (the CH in this case) and the session invitee (an MH). Assuming that the

home SIP server acts as a proxy server, the session-setup delay in TI-MIPv6-SIP is given by

$$T_{SS}^{TI-MIPv6-SIP} = T_{Invite}^{c} + T_{Invite}^{b} + T_{Ok}^{b} + T_{Ok}^{c} + T_{Ack}^{a}. \quad (5.25)$$

This is also the delay in LI-SYLU as the SIP HS always have the updated location information for an invited MH.

$$T_{SS}^{LI-SYLU6} = T_{SS}^{TI-MIPv6-SIP}. \quad (5.26)$$

On the other hand, in LI-ODLE additional delays are incurred if the cached information for the invited MH has expired, i.e.,

$$T_{SS}^{LI-ODLE6} = T_{SS}^{LI-SYLU6} + P_{expire} \cdot (T_{BRR}^{d} + T_{BU}^{d}). \quad (5.27)$$

In the Hyb6, the SIP session-setup delays are computed by (5.25) as well if the SIP HS receives AOR-based INVITE messages. However, when a CH sends an INVITE towards an NAI-based SIP URI, the SIP server that is located in the domain indicated by the NAI then receives the message, and queries its associated location service for mapping the NAI to the current location of the callee. As the MIP NAI is unlikely to have been registered there, the query would fail and so would the session setup. In another scenario, when receiving an INVITE towards a SIP URI based on an MH's HoA, the SIP server located in the domain determined by the prefix of the HoA (normally this SIP server is the SIP HS for the MH) would encounter the same problem of no matched results. These failures are avoided in the TI-MIP-SIP and the LI-MIP-SIP architectures thanks to the integration and interactions between the MIP HA and the SIP HS. In [Lee etc 2003], a scheme (referred to as INT-HoA) is proposed to enable the SIP HS to forward the HoA-based SIP URI to the MIP HA, which in turn tunnels the INVITE to the CoA of the MH. However, the INT-HoA scheme has a few drawbacks. Firstly, it essentially demands that a SIP proxy server (usually the SIP HS) be located in the same domain where the MIP HA is deployed to receive a HoA-based INVIATE. Secondly, the global tunnelling introduces

additional signalling overhead and the enlarged INVITE takes longer time to reach the MH, and thus incurs longer session-setup delays. Thirdly, it is desirable to populate SIP messages like INVITE among SIP entities, e.g., for application-level routing and AAA purposes. In INT-HoA, the MIP HA tunnels the INVITE to the current address of the MH directly, bypassing any SIP servers that may be desired to process (e.g., record) the INVITE in the transaction.

5.4.3. Handoff Packet Loss Analysis

Handoff packet loss corresponds to the number of packets lost during a handoff. We measure this metric at the MH by labelling each packet sent from the CH with a sequence number and comparing the sequence numbers of the two packets received at the beginning and the end of a handoff.

For an analysis, we consider a UDP-based media streaming flow with a constant packet arrival rate R (in terms of packets/s), and a TCP-based file transfer flow. Both flows are sent from the CH towards the MH. For the UDP application, the handoff packet loss (HPL) is given by the product of the packet arrival rate and the handoff delay using SIP mobility scheme in protocol X, i.e.,

$$HPL_{UDP}^{X} = R \cdot T_{HO-SIP}^{X}. \tag{5.28}$$

For the TCP application, with the MIP mobility scheme in protocol X the HPL can be approximated by [Eom etc 2002],

$$HPL_{TCP}^{X} = \min(\frac{MWS}{RTT} \cdot T_{HO-MIP}^{X}, MWS), \tag{5.29}$$

where MWS is the maximum window size (in terms of packets here) of the TCP connection and RTT is the round trip time (between the CH and the MH) of the TCP connection. In our system model, for MIPv6 (with built-in route optimisation) we assume that the CH has

a valid cache of the MH's binding so that it can send packets directly to the MH, bypassing the HA/HMS. Therefore, the *RTT* is estimated by

$$RTT^{MIP6} = T^{a}_{TCP6data} + T^{a}_{TCP6Ack}, \tag{5.30}$$

where $T^{a}_{TCP6data}$ is the one-way delay for an IPv6-based TCP packet sent from the CH to the MH, and $T^{a}_{TCP6Ack}$ is the one-way delay for an IP-based TCP acknowledgement (ACK) packet sent from the MH to the CH. For MIPv4 (without the RO/ERO option), the TCP packets sent by the CH goes through the triangular routing though the responses from the MH can travel to the CH directly. Thus, the *RTT* is estimated by

$$RTT^{MIP6} = (T^{c}_{TCP4data} + T^{b}_{TCP4data-tunnelled}) + T^{a}_{TCP4Ack}, \tag{5.31}$$

where $T^{c}_{TCP4data}$ is the one-way delay for an IPv4-based TCP packet sent from the CH to the HA, $T^{b}_{TCP4data-tunnelled}$ is the one-way delay for a tunnelled IPv4-based TCP packet sent from the HA to the MH, and $T^{a}_{TCP4Ack}$ is the one-way delay for an IP-based TCP ACK sent from the MH to the CH.

5.4.4. Handoff Reliability Analysis

The proposed ERO option is expected to reduce handoff delay and to increase the handoff reliability at the same time. We evaluate the handoff reliability in terms of successful binding update probability at the CH over networks where the packet loss rate is high. Let p_i denote the packet loss rate of the ith hop, the probability of a successful transmission of a binding update message over each side between the MH, the CH and the HMS or HA is given by

$$P_H = \prod_{i=1}^{H} (1 - p_i), \tag{5.32}$$

where $H = a, b, c$ as defined in Figure 5.12.

In MIPv6 (or SIP), the MH directly informs the CH of its binding update. Hence, the probability that a successful binding update takes exactly K transmissions (including K-1 retransmissions) is

$$p_{MIPv6}^{K} = (1 - P_a)^{K-1} \cdot P_a . \tag{5.33}$$

Accordingly, the average transmission times for a successful BU in MIPv6 are given by

$$E[K]_{MIPv6} = \sum_{K=1}^{\infty} (K \cdot p_{MIPv6}^{K}) = \sum_{K=1}^{\infty} (K \cdot (1 - P_a)^{K-1} \cdot P_a) . \tag{5.34}$$

The corresponding cumulative distribution function (CDF) of is computed by

$$P_{MIPv6}^{K} = \sum_{i=1}^{K} p_{MIPv6}^{i} = \sum_{i=1}^{K} ((1 - P_a)^{i-1} \cdot P_a) . \tag{5.35}$$

When a binding update of MIPv4-RO style is used, it takes two steps and thus at least two transmissions for each end-to-end (MH → HA → CH) binding update. Therefore, the probability that a successful binding update takes exactly K transmissions (K-2 retransmissions) is expressed as

$$p_{MIPv4-RO}^{K} = \sum_{i=1}^{K-1} \{((1 - P_b)^{i-1} \cdot P_b)((1 - P_c)^{K-i-1} \cdot P_c)\} , K \geq 2. \tag{5.36}$$

Similarly, we can also calculate the average transmission times and the CDF.

In the proposed integrated MIPv6-SIP (with the ERO option), both routes (MH → CH and MH → HMS/HA → CH) are taken concurrently and independently for the first time transmission (K = 1). Since this dual BUs will greatly increase the handoff reliability for just one time (0 retransmission) as will be shown, from the second time (if both BUs in the first transmission fail to reach the CH) MIPv6-SIP only sends a BU from the MH to the CH and does not ask the HA to forward the BU to the CH. Thus, the successful probability of transmission for just one time (K = 1) is given by

$$p_{MIPv6+}^{1} = 1 - (1 - p_{MIPv6}^{1})(1 - p_{MIPv4-RO}^{2}) = 1 - (1 - P_a)(1 - P_b P_c) . \tag{5.37}$$

For $K \geq 2$, it is given by

$$p_{MIPv6+}^{K(\geq 2)} = (1 - p_{MIPv6+}^{1})(1 - P_a)^{K-2} P_a \qquad (5.38)$$
$$= P_a(P_a + P_b P_c - P_a P_b P_c)(1 - P_a)^{K-2}, K \geq 2.$$

Thus, the average transmission times and the CDF are given, respectively, by

$$E[K]_{MIPv6+} = \sum_{K=1}^{\infty}(K \cdot p_{MIPv6+}^{K}) = p_{MIPv6+}^{1} + \sum_{K=2}^{\infty}(K \cdot p_{MIPv6+}^{K(\geq 2)}); \qquad (5.39)$$

$$P_{MIPv6+}^{K} = \sum_{i=1}^{K} p_{MIPv6+}^{i}. \qquad (5.40)$$

5.5 Analytical Results

In this section, we present and analyse the numerical results obtained based on the theoretical analyses in Section 5.4.

5.5.1. Signalling Costs

5.5.1.1. Parameter Configuration

To obtain the signalling costs generated in the concerned architectures based on the cost analysis in Section 5.4.1, we apply the typical (default) values listed in Table 5.2 and those in Section 4.5.2 of Chapter 4 as input parameters.

Table 5.2 Parameters for cost analysis

Symbol	*Input Parameter*	*Typical (Default) Value*
v	Mean speed of MHs	100 km/hr
ρ	Density of powered-on MHs	50 /km^2
L_C	Perimeter of a cell (subnet)	10 km
K	Number of rings in a domain	5
P_V	Ratio of visiting MHs in a domain	10%
N_P	Number of powered-on MHs in a domain	$\rho \cdot A_D$
N_V	Number of powered-on visiting MHs in a domain	$N_P \cdot P_V = \rho \cdot A_D \cdot P_V$
λ_R	Refresh rate of home registration at MIP HA or SIP HR or MIP-SIP HMS	2 /hr (home registration lifetime is 1800 sec)
a	Distance between an MH and the CH	20 hops
b, c	Distance between an MH or its CH and the MH's MIP HA (or SIP HS or MIP-SIP HMS)	15 hops
d	Distance between an the MIP HA and the SIP HS	15 hops

In the following, we present the signalling costs results in the IPv6 and IPv4 contexts, respectively. For simplicity, we assume that query results are not cached in LI-ODLE.

5.5.1.2. IPv6 signalling costs

First, with the default mobility rate (λ_M) we investigate the influence of the erlangs (the product of session holding time and session arrival rate) on the signalling costs as shown in Figure 5.13 and Figure 5.14, where the session arrival rate (λ_S) and the session holding time ($1/\mu$) varies alternately.

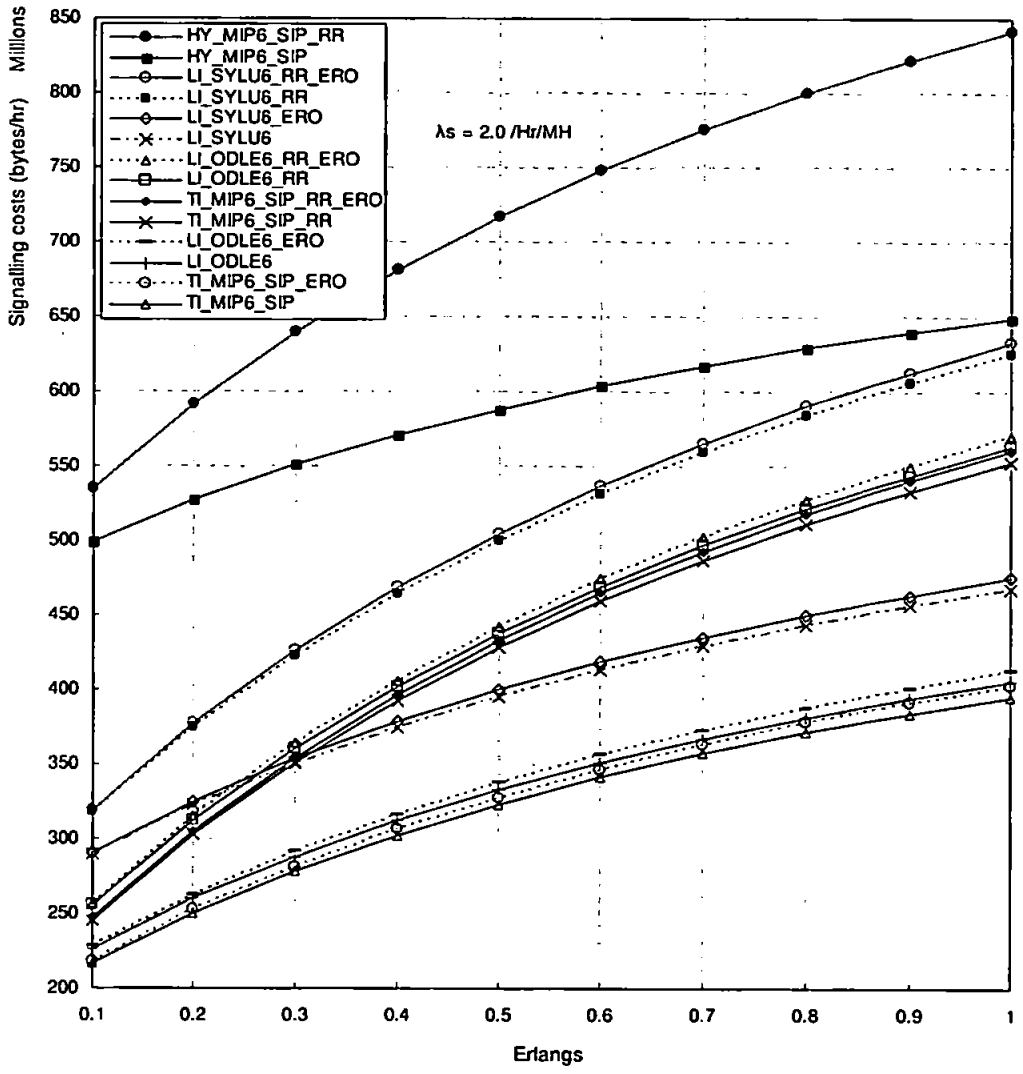


Figure 5.13 Signalling costs vs. erlangs (session holding time is variable) in IPv6 contexts

With the increase of the erlangs, more handoffs take place and thus the signalling costs keep increasing in all the architectures. The costs grow more sharply in Figure 5.14 because the session setup procedures are invoked more frequently. In both cases, HY-MIPv6-SIP and TI-MIPv6-SIP constantly generate the highest and the lowest costs, respectively, whereas the two schemes in LI-MIPv6-SIP also perform greatly as they only provoke slightly more costs than TI-MIPv6-SIP. The integrated architectures can reduce the costs by up to 57% compared to HY-MIPv6-SIP when the erlangs are small (which are more realistic situations) in both cases, though the reduction percentages are more constant when the session arrival rate is fixed as shown in Figure 5.13. LI-ODLE6 performs better than LI-SYLU6 when the erlangs are small (erlangs < 0.8) in Figure 5.14 and regardless of the erlangs in Figure 5.13. These results confirm the fact that when the session arrival rate is low (e.g., at the default value) LI-ODLE6 is more efficient than LI-SYLU6, and vice versa. The underlying reasons for this change are the fact that LI-ODLE6 generates more overheads in session setups whereas LI-SYLU6 incurs more overheads in handoffs and location updates.

Furthermore, the protocols with the RR process (e.g., HY-MIPv6-SIP) generate obviously more costs than their counterparts without the RR process (e.g., HY-MIPv6-SIP-RR) do. On the other hand, the protocols with the ERO option (e.g., TI-MIPv6-SIP-ERO) only invoke insignificantly more costs than their counterparts without the ERO option (e.g., TI-MIPv6-SIP) do. The added costs due to the RR process can reach over 30% whereas only around 1% due to the ERO option. That is why the TI-MIPv6-SIP-RR and the LI-ODLE6-RR protocols can produce more costs than LI-SYLU6 do when the erlangs become larger in the case shown in Figure 5.13. For presentation clarity, the costs in the protocols with the MIP PMA and MIP PMA-ERO options are not shown in Figure 5.13 or Figure 5.14. These costs are slightly higher than the costs in the protocols with the ERO

option only. In addition, the protocols with both the RR process and the ERO option (or/and the MIP PMA option), not shown either, generate slightly more costs than those with the RR process only do.

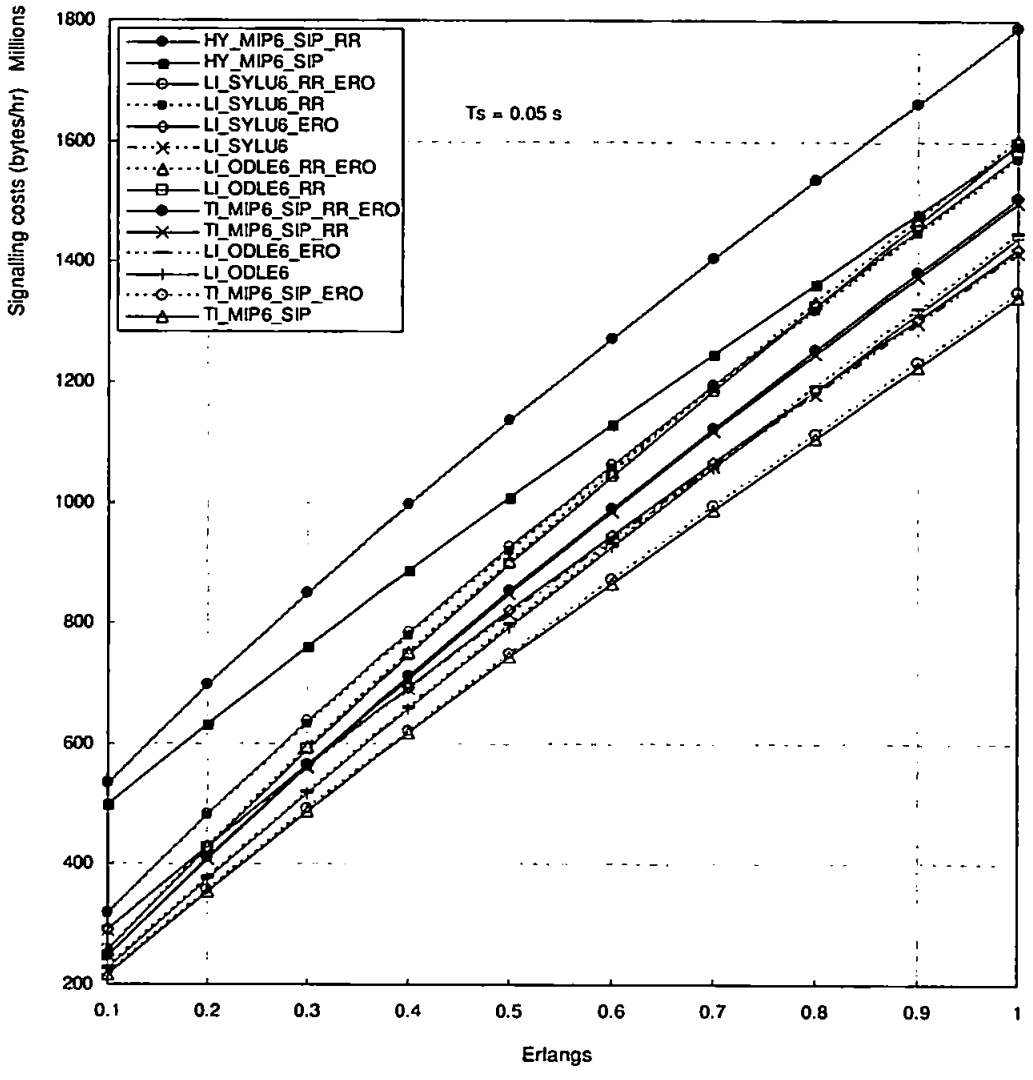


Figure 5.14 Signalling costs vs. erlangs (session arrival rate is variable) in IPv6 contexts

Next, Figure 5.15 and Figure 5.16 demonstrate the effects of CMR on the signalling costs. The concerned CMR ranges from 0.5 to 5.0, whilst the session arrival rate and the mobility rate are fixed using the default value, alternately. The corresponding erlangs are 0.1 in Figure 5.15, and 0.1 to 1.3 in Figure 5.16, respectively.

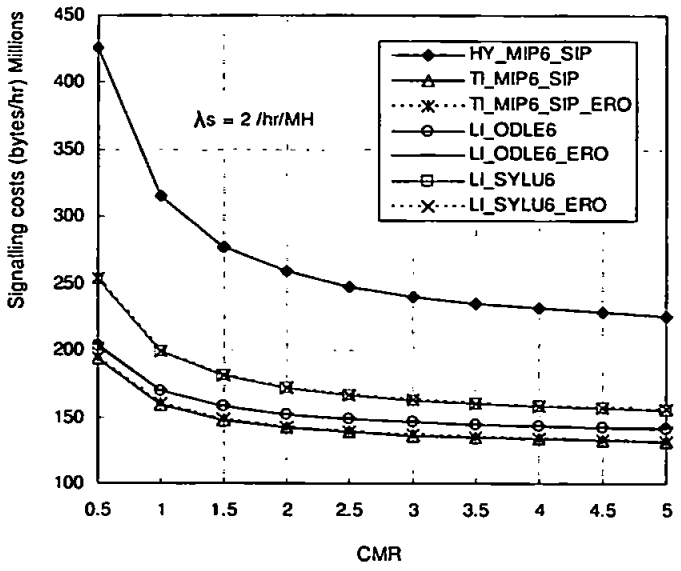


Figure 5.15 Signalling costs vs. CRM (mobility rate is variable) in IPv6 contexts

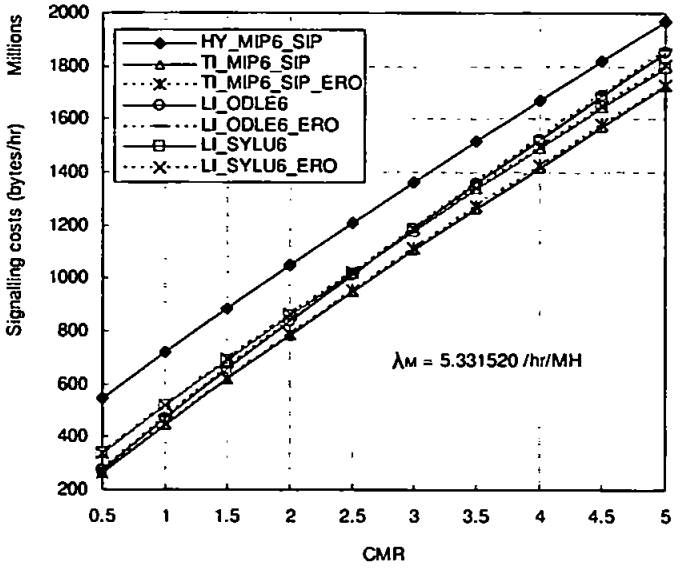


Figure 5.16 Signalling costs vs. CRM (session arrival rate is variable) in IPv6 contexts

As seen from Figure 5.15, the signalling costs decrease with the increase of CMR because the mobility rate declines (i.e., the MHs tend to be more and more static) when the session arrival rate is constant. On the contrary, in Figure 5.16, the signalling costs boost up as the CMR becomes larger since the session arrival rate is rising (i.e., more session

setup costs are incurred) when the mobility rate is fixed. LI-ODLE6 performs better than LI-SYLU6 in most cases with the exception in Figure 5.16 when the CMR is high (CMR > 2.5). In contrast to HY-MIPv6-SIP, when the CMR is small TI-MIPv6-SIP can reduce up to 54% costs in Figure 5.15, or up to 52% in Figure 5.16. LI-MIPv6-SIP, especially LI-ODLE6, can achieve similar results because the introduced interactions are based on the compact MIP messages. Moreover, similar to what have been discussed for Figure 5.13 and Figure 5.14 the protocols with the ERO option (or/and the MIP PMA option, not shown) only generate trivially more costs than their counterparts without the option(s) do.

5.5.1.3. IPv4 signalling costs

Similarly, we study the signalling costs in the IPv4 contexts. The influences of the erlangs are shown in Figure 5.17 (a) and (b), where the session holding time and the session arrival rate is fixed alternately. TI-MIPv4-SIP can reduce the costs by up to 61% compared with HY-MIPv4-SIP when the erlangs are small. Figure 5.18 (a, b) demonstrates the effects of CMR with the mobility rate and the session arrival rate vary alternately. Compared with HY-MIPv4-SIP, TI-MIPv6-SIP can reduce up to 58% costs in (a), or up to 56% in (b) when the CMR is small. Other observations are similar to those in the IPv6 contexts. For example, LI-ODLE4 only produces marginally more costs than TI-MIPv4-SIP does, and performs better than LI-SYLU4 when the CMR is small. Note that there is no RR process involved in the IPv4 contexts and no options are assumed in this study.

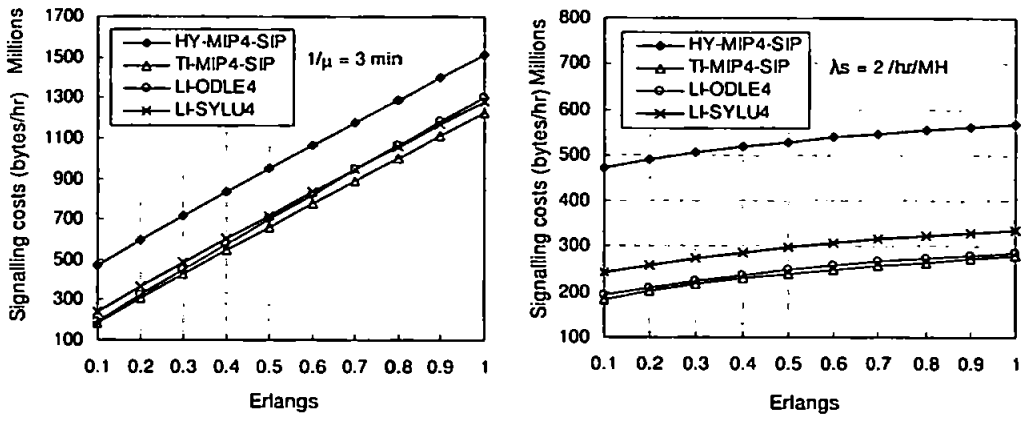


Figure 5.17 Signalling costs vs. erlangs in IPv4 contexts

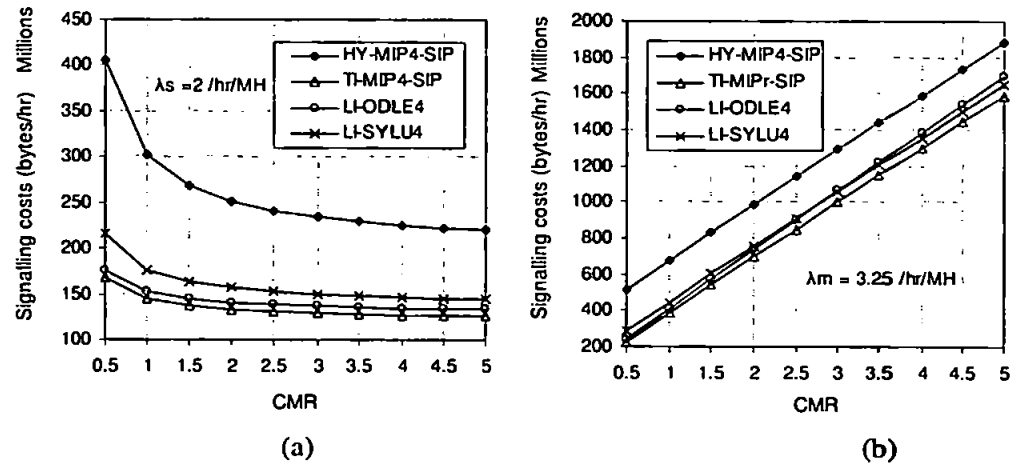


Figure 5.18 Signalling costs vs. CMR in IPv4 contexts

5.5.2. Handoff Delay

To obtain numerical results of delays, we have the following configurations as shown in Table 5.3. The default values are mainly adopted or inferred from the literature [Lo etc 2004, Nakajima etc 2003, Banerjee etc 2004]. The service rate for a SIP message is lower than that of a MIP message at an end host (an MH or a CH) due to the additional application-layer delays, though this difference is negligible for high-speed network

entities like servers and routers. In addition, the default values for the distance parameters are the same as listed in Table 5.2.

Table 5.3 Parameter configurations for delay analysis

Parameter	*Default values*
Number of wireless hops	1 (between MH and the access network)
Bandwidth of a wired hop	100 Mbps
Latency of the wireless hop	2.0 ms
Latency of a wired hop	0.5 ms
Mean service rate of routing an IP message at a router	$5 \cdot 10^6$ messages/s
Mean service rate of a MIP message at HA or HS	5,000 message/s
Mean service rate of a SIP message at HS	5,000 message/s
Mean service rate of a MIP message at an end host	1,000 messages/s
Mean service rate of a SIP message at an end host	400 messages/s
Utilisation (or load) at an end host	0.5
Delay for L2 handoff	12.5 ms
Minimum delay for L3 handoff detection	30.0 ms
Minimum delay to obtain a new valid IPv6 CoA	1,000.0 ms

Given these configurations, the handoff delays are mainly governed by the bandwidth of the wireless link, which is the transmission bottleneck, and the utilisation of the network entities. Figure 5.19 and Figure 5.20 display the corresponding handoff delays, respectively, when the wireless bandwidth or the network utilisation varies alternately whilst the other parameter remains unchanged.

Firstly, we examine the results presented in Figure 5.19, where the network utilisation is fixed at 0.8 whilst the wireless bandwidth varies. With the increase of the wireless bandwidth, the handoff delays decrease sharply in all the protocols. This reflects the fact that the narrowband wireless systems result in significantly higher delays than the wideband ones do. Figure 5.19 (a) provides an overview of the picture. The delays generated in the Hyb6 and the Hyb6-RR protocols for UDP mobility (Hyb6 UDP and Hyb6-RR UDP) are by far higher than those in the other protocols are due to the inefficient OS polling for new IP address detection in the default SIP mobility implementation. The delay details of the other protocols are better shown in Figure 5.19 (b). The Int6 and the Hyb6 for TCP mobility (Int6 TCP and Hyb6 TCP) produce same handoff delays when using the standard MIPv6. The handoff delays in the Int6 for UDP mobility (Int6 UDP)

decrease faster than those in the Int6 TCP and the Hyb6 TCP do, and the Int6 UDP actually performs better except when the bandwidth is very low. This result is desirable as the real-time applications (based on UDP) can be better supported in most cases. The Int6-RR UDP produces higher delays than the Int6 UDP does, though these delays are still much lower than those in the Hyb6 UDP and the Hyb6-RR UDP discussed earlier. Furthermore, the Int6-ERO and the Int6-PMA generate the lowest handoff delays consistently. Note that the delay relation between these two protocols depends on the one-way delays of a BU sent from an MH to the CH directly or via the HA or the HMS. The lower delays generated from them are the delays in the Int6-PMA-ERO, not shown here. It is worth noting that the delays in the Int6-PMA are not necessarily lower than those in the Int6-ERO are.

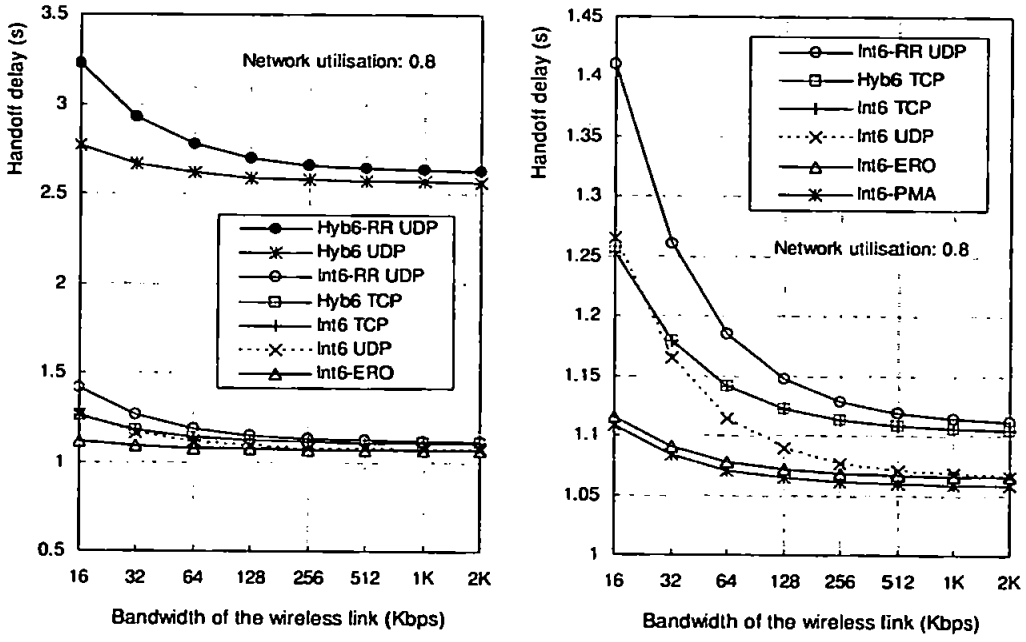


Figure 5.19 Handoff delay vs. wireless bandwidth in IPv6 contexts

Secondly, Figure 5.20 (a) and (b) show the handoff delays of the protocols when the wireless bandwidth is fixed at 128 Kbps whilst the network utilisation varies. Generally, the value relations among them remain the same as shown in Figure 5.19 (a) and (b). For

example, whilst the Int6-ERO and the Int6-PMA produce the lowest delays, the Hyb6 UDP and the Hyb6-RR UDP generates the highest delays, which are displayed in Figure 5.20 (b) separately for clarity. Among these protocols, the Int6 UDP, the Int6-PMA, and the Hyb6 UDP produce almost constant delays regardless of the network utilisation. This is because that in these protocols the handoff binding update message (re-INVITE or BU) is sent from the MH to the CH directly via standard routers, which are far less sensitive to the changes of their utilisations than the home mobility servers are. On the other hand, the handoff delays in the other protocols increase with the growth of the network utilisation. However, these increases are rather small except when the network utilisation becomes near 1.0. In other words, the network utilisation does not affect the handoff delays significantly until the involved servers are almost fully occupied. For this reason, in the following analysis for session-setup delays we only consider the influence of the wireless bandwidth and assume that the network utilisation is 0.8.

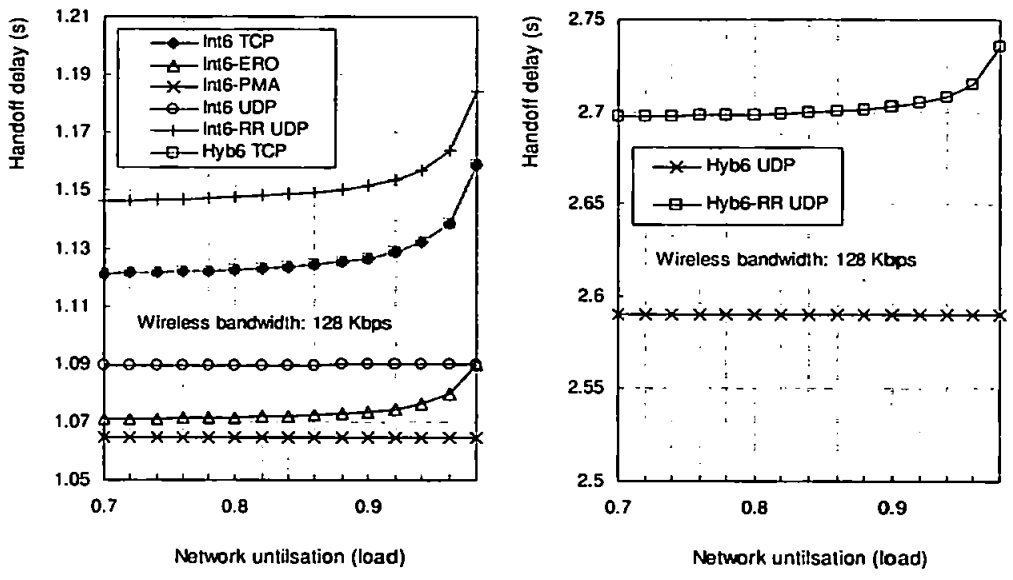


Figure 5.20 Handoff delay vs. network utilisation in IPv6 contexts

5.5.3. Session-Setup Delay

We now move to the results of the session-setup delays, and compare three distinctive approaches, TI-MIPv6-SIP, LI-ODLE6, and INT-HoA6 [Lee etc 2003], as shown in Figure 5.21.

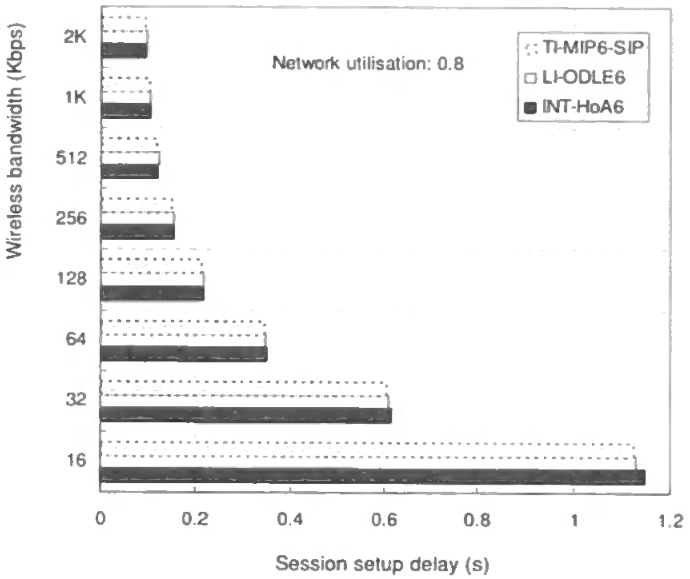


Figure 5.21 Session-setup delay vs. wireless bandwidth in IPv6 contexts

The session-setup delays of all the protocols increase sharply with the decrease of the wireless bandwidth, though for a given wireless bandwidth the differences in these delays of the three protocols are not significant. For a closer examination, TI-MIPv6-SIP consistently invokes the lowest delays, whilst LI-ODLE6 performs better than INT-HoA6 in narrow wireless-bandwidth scenarios because the encapsulated INVITE in INT-HoA6 incurs relatively more delays than the added delays in LI-ODLE6 for location query. For the opposite reasons, INT-HoA6 invokes lower delays than LI-ODLE6 does in wide-bandwidth wireless systems. In contrast to TI-MIPv6-SIP, according to another calculation (not plotted here), the added delays in LI-ODLE6 are under 10 ms and 20 ms when the hops between the MIP HA and SIP HS are 2 and 15, respectively. These two

configurations correspond to the scenarios where the HA and the HS are located in a same domain and in different domains far away from each other, respectively. In practice, theses marginal delays in either scenario can hardly be perceptible to the session initiator, and thus in effect LI-ODLE6 does not deteriorate the session-setup performance.

Furthermore, HY-MIPv6-SIP and LI-SYLU6 provoke the same session-setup delays as TI-MIPv6-SIP does assuming the same service rate for an INVITE message at the SIP servers (including the integrated servers). Therefore, these two protocols are not compared in Figure 5.21. However, HY-MIPv6-SIP can hardly support session invitations based on an HoA or an NAI without extensions like INT-HoA6.

The above observations remain true in the IPv4 context. In addition, when the size of an INVITE message is close to the IPv4 Maximum Transfer Unit (MTU, 576 bytes by default), the encapsulation performed in INT-HoA4 may cause an IP fragmentation at the MIP HA. In that case, the encapsulated INVITE is fragmented to two IP packets, and thus further additional signalling costs and session-setup delay are generated.

5.5.4. Handoff Packet Loss

To obtain numerical results from the analysis, we have the following assumptions, regardless of IPv4 or IPv6. For the UDP application, the packet arrival rate R ranges from 6.25 ~ 50 packets/s, which correspond to real-time flows of 8 ~ 64 Kbps assuming that the IP payload of a packet is 160 bytes (e.g., 160 bytes/packet · 50 packets/s · 8 bits/byte = 64 Kbps). For the TCP application, the sizes of the IP payloads of a TCP data packet and an ACK packet are assumed 552 bytes (512-byte TCP payload, 20-byte standard TCP header plus 20-byte TCP options) and 40 bytes (0-byte TCP payload), respectively. The maximum TCP window size ranges from 2 ~ 32 packets, corresponding to 1 ~ 16 Kbytes. For the sizes of IPv6 and IPv4 packets, 40-byte IPv6 header and 20-byte standard IPv4 header are

added, respectively. The wireless bandwidth is set to be 16 Kbps. Under these configurations in the IPv6 contexts, we demonstrate the analytical results of UDP and TCP handoff packet loss in Figure 5.22 (a) and (b), respectively. The findings in the IPv4 scenarios are similar and thus omitted here.

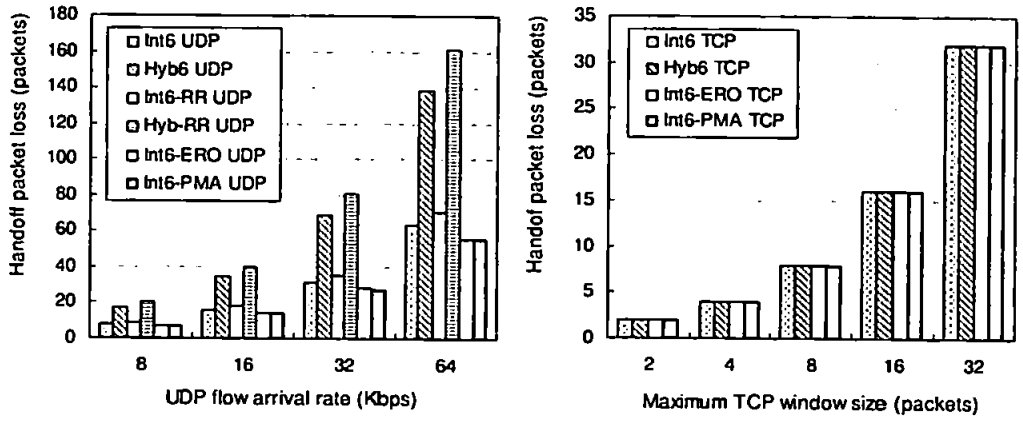


Figure 5.22 Handoff packet loss in IPv6 contexts

For the UDP scenarios, the handoff packet losses in all the protocols increase proportionally to the growth of the flow arrival rates. Not surprisingly, the increases in the Hyb6 protocols are much sharper than those in the Int6 ones are due to the much longer handoff delays. Actually, compared with the Hyb6 and the Hyb6-RR, the Int6 and the Int6-RR protocols can save 59% and 58% packet loss during a handoff, respectively.

On the other hand, in the TCP case the handoff packet losses appear to be determined by the maximum window sizes (MWSs) only because the TCP handoff delays in all the protocols are larger than the round trip time (RTT). Recall (5.30) under this condition, the handoff packet loss in a protocol is equal to the MWS at that instant. Thus, equal handoff losses are expected. Though these lost packets can be recovered by TCP retransmissions later, the retransmissions add data-delivery costs to the system and the packets dropped in the previous domain consume unnecessary system resources. Both the retransmission costs

and the resource consumption are proportional to the handoff packet loss. To reduce the current handoff packet loss, other schemes must be introduced. In fact, this is one of the motivations for our micro-mobility protocol design. Notably, some mechanisms devised in the micro-mobility protocol are applicable to the macro-mobility situations as well. More discussions are performed in Chapter 6.

Furthermore, it is noteworthy that the proposed Int6-ERO, Int6-PMA and Int6-PMA-ERO schemes are useful to macro-handoff designs that can reduce the delay for new IPv6 address configuration to around one-way delay between an MH and the CH. Under that condition, the handoff delays can be lower than the RTT so that the resultant handoff packet loss is a fraction of the MWS, as indicated by (5.30). For instance, the Optimistic DAD scheme [Moore 2005] proposes to skip the DAD process when the probability that an MH fails the DAD is very low. Figure 5.23 (a) and (b) demonstrate the handoff packet loss when the DAD is skipped. Compared with Figure 5.22, the packet loss is further reduced in the UDP case; and what is more, in the TCP case the Int6-ERO and the Int6-PMA this time can reduce the packet loss by half in contrast to the Int6 TCP and the Hyb6 TCP, whose handoff delays are still higher than the RTT.

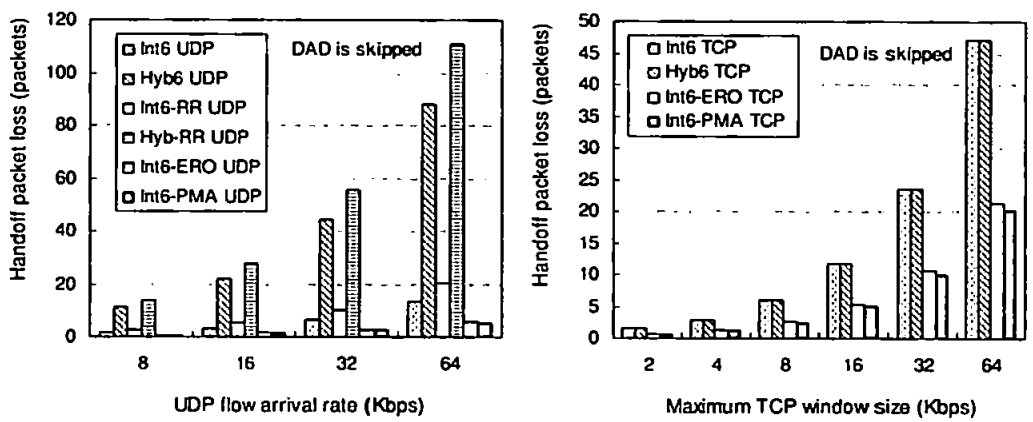
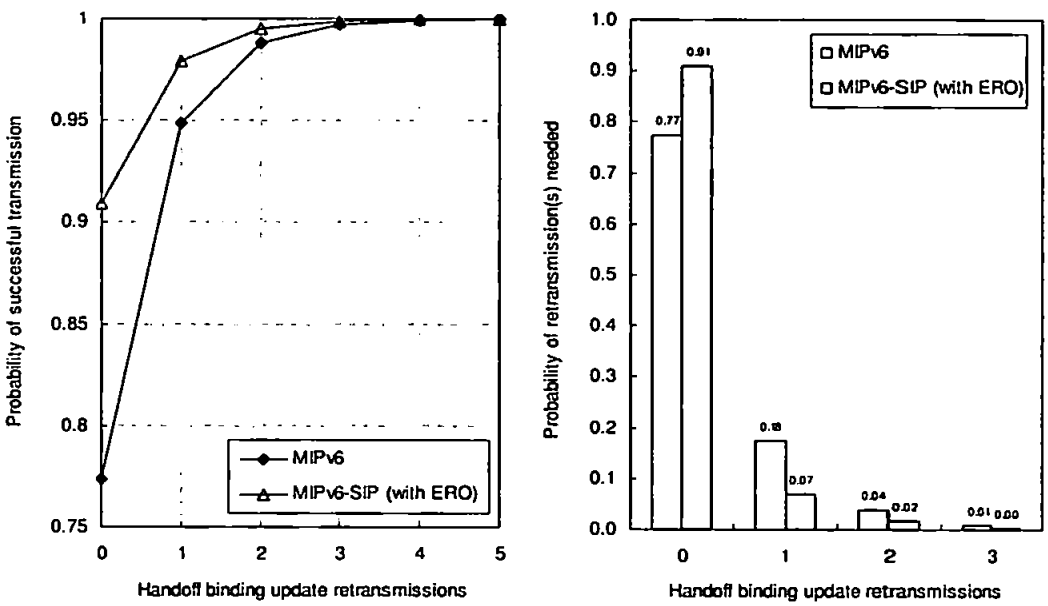


Figure 5.23 Handoff packet loss in IPv6 contexts (DAD is skipped)

In addition, in some designs (e.g., [Faccin etc 2004]), MIPv6 alone is used for universal mobility management and both TCP and UDP handoffs are dealt with MIPv6. In these designs, the Int6-ERO, Int6-PMA and Int6-PMA-ERO schemes can also directly reduce the UDP handoff packet losses, which are proportional to the corresponding reduced handoff delays.

5.5.5. Handoff Reliability

To give an example of the improvements in handoff reliability, we assign a, b, c = 5 and the average packet loss rate over each hop is 0.05. Figure 5.24 demonstrates the results in the proposed integrated MIPv6-SIP (with ERO option) and in MIPv6. Zero retransmission means no retransmission is needed.



(a) (b)

Figure 5.24 Handoff reliability

According to Figure 5.24(a), the probability of successful transmission of the binding update in each protocol increases with the number of retransmission attempts. However, for a given retransmission time, the success probability is significantly higher in MIPv6-

SIP. This results in lower probability in retransmission requirements as shown in Figure 5.24(b). Notably, in both MIP and SIP, the retransmission interval is exponentially configured and thus each retransmission attempt dramatically increases the handoff delay. Therefore, by reducing the probability of retransmissions, the integrated MIPv6-SIP (with ERO option) can help to reduce the handoff delay considerably compared with MIPv6.

5.5.6. Summary of Analytical Results

Finally, before we conclude this section we summarise the major findings in the analytical evaluation by presenting the results in a qualitative way in Table 5.4.

Table 5.4 Comparison of macro-mobility protocols based on joint MIP and SIP

	Hybrid MIP-SIP	*TI-MIP-SIP*	*LI-MIP-SIP ODLE*	*LI-MIP-SIP SYLU*
Signalling costs	Very high	Lowest	Very low	Low
Handoff delay	UDP: high TCP: normal	UDP: low TCP: low (with option, e.g., ERO)	UDP: low TCP: low (with option, e.g., ERO)	UDP: low TCP: low (with option, e.g., ERO)
Session-setup delay	Normal with limited SS functionality	Normal with extended SS functionality	Trivially higher than normal with extended SS functionality	Normal with optional extended SS functionality [a]
Handoff packet loss	UDP: high TCP: high	UDP: low TCP: lower [b]	UDP: low TCP: lower [b]	UDP: low TCP: lower [b]
Handoff reliability	Normal	Higher (with ERO)	Higher (with ERO)	Higher (with ERO)
Server location requirements	No [c]	MIP HA and SIP HS converged in the home domain	No [c]	No [c]
Deployment	Promptest or temporary deployment	Long-term deployment	Prompt deployment, esp. beneficial in small CMR situations	Prompt deployment esp. beneficial in large CMR situations

a. It is optional to enhance a SIP HS with the HoA and NAI records in LI-SYLU

b. When the DAD and L2 handoff delays are negligible, and the wireless bandwidth is narrow

c. Though the MIP HA and the SIP HS usually located in the same home domain

To sum up, the HY-MIP-SIP architectures are best in deployment promptness whereas worst in system performances; in contrast, the TI-MIP-SIP architecture outperforms all the other architectures as the most cost-efficient approach despite its deployment difficulty in

the short term. The LI-MIP-SIP architecture is a trade-off design of these two extremes, which achieves sub-optimal efficiency at the cost of mild modifications on the protocol operation. Note that the SIP standard does not define the interactions between an HS and its location services, and thus the LI-MIP-SIP architecture can be deemed as an enhanced scenario where SIP utilises MIP as a location service. From this perspective, the requirement for a SIP HS to employ MIP location management messages for location service may be deemed as a natural enhancement to SIP.

5.6 Simulation Results

To complement the analytical results, simulations are performed. This section presents the simulation results and discussions.

5.6.1. Simulation Configurations

5.6.1.1. Simulation Scenarios and Configurations

The simulation software is OPNET Modeller 11.0. Figure 5.25 illustrates the network layout in the simulations. The MH is initially located in its home domain and managed by its HMS (Home Mobility Server). When the simulation is started, the MH stays for 90 s in its home domain. Afterwards, it moves in an anti-clockwise way, passes by three foreign access routers (FRs) administrated by three foreign domains serially before finally returning home. The three FRs are denoted by FR1, FR2 and FR3, respectively. The MH stays for 60 s whenever it reaches near an FR. The moving speed of the MH is 60 mile/hr (97 Km/hr), the moving direction follows a roughly straight line between two adjacent FRs, and the radius of each subnet is 1000 m. Therefore, it can be derived that in each simulation four handoffs occur at an interval of about 134 s (though the actual handoff intervals range from 114 to 144 s because the trajectory is not a square). The simulated time of the whole process in each scenario is 10 min. All the wired links are OC48 (2.5

Gbps) links and thus the transmission delay along these links are negligible. Therefore, we can affect the end-to-end delays across domains by varying the core network delay only, and thus we can analyse the delay-related metrics more conveniently. The HMS and the FRs are also IEEE 802.11b WLAN access points. Two CHs, CH1 and CH2, are located in the third foreign domain. CH1 is a static wireless host running a video conference (real-time application), whilst CH2 is a wired FTP server (non-real-time application). Simulations are repeated in each scenario and the averaged results are collected.

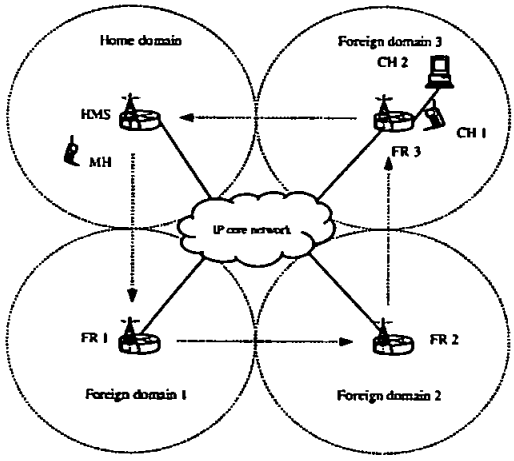


Figure 5.25 Simulation network layout

5.6.1.2. Evaluation Metrics

The following metrics are defined for the evaluation:

- Protocol handoff delay: Time elapsed between an L3 handoff protocol is triggered by new router detection to the time the CH (or the HA if RO is disabled) is notified of the MH's new CoA so that subsequent session packets can be rerouted correctly.
- Handoff packet loss reduction: The reduction in packet loss during a handoff when two handoff protocols are compared with each other. It is obtained by subtracting the handoff packet loss occurred in one handoff protocol (the proposed protocol) from that in another protocol (the one to be compared with).

- End-to-end delay: Time elapsed between a packet is sent out by a calling party to the time the packet reaches a called party. This statistic is collected on a caller basis.
- Delay variation: Variation among end-to-end delays for packets received by this node. End-to-end delay for a packet is measured from the time it is created to the time it is received. This statistic is collected on a per caller basis.
- TCP retransmissions: Number of TCP retransmissions on this node. Written when data is retransmitted from the TCP unacknowledged buffer.

5.6.2. Performance Comparison

Wherever appropriate, we compare the proposed integrated MIPv6-SIP approach (including TI-MIPv6-SIP and LI-MIPv6-SIP, briefly MIPv6-SIP) with one or more of the following protocols: MIPv6 with RO (MIPv6 w RO), MIPv6 without RO (MIPv6 w/o RO), and an optimised MIPv6 (oMIPv6) which skips the DAD process in MIPv6 w RO.

5.6.2.1. Simulation Setting for the Video Conference Application

The simulation configurations for the video conference are tabulated in Table 5.5.

Table 5.5 Simulation configurations: video conference

Parameter	Value
Video conference frame size	200 bytes
Video frame inter-arrival time	Default value: Constant(0.1)s, i.e., 10 frames/s
WLAN data rate	1 Mbps
Number of lost RAs that constitute an L3 handoff indication	2
Interval between two consecutive RAs	Uniform distribution on [0.1, 0.5] s
DAD delay	Uniform distribution on [1, 1.5] s
Core network delay	Default value: 0.1 s
Application start time	Uniform distribution on [50, 60] s
Application end time	End of the simulation

5.6.2.2. Protocol Handoff Delay

As aforementioned, four consecutive handoffs take place in each simulation. The simulation configurations actually enable four kinds of inter-domain handoffs. The first one is from the MH's home domain to a foreign domain, the second one is from one

foreign domain to another, the third one is from a foreign domain to the CH's domain (which is still foreign to the MH), and the fourth one is from the CH's domain to the MH's home domain. Let us denote these four kinds of handoffs as H->F1, F1->F2, F2->F3&Hc, and F3&Hc->H, respectively. Figure 5.26, Figure 5.27, and Figure 5.28 show the protocol handoff delays in MIPv6 w RO, oMIPv6 and MIPv6-SIP, respectively, with the increase of the core network delays. The changes of core network delays can stand for scenarios of different scaled core networks or routes changes in a certain core network (e.g., due to network congestion).

When Figure 5.26 is concerned, it appears that the protocol handoff delays in MIPv6 w RO increase slowly with the increase of the core network delays. The increase is because that the transmission delay for macro handoff signalling becomes higher when the delay along the route is higher. The slowness is because that the DAD delay is so large that it overshadows the relatively low increase in the signalling delay.

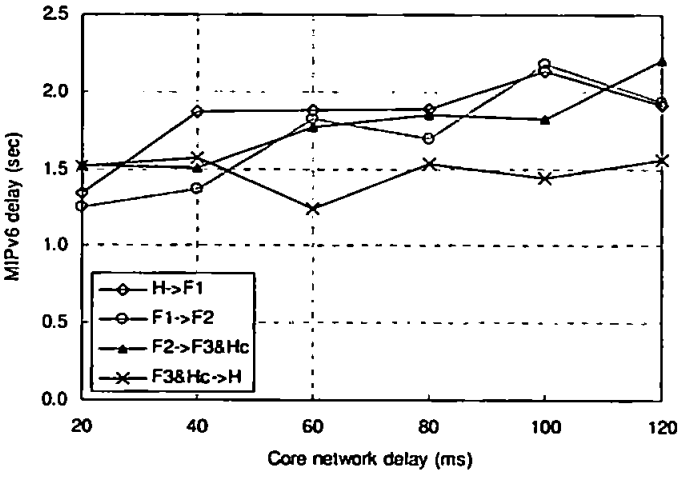


Figure 5.26 MIPv6 w RO handoff delay

Next, we consider the oMIPv6 handoff delays as shown in Figure 5.27. In the H->F1 and F1->F2 scenarios, the MH visits a foreign domain that is foreign to the CH as well. In both cases, the RR and the CH binding take the longest time, and thus the handoff delays

are higher than the other two cases. In the F2->F3&Hc scenario, the RR and the CH binding processes can be completed more quickly as the MH and the CH are in the same domain. In the F3->Hc->H case, the MH deregisters its CoA at the HA/HMS and the CH on returning home domain. The RR tests are finished very quickly as the MH is at home because both the home tests and the care-of tests take an RTT (round trip time) to finish almost simultaneously. Thus, the total protocol handoff delay is roughly 1.5 RTTs or 3 one-way end-to-end delays between the MH and the CH. For example, when the core network delay is 100 ms, the protocol handoff delay is about 100 ms * 3 = 300 ms, i.e., 0.3 s as shown in the figure.

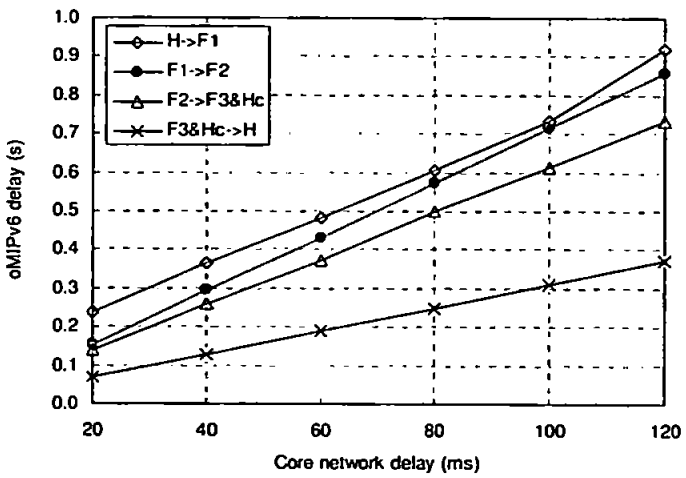


Figure 5.27 oMIPv6 handoff delay

In MIPv6-SIP, the MH performs SIP-style CH binding directly upon a handoff as a security association has been established e.g., during session setup stage, and thus the protocol handoff delay roughly corresponds to the one-way end-to-end delay from the MH to the CH except in the F2->F3&Hc case. In the F2->F3&Hc scenario, the MH and the CH are in the same domain, thus the end-to-end delays between them are lowest compared to other scenarios as shown in Figure 5.28.

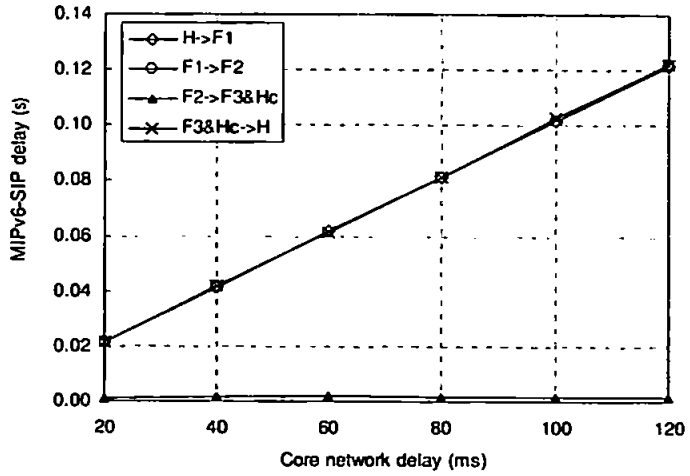


Figure 5.28 MIPv6-SIP handoff delay

To compare the protocol handoff delays, we take the delays in the F1->F2 scenario as an example as illustrated in Figure 5.29.

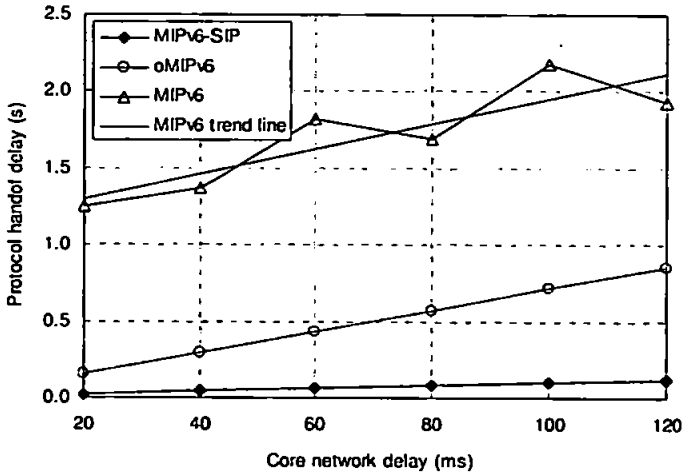


Figure 5.29 Comparison of protocol handoff delay (F1->F2 case)

Clearly, the proposed MIPv6-SIP yields the lowest delays consistently whereas the MIPv6 w RO generates the highest delays. The differences between the MIPv6 trend line and the oMIPv6 are around 1.25 s, corresponding to the mean value of the uniform distribution of the DAD delay over [1 s, 1.5 s]. The protocol handoff delays in MIPv6 w/o RO are similar to those in the integrated MIPv6-SIP since on a handoff the MH notifies its HA directly so that the HA can tunnel the following packets to the MH's new CoA.

Consequently, the handoff packet losses in these two protocols are also similar and thus are not compared here.

5.6.2.3. Handoff Packet Loss Reduction

Thanks to the reduced handoff delays in the proposed MIPv6-SIP approach, the handoff packet loss is reduced accordingly. Figure 5.19 demonstrates the handoff packet loss reduction when MIPv6-SIP compares with oMIPv6 and MIPv6 w RO. With the increase the video frame arrival rate, the reductions increase proportionally. The reductions in MIPv6 w RO are more significant since the handoff delays are higher in MIPv6 w RO than those in oMIPv6 are.

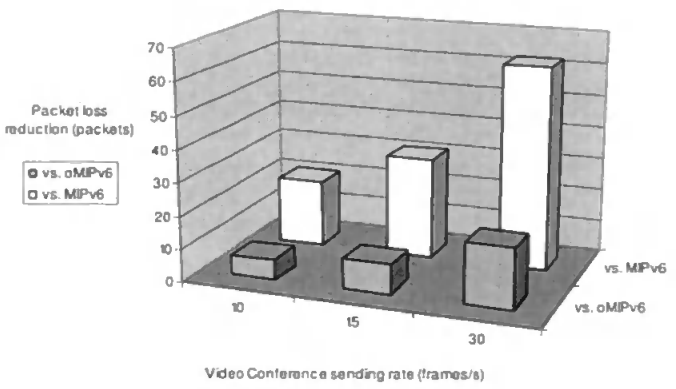


Figure 5.30 Comparison of handoff packet loss reduction (in the F1->F2 case)

5.6.2.4. End-to-End Delay

Figure 5.31 and Figure 5.32 show the end-to-end delay of the video conference packets from the CH to the MH and vice versa, respectively. Generally, MIPv6 w/o RO yields longer end-to-end delays than those in the other protocols except when the MH is in the home domain (0 ~ 144 s and 540 ~ 600 s). As to the other protocols, they tend to result in comparable end-to-end delays overall though the delays in MIPv6 w RO are larger than those in MIPv6-SIP and oMIPv6 upon handoffs, which occurred at about 144, 258, 402

and 540 s, respectively. In addition, it is found in the simulations that the service disruption time on the second handoff is the largest among the four handoffs, due to the slowest router detection in the foreign to foreign inter-domain handoff.

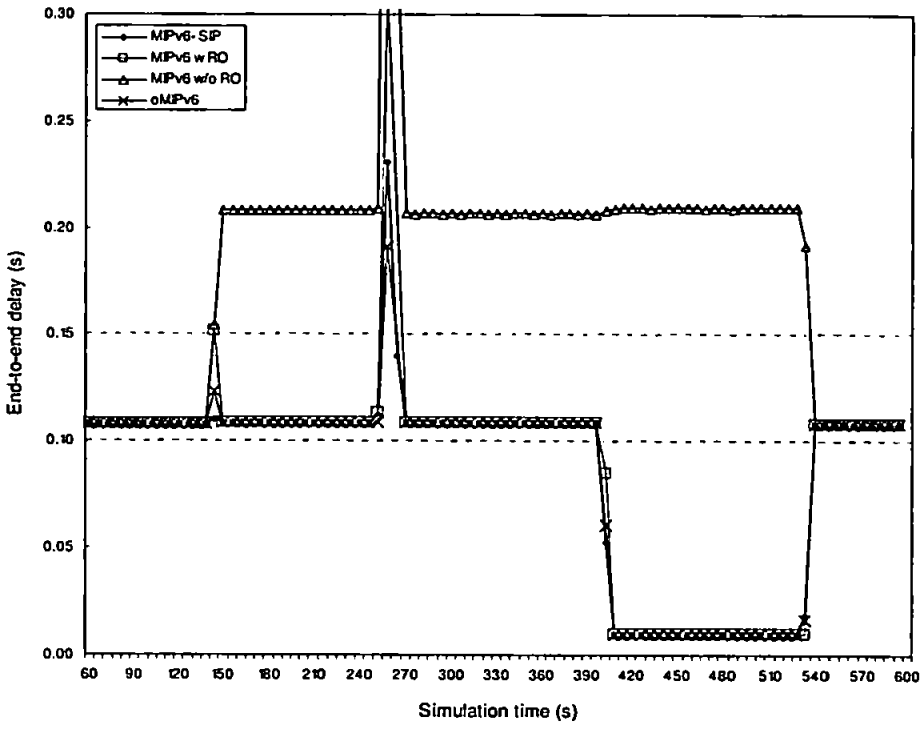


Figure 5.31 Comparison of CH to MH end-to-end delay

In MIPv6 w/o RO, bi-directional tunnelling is used. Therefore, the traffic between the MH and the CH in each direction has to pass by the HA. Unless the MH is at home domain, the end-to-end data delivery follows a triangular route. When the MH is at home, the end-to-end delay can be approximated by one core network delay, i.e., about 0.11 s; when it is away, this delay doubles, i.e., about 0.21 s. Since real-time applications usually require a bounded end-to-end delay (e.g., [ITU114]), MIPv6 w/o RO is not a good mobility support candidate for real-time applications running in a mobile environment.

In other protocols, the end-to-end delays correspond to one-way delay between the MH and the CH thanks to the RO. When the MH and the CH is not in the same domain,

this delay is roughly equal to one core network delay; otherwise, this delay is negligible (about 0.01 s) since the MH and the CH communicate with each other locally and the traffic between them does not traverse the core network at all.

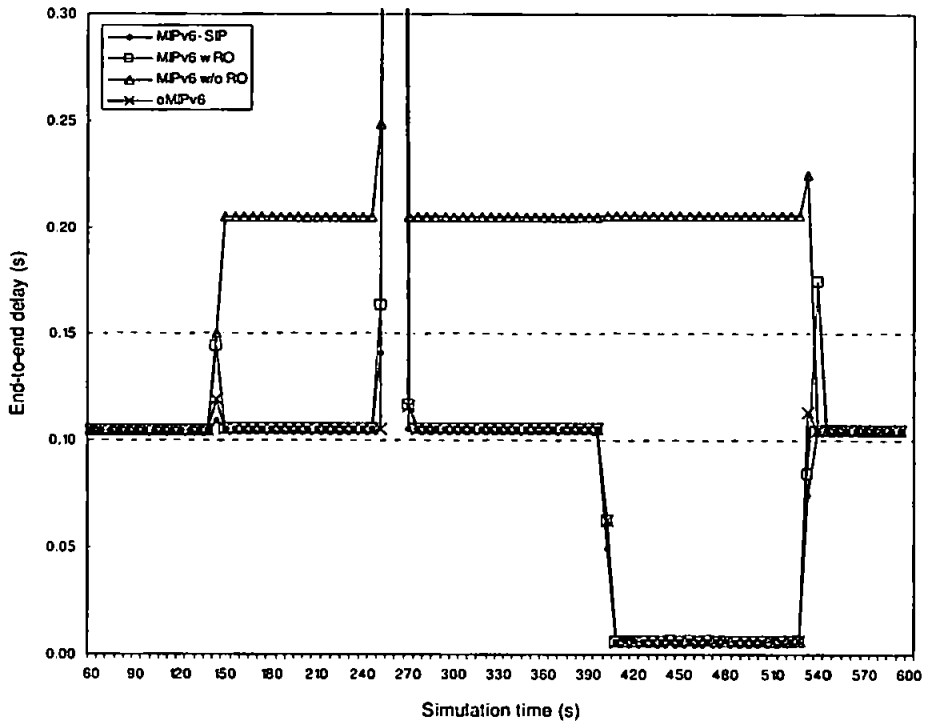


Figure 5.32 Comparison of MH to CH end-to-end delay

5.6.2.5. Delay Variation

The delay variations are measured at both the MH and the CH sides and are illustrated in Figure 5.33 and Figure 5.34, respectively.

Firstly, we analyse the MH side performances. Generally, the delay variations increases sharply during a handoff due to the service disruption and the variations start to drop on the completion of the handoff. For RO-enabled protocols, the sharpest increase starts upon the third handoff when the MH enters the CH's domain because the end-to-end delay changes from the normal value (0.11 s) to a negligible one (0.01 s). On the other hand, for MIPv6 w/o RO, the sharpest increase happens on the first handoff when the MH

moves out of its home domain since the end-to-end delay changes from 0.11 s to 0.21 s. A similar sharp change also happens on the fourth handoff when the MH returns home.

Next, we discuss the delay variations at the CH side. The largest change in delay variation occurs on the second handoff, where the MH moves from one foreign domain to another, and both of the involved domains are foreign to the CH, too. This is because that from the CH's perspective only the MH's second handoff causes a significantly noticeable service disruption. The delay variation begins to decrease on the completion of that handoff.

In both cases, the delay variations in MIPv6 (w or w/o RO) are considerably higher than those in the integrated MIPv6-SIP and oMIPv6 are. Regarding the integrated MIPv6-SIP and oMIPv6, their delay variations are comparable at the MH side whilst those in the integrated MIPv6-SIP are significantly lower than oMIPv6's at the CH side.

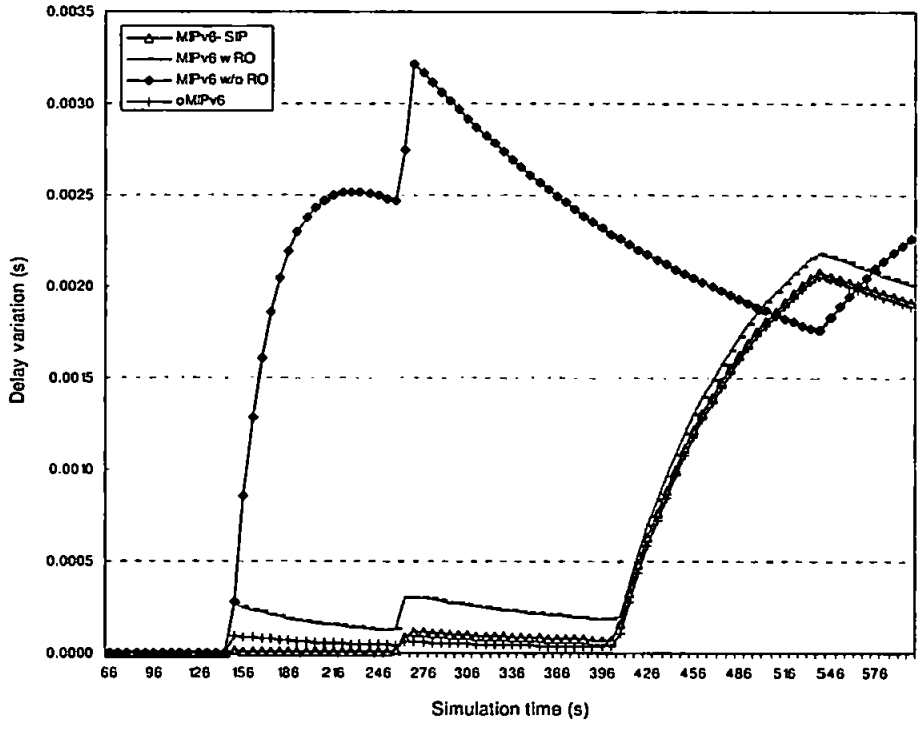


Figure 5.33 Comparison of delay variation at the MH

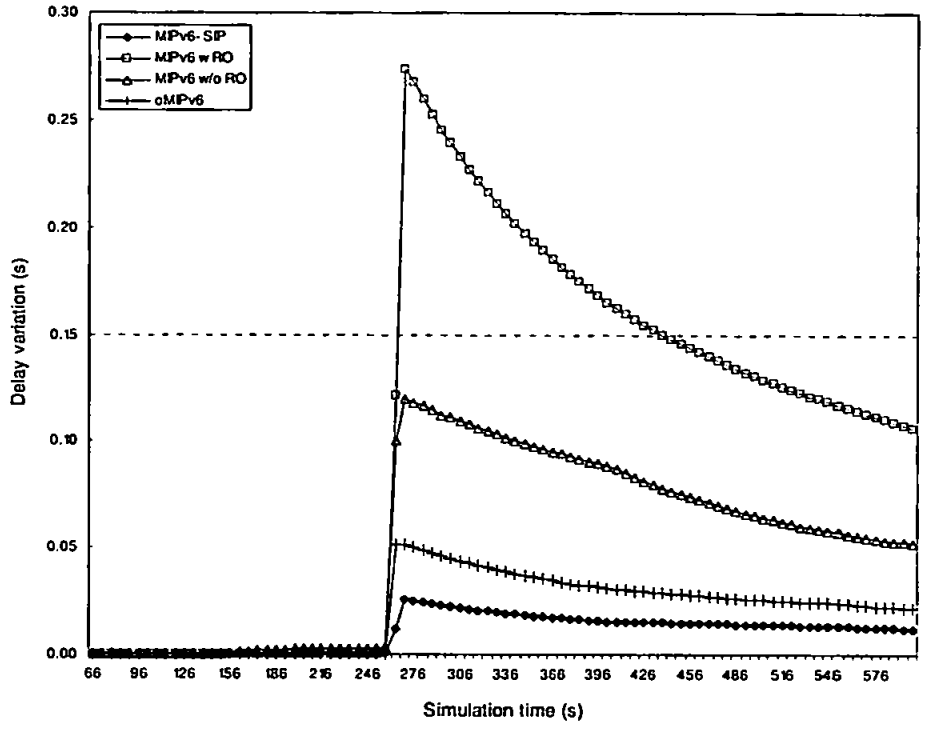


Figure 5.34 Comparison of delay variation at the CH

5.6.2.6. TCP Retransmissions

Simulations are also performed to evaluate the mobility support performances of the non-real-time application FTP downloading from the FTP server (CH2). Four successive FTP downloading operations are arranged and each begins before the corresponding handoff and ends after the handoff. The configurations are listed in Table 5.6.

Table 5.6 Simulation configurations: FTP

Parameter	Value
FTP file size	1 Mbytes
FTP inter-request time	Constant(120 s)
WLAN data rate	1 Mbps
Number of lost RAs that constitute an L3 handoff indication	2
Interval between two consecutive RAs	Uniform distribution on [0.1, 0.5] s
DAD delay	Uniform distribution on [1, 1.5] s
Application start time	120 s

The total TCP retransmissions performed by both the MH and the FTP server are collected during a series of simulations in MIPv6-SIP (utilising oMIPv6) and the standard

MIPv6 (MIPv6 w RO), and the results are shown in Figure 5.35. We can find that almost in all the cases (expect the fourth simulation) MIPv6-SIP invokes less retransmissions than MIPv6 does thanks to its shorter handoff delays. The average retransmissions in MIPv6-SIP and MIPv6 are 31.3 and 34.0, respectively. Thus, MIPv6-SIP also appears a better solution for TCP non-real-time applications by reducing 7.8% retransmissions.

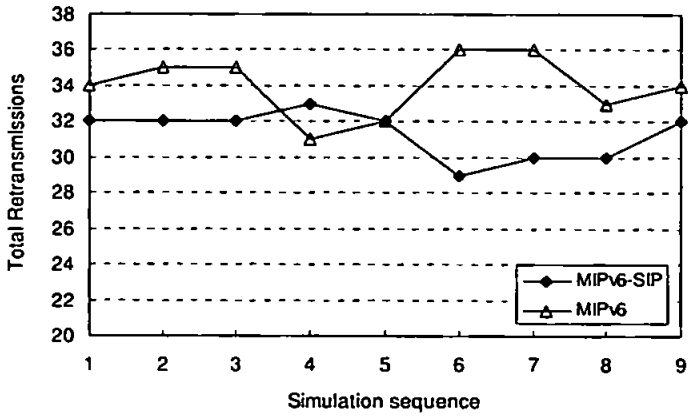


Figure 5.35 Comparison of TCP retransmissions

5.6.2.7. Performance Comparison Summary

Table 5.7 summarises the comparisons, mainly based on the discussed simulation results. Overall, the proposed integrated MIPv6-SIP architectures (TI/LI-MIPv6-SIP) outperform the other approaches in terms of improved handoff performances and advanced mobility support. Thus, it appears as a promising mobility support solution for both real-time and non-real-time applications.

Table 5.7 Performance comparison summary

	TI/LI-MIPv6-SIP	*oMIPv6*	*MIPv6 w RO*	*MIPv6 w/o RO*
Protocol handoff delay	lowest	low	high	Comparable to TI/LI-MIPv6-SIP
Handoff packet loss	lowest	low	high	Comparable to TI/LI-MIPv6-SIP
End-to-end delay	low	low	low	highest
Delay variation	low	low	high	high
TCP retransmissions	fewer	Same as TI/LI-MIPv6-SIP	more	N/A
Advanced capabilities e.g., session renegotiation	Yes	No	No	No

5.7 Concluding Remarks

In this chapter, we have proposed the Loosely Integrated MIP-SIP architecture (LI-MIP-SIP) as an alternative macro-mobility architecture to the Tightly Integrated MIP-SIP architecture (TI-MIP-SIP) proposed in Chapter 4. Two protocols are devised to establish the desired interactions between MIP and SIP servers for efficient and enhanced mobility support: in the LI-ODLE protocol, only MIP HA tracks the location of an MH, and SIP HS uses MIP HA as a location service; in the LI-SYLU protocol, MIP HA updates SIP HS on behalf of an MH. Both theoretical analyses and simulations are conducted to evaluate the proposed architectures, LI-MIP-SIP and TI-MIP-SIP, under a set of metrics.

The analytical results show that the LI-MIP-SIP and the TI-MIP-SIP architectures, together with the useful options such as the ERO (Enhanced Route Optimisation) option, improve the system performances significantly by reducing the signalling costs, handoff delay, and handoff packet loss compared with the traditional hybrid MIP-SIP approach. Furthermore, the system functionality is also extended by introducing enhancements such as the extended support for SIP session setup and the SS (session setup) option for MIP sessions. The enhancements for the session setup procedure facilitate the network to track a called user more effectively since a user can be identified by either SIP or MIP identifiers.

Simulations are also designed and performed with OPNET to evaluate the performances of the LI-MIPv6-SIP and TI-MIPv6-SIP architectures, compared with standard MIPv6 and its variants. The overall simulation results demonstrate that the proposed architectures outperform other approaches in supporting macro handoffs for both real-time and non-real-time applications.

Surely, there is an overhead to implement the protocol integration or interactions through the proposed designs. We presume that the added complexity to the system is

outweighed by the performance improvements and functionality enhancements. From a network operator's view, the great reduction in signalling overheads increases the scalability of the whole network and thus decreases the maintenance costs. From a service provider's perspective, the improved quality of service caters for more subscribers and thus generates more revenue. From a user's standpoint, he/she would like to have a better roaming experience when engaged in live real-time and/or non-real-time applications.

Regarding the two proposed architectures, the TI-MIP-SIP approach can prove more cost-efficient in a long run and thus it is suitable for a future-generation deployment. In contrast, the LI-MIP-SIP architecture, especially the LI-ODLE protocol, may be preferred in the near-future stage since this approach does not modify the physical entities or constrain the physical locations of the entities whilst being capable of achieving similar performance improvements and functionality enhancements.

Chapter 6

The Optimised Micro-Mobility Architecture

In this chapter, we propose and evaluate an IP-centred micro-mobility architecture, based on an optimisation and integration of hierarchical Mobile IP and fast handoffs. This chapter is partially based on a publication [Wang and Abu-Rgheff 3G2005].

6.1 Introduction

As we have discussed in previous chapters, MIP (MIPv4 [RFC3344] and MIPv6 [RFC3775]) and SIP [RFC3261] are the two dominant mobility management protocols for IP applications, and they can cooperate with each other in supporting various mobility scenarios. In general, both MIP and SIP are macro-mobility protocols, relying on location tracking through a MIP HA or SIP home servers in the home domain of an MH. Since the home domain is typically far away from the foreign domain the MH is visiting, mobility messages have to traverse globally on each handoff (or location update when the MH is in the idle mode), which leads to laggard response to user mobility and huge traffic burden on the core network as well as the home domain. The situation is aggravated in the 3G and beyond systems, where micro and pico cells are introduced to increase the system capacity, and thus handoffs occur more frequently.

Therefore, a number of micro-mobility protocols have been proposed for the last few years. These protocols can be largely classified into two categories [Campbell and

Castellanos 2000, Campbell etc 2002, Akyildiz etc 2004]: host-specific protocols represented by Cellular IP [Campbell etc 2000, Shelby etc 2001] and HAWAII [Ramjee etc 2002 A], and tunnelling-based ones represented by Hierarchical MIPv6 (HMIPv6) [RFC4140] and MIPv4 Regional Registrations (MIPv4-RR) [Gustafsson etc 2004]. As discussed in Chapter 2, the differences between these two approaches mainly lie in the deployment considerations since they yield similar performances. As far as deployment is concerned, the tunnelling approach is advantageous because it does not require that the intermediate routers (routers located between the gateway and the ARs) are mobility-aware as the host-specific approach does. Thus, we propose to exploit the tunnelling approach for its deployment advantage.

By introducing virtual home mobility entities such as MAP (Mobility Anchor Point) and GFA (Gateway Foreign Agent) locally, HMIPv6 and MIPv4-RR can quickly respond to intra-domain mobility and largely confine mobility signalling within the domain, and thus can expedite handoffs and reduce global signalling overhead. In addition, fast handoff schemes such as Fast Handovers for MIPv6 (FMIPv6) [RFC4068] and Low Latency Handoffs in MIPv4 (LL-MIPv4) [Malki 2004] have been designed to expedite the L3 handoff by exploiting L2 triggers. All the mentioned protocols are reviewed in Chapter 2.

Regarding IPv6-based mobility protocols, FMIPv6 operates over MIPv6 by default and thus a costly global location update at the HA is performed on each subnet crossing even within a domain. Therefore, an integration of both FMIPv6 and HMIPv6 could combine their complementary merits. However, existing FMIPv6 over HMIPv6 schemes do not seem cost-effective or suitable for large domains. Moreover, there is a requirement for QoS support in the mobile environments [RFC3583], and hence interworking between mobility management and QoS protocols is needed. Nevertheless, an optimisation of such interworking is still missing. More discussions are provided in Chapter 2.

Motivated by the above observations, among others, we propose an efficient and fast micro-mobility architecture for all IP networks, focusing on IPv6. The micro mobility is achieved by dynamically integrating HMIPv6 and FMIPv6 with a two-phased handoff scheme. A number of fast and smooth handoffs take place along an extended QoS route in the first phase whilst an optimised QoS route is dynamically performed in the second phase. The remainder of the chapter is organised as follows. Section 6.2 expounds the design of the proposed micro mobility architecture, followed by an interworking with the proposed macro-mobility architectures in Section 6.3. An analysis under a set of evaluation metrics is provided in Section 6.4. Analytical and simulation results and further discussions are then presented in Section 6.5. Finally, Section 6.6 concludes this chapter.

6.2 System Structure of the Proposed Micro-Mobility Architecture

From now on, we present the design of the proposed micro-mobility architecture, whose structure is outlined in this section.

For macro-mobility support, as proposed in Chapter 4 and Chapter 5, respectively, the home MIP and SIP servers can be either merged into a unified MIP-SIP mobility server called HMS in the TI-MIP-SIP architecture, or kept separated physically whereas combined functionally as a virtual HMS using necessary interactions in the LI-MIP-SIP architecture. Similarly, for micro-mobility support, in a foreign domain to an MH the local SIP servers can be integrated tightly or loosely with an HMIPv6 MAP (or a MIPv4-RR GFA in the IPv4 context). For higher efficiency, a tight integration can be adopted to construct a unified MIP-SIP foreign mobility server (FMS), following the same methodology to construct an HMS. For easier deployment, a loose integration through a collocation of SIP and HMIPv6 servers can be achieved and the resultant collection of

these servers can be deemed as a virtual FMS. Furthermore, an FMS, virtual or not, is preferably collocated with the domain gateway (GW) and collectively called as a GW-FMS. Through this deployment, session traffic and mobility or QoS signalling flows can avoid triangular route via a third party between the GW and an MH. The signalling and operations within a GW-FMS are deemed as internal and assumed to have negligible impacts on signalling costs or delays. Generally, the structure and operations of an FMS (or a virtual FMS) resemble those of an HMS (or a virtual HMS) depicted in Figure 4.2 (or Figure 5.6), and thus is not illustrated here. The main differences, though, are listed as follows. Firstly, the MIP HA is replaced by a domain HMIPv6 MAP. Secondly, a local AAA server replaces the role of AAAH. Thirdly, interfacing with the macro-mobility architecture is introduced, e.g., the signalling and data delivery operations involve additional address translations between local and global addresses, and the data flows are differentiated from the perspective of a foreign domain rather than a home domain. The details of the third aspect are discussed later.

After placing an FMS with a domain GW, we further push the mobility-awareness intelligence to the other side of the domain edges, i.e. the access routers, to make use of micro-mobility enhancements. These arrangements also increase the GW's scalability by distributed computing and registration in the AR level. For higher scalability and reliability considerations, multiple GW-FMS entities may be deployed within a domain, though we demonstrate the single-GW-FMS case in the design. All the other intermediate nodes within the domain are standard routers, unaware of mobility.

Real-time applications are focused on in this micro mobility context. For demonstration clarity, we assume the scenario where an MH is receiving multimedia streaming from a stationary CH during its movements, though the design can be easily extended to bidirectional communications between two mobile hosts. RSVP is ready for

the QoS signalling for such real-time applications. In our architecture, RSVP with mobility extension is running within access domains, and standard RSVP over DiffServ or Aggregated RSVP is operating in the IP core network.

For cost-effective QoS-aware micro handoffs, we propose a two-phased scheme. In the first phase, a series of QoS route extension and fast handoffs are performed between consecutive ARs; in the second phase, which is dynamically triggered, QoS route optimisation is initiated to balance the costs for data delivery and those for QoS and handoff signalling. In our architecture, location management (in the idle mode) follows HMIPv6 in principle, and thus we focus on the handoff management.

Figure 6.1 illustrates the whole picture of the system. In addition to the mentioned building blocks, the figure demonstrates an MH's trajectory during an ongoing streaming session. The MH's movement provokes both a macro handoff between two foreign domains and a number of micro handoffs within each of the foreign domains. Though the micro handoff management is emphasised (Section 6.3), the operations of a macro handoff with QoS signalling in the presence of the proposed micro-mobility architecture are also discussed later (Section 6.4).

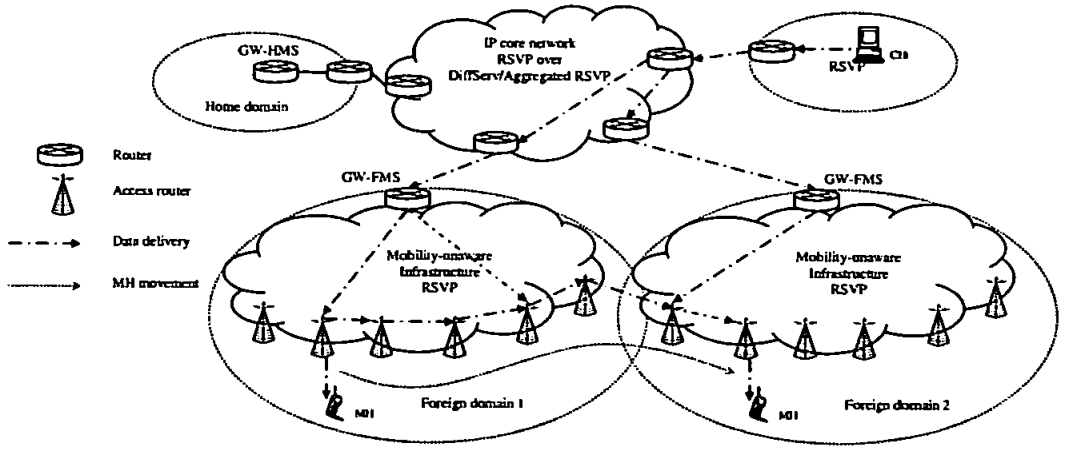


Figure 6.1 Network model and system overview

6.3 The Proposed Micro-Mobility Handoff Management

In this section, we expound the design of handoff management in the micro-mobility architecture. After an overview provided in Section 6.3.1, Section 6.3.2 presents a new scheme to expedite the standard IPv6 address auto-configuration. Subsequently, Sections 6.3.3 and 6.3.4 describe the operations in Phase I and Phase II, respectively. In Section 6.3.5, w derive the expressions to apply a cost-driven algorithm to trigger Phase II dynamically.

6.3.1. Overview

To solve the problems stated in Sections 2.7.4 to 2.7.7 in Chapter 2 and cater for mobile users with high mobility, we propose a two-phased handoff scheme outlined as follows.

In Phase I, the valid unique LCoA obtained when the MH enters a new domain (or after a route optimisation in Phase II), referred to as the primary LCoA, is maintained when the MH moves across ARs within a domain. Notably, it is required that an IPv6 host use a topologically correct source address for outgoing packets [RFC3775]. Thus, for bidirectional IP-level packet transportation convenience, a new transient LCoA is obtained through the FMIPv6 enhanced with an optimised IPv6 address auto-configuration scheme, and this transient LCoA is only registered in the new and the last ARs for packet delivery between them.

In Phase II, a route optimisation is triggered to establish an optimised route between the current AR (and thus the MH) and the GW-FMS. Further, as aforementioned, our consideration on QoS interactions also lead to a two-phased management design. Clearly, these two considerations match each other perfectly and can thus be dovetailed gracefully.

Figure 6.2 depicts the outline of the two-phased micro handoff scheme. The LCoAs shown in the figure are primary LCoAs.

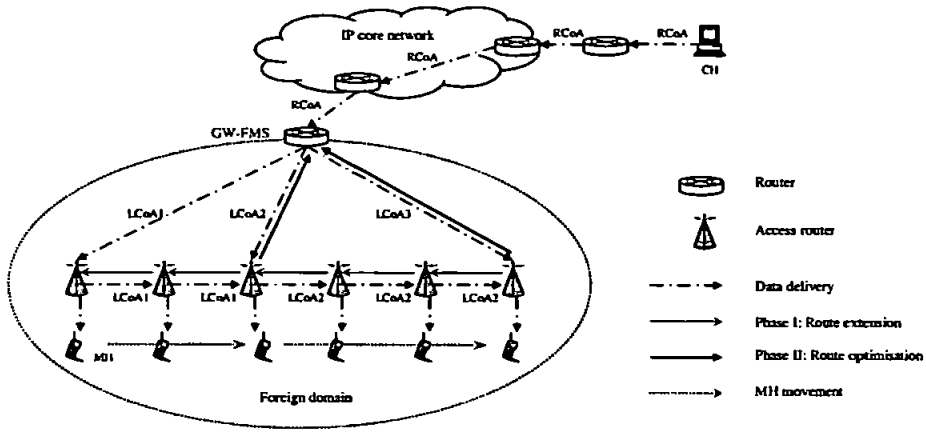


Figure 6.2 The two-phased intra-domain handoff

6.3.2. Acceleration of IPv6 Address Auto-Configuration

HMIPv6 and FMIPv6 (over MIPv6) both rely on the lengthy DAD to verify the uniqueness of a TLCoA or TCoA and possibly another TLCoA or TCoA (or even more in exceptional cases) if the proposed one(s) fail(s) the verification. Particularly, in FMIPv6, the NAR (New AR) is in charge of verifying the TCoA on behalf of an MH through the DAD process when the NAR receives the TCoA carried in the HI message. Therefore, it would be desired to find a new scheme that can facilitate the NAR to complete the address verification more quickly than the DAD does; and at the same time, the new scheme should fulfil the task in an equally safe way as the DAD to avoid the address collision risk imposed by the oDAD proposal [Moore 2005].

Therefore, we devise a new scheme called Prompt Address Verification and Complementary Replacement (PAVER), based on the combination of the in-advance valid address generation method [Vatn and Maguire 1998, Hwang etc 2004] and the distributed host registration at the ARs. As a local registrar, each AR maintains a registration record

(primary LCoA, transient LCoA, L2 address) of all the hosts in its subnet, and this host database is utilised for prompt verification of a proposed LCoA (when FMIPv6 operates over HMIPv6, only LCoAs are needed). In addition, each AR also generates a small pool of very limited complementary LCoAs and verifies them using the standard DAD process as a background operation in advance, so that it can assign a valid LCoA to an MH just in case the proposed LCoA turns out to be invalid (already in the host record). The detailed PAVER operation flow at a PAVER-enabled AR is illustrated by Figure 6.3.

Note that the PAVER scheme may also be applicable to other mobility servers such as GW-FMS or MAP, which need to check the validity of IPv6 addresses to be registered. Thereby, the related registration latency would be significantly reduced. This decreased latency in turn will benefit the handoff performance. Also note that in existing proposals such as [Vatn and Maguire 1998, Hwang etc 2004] each DHCP server or AR keeps generating, verifying (using DAD) and reserving a great amount of valid CoAs/LCoAs for the expected number of MHs in its subnet, and thus considerable costs are invoked even only considering the DAD consummation of the valuable wireless bandwidth. Moreover, in these proposals by default an MH would ask for a valid CoA/LCoA from the NAR other than propose a TCoA itself and have it validated by the NAR as defined in FMIPv6, thus modifications to the standard FMIPv6, both the host and the server modules, must be made. In contrast, the proposed PAVER scheme only modifies the server module at the NAR.

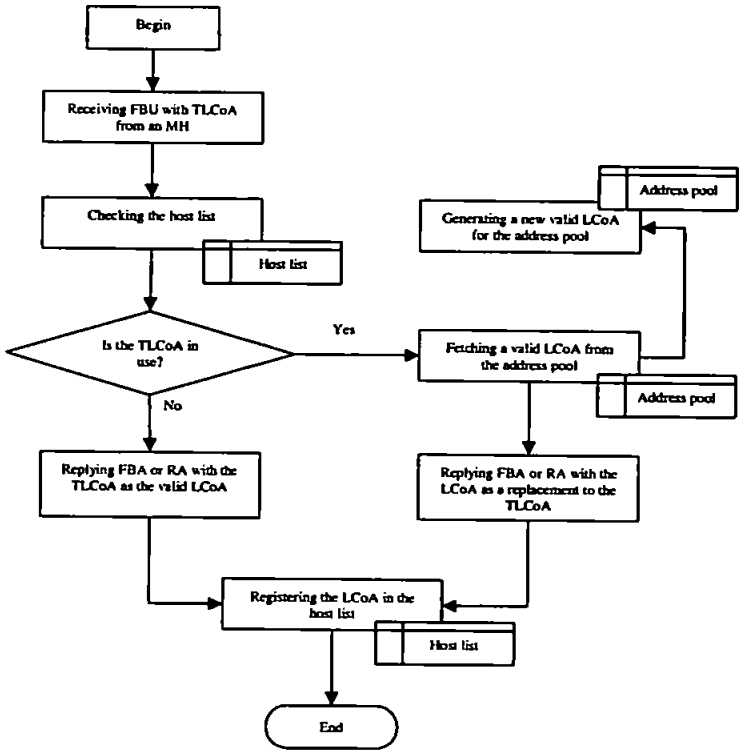


Figure 6.3 PAVER operation at an AR

6.3.3. Phase I Operations

As mentioned before, we apply a two-phased handoff scheme in the proposed micro-mobility architecture. Phase I is based on the FMIPv6 enhanced with the proposed PAVER scheme, and its operations are shown in Figure 6.4. In the figure, AR_0 denotes the AR that serves an MH on its entering the domain or the AR where the last Phase II is initiated; AR_{k-1} and AR_k correspond to the PAR (previous AR) and the NAR in FMIPv6, respectively. The operation sequence is detailed as follows.

Step 1: The MH performs new router detection by exchanging the Router Solicitation for Proxy Advertisement (RtSolPr) and Proxy Router Advertisement (PrRtAdv) messages with the PAR (the current serving AR), and formulates a proposed LCoA by appending its interface identifier to NAR's subnet prefix derived from the PrRtAdv. This handoff

anticipation is enabled by proactive L2 candidate access points probing (scanning). When triggered by an imminent L2 handoff switch, the MH sends a Fast BU (FBU) to PAR. In the FBU, the source address is the current transient LCoA, the proposed LCoA is placed in the Alternate Care-of Address option, and the primary LCoA is in the Home Address option. While starting to buffer the incoming packets meant to the MH, the PAR also initiates its Step-2 operations on receiving the FBU. Meantime, the MH starts the due L2 switch without waiting for an FBA from the network. Note that this timing corresponds to a trade-off of the typical proactive mode and the typical reactive mode in FMIPv6, and is recommended for a couple of reasons. For one thing, the FBU is only triggered by an imminent handoff to ensure the handoff is really happening. For another, for a fast-moving MH there is perhaps no delay allowed to wait for the FBA before the imminent handoff. Therefore, this operation mode (categorised in the reactive mode in FMIPv6) appears more practical than a standard FMIPv6 proactive mode. The involved L2 triggers are enabled by cross-layer signalling mechanisms. Note that L2 switch is the last step of an L2 handoff, where candidate access points probing accounts for the most of the total L2 handoff delay (e.g., more than 90% in IEEE 802.11 WLAN [Mishra etc 2003]). For FMIPv6-enabled schemes, only L2 switch is factored into the total L3 handoff thanks to the handoff anticipation if the MH can keep communicating with the PAR while scanning for candidate NAR(s). In contrast, if the capability of simultaneous scanning or FMIPv6 is unavailable, the total L2 handoff delay has to be added to the L3 handoff delay. In an 802.11b WLAN, the "ad hoc mode" can be configured to facilitate such capability [Bernardos etc 2005]. Optimisation work on the 802.11b driver is also underway in the EU IST Moby Dick project [MobyDick]. The preliminary tests has shown that the total L2 delay can be dramatically reduced even in the default "infrastructure mode", and the total

L2 and L3 handoff delay is as low as between 0 and 15 ms using FMIPv6 with the DAD skipped [Bernardos etc 2005].

Step 2: On receiving the FBU from the MH, the PAR sends an enhanced HI (Handover Initiate) message called HI^+ to the NAR, incorporating the MH's proposed LCoA and L2 (MAC) address, together with mobility-related transferable contexts such as some parameters used in the algorithm to trigger Phase II. On receiving the HI^+, the PAVER scheme is started. The NAR checks its host database for prompt address verification. If the proposed LCoA is not in use, it is valid. Otherwise, the NAR randomly picks an LCoA from its address pool and assigns it to the MH. In either case, the NAR sends a HAck (Handover Acknowledge) message to the PAR and will send a FBA (or a RA with the Neighbour Advertisement Acknowledge option in the latter case) to the MH with the valid new transient LCoA enclosed in both messages. As the MH is probably in the progress of the L2 handoff, the NAR delays the sending of the FBA until it receives a Fast Neighbour Advertisement (FNA) from the MH. Actually, the MH is notified by another L2 trigger immediately after the L2 handoff to send an FNA with the same FBU encapsulated to the NAR. On receiving the FBA (or RA) the MH configures the valid LCoA to its interface and sends a BU to the NAR indicating that it has regained the normal IP connectivity (with a unique topologically correct LCoA). This is later acknowledged by a Path message from the NAR.

On receiving the HAck, the PAR initiates the resource reservation for the route extension, so that the buffered and future packets meant to the MH can be forwarded (by address replacement) or tunnelled (by encapsulation) to the MH with the committed QoS. When the route extension is ready, the PAR starts tunnelling the buffered and following packets to the MH's new transient LCoA directly. When sending packets including resource reservation refreshments, the MH uses the reverse route extension. Note that no

BU is sent beyond the ARs except the infrequent home registration refreshments (e.g., 2 times/hr is a typical value [RFC3344]). In that case, the MH encapsulates the BU. The source and destination addresses of the outer header are the new transient LCoA and the GW-FMS address, and those of the inner header are the RCoA and the HMS address.

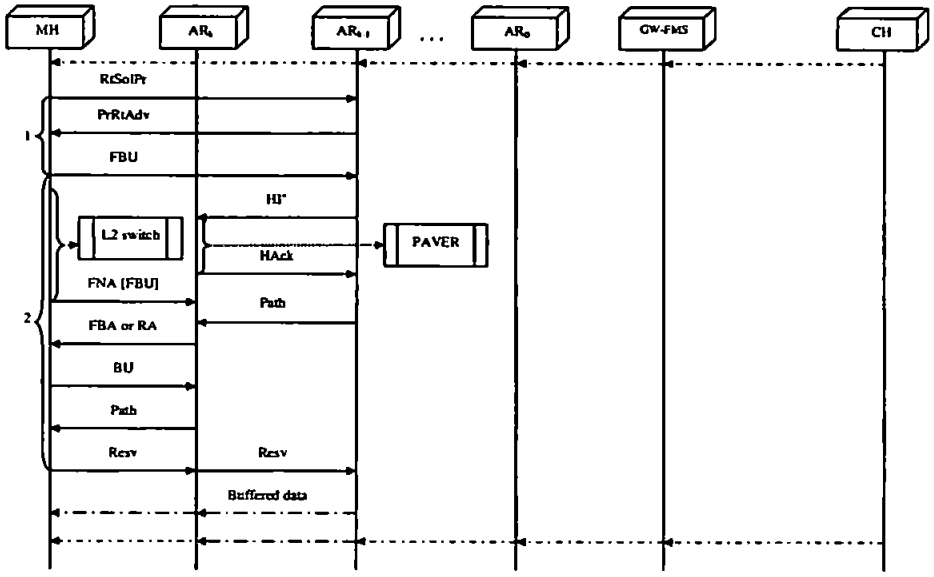


Figure 6.4 Phase I of intra-domain handoff signalling:

fast handoff and QoS route extension

It is worth noting that the MH may revisit one of the ARs (not necessarily the previous one) on a micro handoff and a route loop could be formed. Thus, prior to the route extension in Step 2, potential loop detection and removal should be performed. The current AR can fulfil this task by checking if the MH has registered itself in the binding table. If so, Step 2 is not performed; instead, the current AR initiates the teardown of the looped route and the release of the associated resources (not shown in for clarity).

Though each Phase-I procedure is fairly efficient, after a number of such operations the extended route form a triangular routing and the associated cumulative effects may cancel its benefits and make it no more cost-effective. When this happens, it is time to initiate the Phase-II operations.

6.3.4. Phase II Operations

After (or even during) each Phase-I fast handoff, while incoming packets are delivered to the MH through the extended QoS route the system starts to check if the Phase-II operations should be invoked. In our scheme, the NAR starts monitoring this on behalf of the MH on receiving the HI^+ in Phase I, since an MH is normally power-limited and computation-capability-confined. The following operations are performed, as shown in Figure 6.5. The cost-efficient policy to trigger Phase II is described in the next subsection.

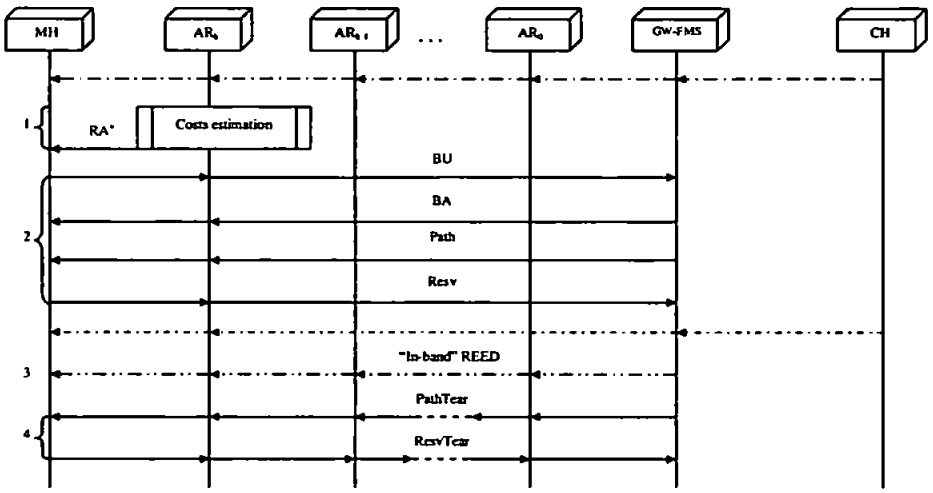


Figure 6.5 Phase II of intra-domain handoff signalling:

regional registration and QoS route optimisation

Step 1: The NAR computes the estimated overall costs for the MH. When the accumulative costs reach a threshold, the NAR sends an enhanced Route Advertisement (RA^+, a RA with a proposed flag set in the Reserved field) to the MH to trigger the route optimisation. The MH then replaces the current primary LCoA with the current transient LCoA.

Step 2: The MH registers its new primary LCoA with the NAR and the GW-FMS using a BU message. The GW-FMS then performs QoS route optimisation between the

GW-FMS and itself and sends a Path message towards the MH. An optimised route is actually created by the Resv message from the MH to the GW-FMS along the reverse way.

Step 3: Now the incoming packets arriving at the GW-FMS are ready to be forwarded to the MH after address conversion along the optimised route. However, packets forwarded from the extended route and the optimised route can be interleaved at the MH. Such out-of-sequence packets lead to wastage of the buffers implemented in real-time applications for stream compensation, and thus this problem should be addressed. Thus, a simple process, referred to as REED (Route Extension End Declaration), similar to those proposed for ATM rerouting in [Kim and Kim 2003] is introduced to deal with the packet out-of-sequence problem. After sending out the Resv message, the GW-FMS stops forwarding incoming packets to the MH's old LCoA and starts sending packets to the MH's new LCoA. Meanwhile, it sends an "in-band" REED message to the MH's old LCoA. Assuming FIFO (First-In-First-Out) data buffers are applied, this message travels as the data packet did along the extended route and reaches the MH finally. At the MH side, it buffers the packets from the optimised route at the IP layer while keeping delivering packets from the extended route. The MH can differentiate these two streams by the source address of an IP packet. On receiving the REED, the MH becomes aware that no more packets will be forwarded through the extended route and it starts to deliver the buffered and following packets from the optimised route to the upper layer. This scheme is also applicable to the macro handoff case. Note that this out-of-sequence problem does not happen in our Phase I scheme because the route is simply extended for a single data stream. However, for other schemes like FMIPv6, this problem occurs in each handoff. An alternative scheme is that the GW-FMS marks the last packet sent along the extended route, e.g., an unused bit in the IP header can be set, to indicate that this packet is the last one.

Step 4: Finally, on receiving the first packet routed directly through the optimised QoS route, the MH initiates the removal of the extended route and release of the involved resources along that route. This happens inside the domain and does not affect the remaining established route outside.

6.3.5. The Cost-Efficient Policy to Trigger Phase II

As described, Phase I is cost-effective in terms of fast and smooth handoff with low signalling costs, nevertheless, the data delivery cost becomes large after a number of consecutive Phase-I procedures; in contrast, Phase II is efficient in data delivery along an optimised route at the price of high signalling loads. There is a trade-off between the data delivery costs and the signalling costs. The cost-efficient policy targets to seek the lowest expected total costs for a sequence of micro handoffs with QoS constrains during a session's lifetime.

The decision whether to trigger the Phase II should be made after the Phase-I operations at AR_i (the new AR) and before the next handoff towards AR_{i+1}(the next AR). Let $A = \{NRO, RO\}$ denote the basic action set on each micro handoff, where *NRO* corresponds to the action that only Phase I (route extension) is performed and *RO* is the action that Phase II (route optimisation) is triggered after Phase I. Let $C(i, NRO)$ denote the estimated signalling and data delivery costs along the current route appended with the extended route from AR_{i-1} (the previous AR, PAR) to AR_i (the new AR, NAR), and $C(i, RO)$ denote the estimated signalling and data delivery costs along an optimised route between the GW-FMS and AR_i if Phase II is triggered. Basically, Phase II should be triggered when $C(i, NRO)$ becomes larger than $C(i, RO)$. The detailed operation flows are depicted by Figure 6.6.

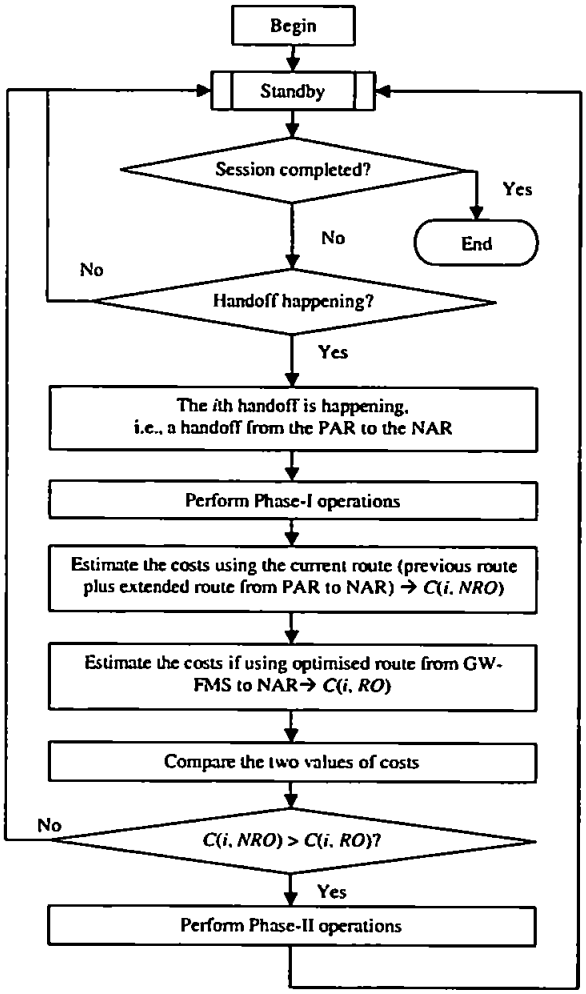


Figure 6.6 Flow chart of the two-phased operations

In the following, we derive the expressions of $C(i, NRO)$ and $C(i, RO)$ to specify the trigger algorithm, and analyse the total accumulative costs during a session's lifetime. Let x_i and x_i' denote the number of hops in the actual route (extended or optimised) and the optimised route (if *RO* is performed at AR_i) between GW-MAP and AR_i respectively, and z_i denote the change in the number of hops after Phase I at AR_i. All excludes the last hop between the MH and the AR it is attached to. Thus,

$$x_i = \begin{cases} x_{i-1} + z_i & \text{If } NRO \text{ is performed;} \\ x_i' & \text{If } RO \text{ is performed.} \end{cases} \qquad (6.1)$$

Let y_i denote the number of hops in the route between AR_{i-1} and AR_i, y_i' denote the reduced number of hops due to a loop removal at AR_i, thus,

$$z_i = \begin{cases} y_i & \text{If } AR_i \text{ is new to the MH;} \\ -y_i' & \text{If } AR_i \text{ is a revisited AR.} \end{cases} \quad (6.2)$$

Note that all the above distances (in hops) are regional parameters within the access domains and can thus be obtained or estimated easily. We assume that such information is available in the involved routers.

Costs trade-off is derived as follows. The signalling costs incurred by the ith handoff is given by

$$C_i^{signalling} = C_i^{I-signalling} + p_{II} \cdot C_i^{II-signalling}, \quad (6.3)$$

where $C_i^{I-signalling}$ and $C_i^{II-signalling}$ are the signalling costs generated from Phase I and II respectively, and

$$p_{II} = \begin{cases} 1 & \text{If } RO \text{ is performed,} \\ 0 & \text{If } NRO \text{ is performed.} \end{cases} \quad (6.4)$$

Referring to Figure 6.4 and Figure 6.5, we can obtain $C_i^{I-signalling}$ and $C_i^{II-signalling}$, respectively. As defined in Chapter 4, the signalling cost incurred by a single message is calculated by the product of the message's IPv6 packet size and the number of hops it traverses. For presentation purpose, we use the name of a message to stand for its size.

For Phase I, the costs are calculated as follows.

If $z_i = y_i$,

$$\begin{aligned} C_{I-signalling}(i) = {} & (RtSolPr + PrRtAdv) + (FBU + FNA + FBA + BU) \\ & + (HI^* + HAck) \cdot y_i + (Path + Resv) \cdot (1 + y_i) \end{aligned} \quad (6.5)$$

If $z_i = -y_i'$,

$$\begin{aligned} C_{I-signalling}(i) = {} & (RtSolPr + PrRtAdv) + (BU + BA) \\ & + (BU + BA + PathTear + ResvTear) \cdot y_i') \end{aligned} \quad (6.6)$$

Similarly, for Phase II, the costs are given by

$$C_{II-signalling}(i) = RA^{+} + (BU + BA + Path + Resv)\cdot(1 + x_i^{'}) + (REED + PathTear + ResvTear)\cdot(1 + x_i). \quad (6.7)$$

Compared with the signalling costs, which are transient on each handoff, data delivery costs are continuously invoked between two handoffs. Let T_s denote the session holding time, T_i denote the subnet residence time at AR_i, T_0 denote the residual time at AR_0 whereby the session is started, K denote the total number of micro handoffs during T_s, and $C_{delivery}$ denote the average data delivery costs (homogenous to signalling costs assumed) per hop. Then we can obtain the data delivery costs at AR_i as

$$C_i^{data} = \begin{cases} C_{delivery}\cdot(x_i + 1)\cdot T_i, & \text{If } 0 < i < K; \\ C_{delivery}\cdot(x_i + 1)\cdot(T_s - \sum_{j=1}^{K-1} T_j - T_0), & \text{If } i = K. \end{cases} \quad (6.8)$$

Therefore, the total costs on the ith handoff are

$$C_i = C_i^{signalling} + C_i^{data}. \quad (6.9)$$

Furthermore, the accumulative costs after K handoffs are given by

$$C_{Accum}^{\pi(K)}(K) = \sum_{i=1}^{K} C_i = \sum_{i=1}^{K} (C_i^{signalling} + C_i^{data}), \quad (6.10)$$

where $\pi(K)$ be the handoff action sequence. Let $P(K)$ be the probability that the MH performs K handoffs during a session, and the expected accumulative costs for a sequence of K handoffs are computed by

$$C^{\pi(K)}(K) = \sum_{i=1}^{K} C_i \cdot P(K) = \sum_{i=1}^{K} (C_i^{signalling} + C_i^{data})\cdot P(K). \quad (6.11)$$

In these K handoffs, on each handoff an action is taken from the action set $A = \{NRO, RO\}$. Our aim is to identify the most cost-efficient sequence of actions, denoted by $\pi_{opt}(K)$, to minimise the (expected) accumulative costs

$$C^{\pi_{opt}(K)}(K) = \min[\sum_{i=1}^{K} (C_i^{signalling} + C_i^{data})\cdot P(K)]. \quad (6.12)$$

Assume that the MH's residence time in an AR area (subnet) follows a general distribution $f_r(t)$ with mean $1/\eta$, and the session holding time follows an exponentially distribution $f_s(t)$ with mean $1/\mu$, then $P(K)$ is given by [Lin etc 1994]

$$P(K)=\begin{cases}\dfrac{\eta}{\mu}[1-f_r^*(\mu)]^2[f_r^*(\mu)]^{K-1} & \text{If } K\geq 1\\ 1-\dfrac{\eta}{\mu}[1-f_r^*(\mu)] & \text{If } K=0,\end{cases} \tag{6.13}$$

where $f_r^*(s)$ is the Laplace transform for $f_r(t)$. Equation (6.13) can be solved when $f_r(t)$ reduces to a specific distribution. To facilitate the implementation of this cost-driven policy, a value L_{opt} can be obtained to serve as the optimised threshold for triggering Phase II. To solve (6.12), a specific algorithm is applied as shown in Figure 6.7.

On the ith micro handoff, compute the following costs:

$C(i,NRO)=C_{I-sig}(i)+C_{delivery}\cdot(x_{i-1}+y_i)\cdot 1/\eta$;

$C(i,RO)=C_{I-sig}(i)+C_{II-sig}(i)+C_{delivery}\cdot x'_i\cdot 1/\eta$.

If $C(i,NRO)>C(i,RO)$, Phase II is triggered after Phase I. The number of handoffs so far since the last-time Phase II or the beginning of the session is used as the L_{opt}. The next-time costs computation may only be performed after another L_{opt} handoffs to save the computation efforts at ARs.

Else, no further actions (except Phase I) are needed.

Figure 6.7 Algorithm to derive a cost-efficient trigger threshold

Notably, all the involved parameters are intra-domain variables and thus we assume that their actual or estimated values are easily obtainable to the NAR. For instance, the distance parameters y_i and x_i' may be derived from routing information, and x_{i-1} can be made available to the NAR by context transfer from the PAR via the HI$^+$ message.

Interestingly, different schemes can be obtained according to different values of L_{opt}, the derived optimised threshold. When $L_{opt} = 1$ the two-phased scheme retreats to a scheme where RO is always performed; whereas when $L_{opt} = 0$ it becomes another scheme where RO is never performed. The latter is equivalent to the deployment where ARs are equipped with MAPs. Hereafter, these schemes are referred to as HMIP-FH-optimisedRO, HMIP-FH-alwaysRO and HMIP-FH-neverRO respectively, and HMIP-FHs collectively.

6.4 Interaction with the Macro-Mobility Proposals

As aforementioned, either of the macro-mobility architectures proposed in Chapter 4 and Chapter 5 respectively can operate as a standalone solution. When the proposed micro-mobility architecture is applied together with one of the macro-mobility architectures, the interaction operations must be specified to ensure a seamless cooperation.

6.4.1. Address Translations in the Involved Protocols

Generally speaking, with the micro-mobility architecture an MH is identified by its HMIPv6 Regional CoA (RCoA) and (primary) on-link CoA (LCoA) to the other network entities outside or inside of the foreign domain, respectively. The following address conversion, by encapsulation or replacement, is conducted to the outgoing/uplink (from the domain to the outside) and incoming/downlink (from the reverse direction) session packets and the messages in the involved protocols at the serving GW-FMS.

HMIPv6 over MIPv6 messages: As defined in HMIPv6, for the outgoing messages, the LCoA is swapped to the RCoA by encapsulating an outer header whose source address and destination address are set to be the RCoA and the CH's address, respectively; for incoming messages, the RCoA is converted to the LCoA similarly. Session packets experience the same translations.

RSVP messages: In addition to the above translations of RCoA and LCoA in the IP headers, RSVP messages contain the communicating IP addresses in their bodies, and must be swapped as well. These operations are defined in [Paskalis etc 2003]. In short, an LCoA-to-RCoA address translation is performed to the Sender_Template object of a Path message or the Session object of a Resv message, respectively, when the message is outgoing; and a reverse translation happens to the Session object of a Path or the Sender_Template object of a Resv, respectively, when the message is incoming. In addition, the corresponding Path State Block or Resv State Block needs to be updated as well.

SIP messages: Similar to RSVP messages, address translations happen to the SIP body in addition to the IP headers. The LCoA to the RCoA translation should be performed to the Contact header of an outgoing SIP message, and the reverse translation to an incoming message. Furthermore, similar operation should be conducted to the IP address in the SDP (Session Description Protocol) 'c' (connection information) session description if included in a SIP message.

In the next subsection, the above address translations are further contextualised in the interactions between macro- and micro-mobility signalling with QoS interworking.

6.4.2. QoS-Enhanced Macro-Mobility in the Presence of Micro-Mobility

As specified in MIPv6, the CH sends packets to the MH's current RCoA directly by setting the RCoA in the destination address field and the MH's home address (HoA) can be contained in the Type-2 routing header. When an MH enters a new foreign domain with an ongoing session with its CH, a macro handoff occurs and a sequence of operations is performed as shown in Figure 6.8. We demonstrate the operations in the TI-MIP-SIP

architecture though it is also applicable to the LI-MIP-SIP architecture with minor modifications (e.g., MIP-based interactions between MIP HA and SIP HR in Step 2).

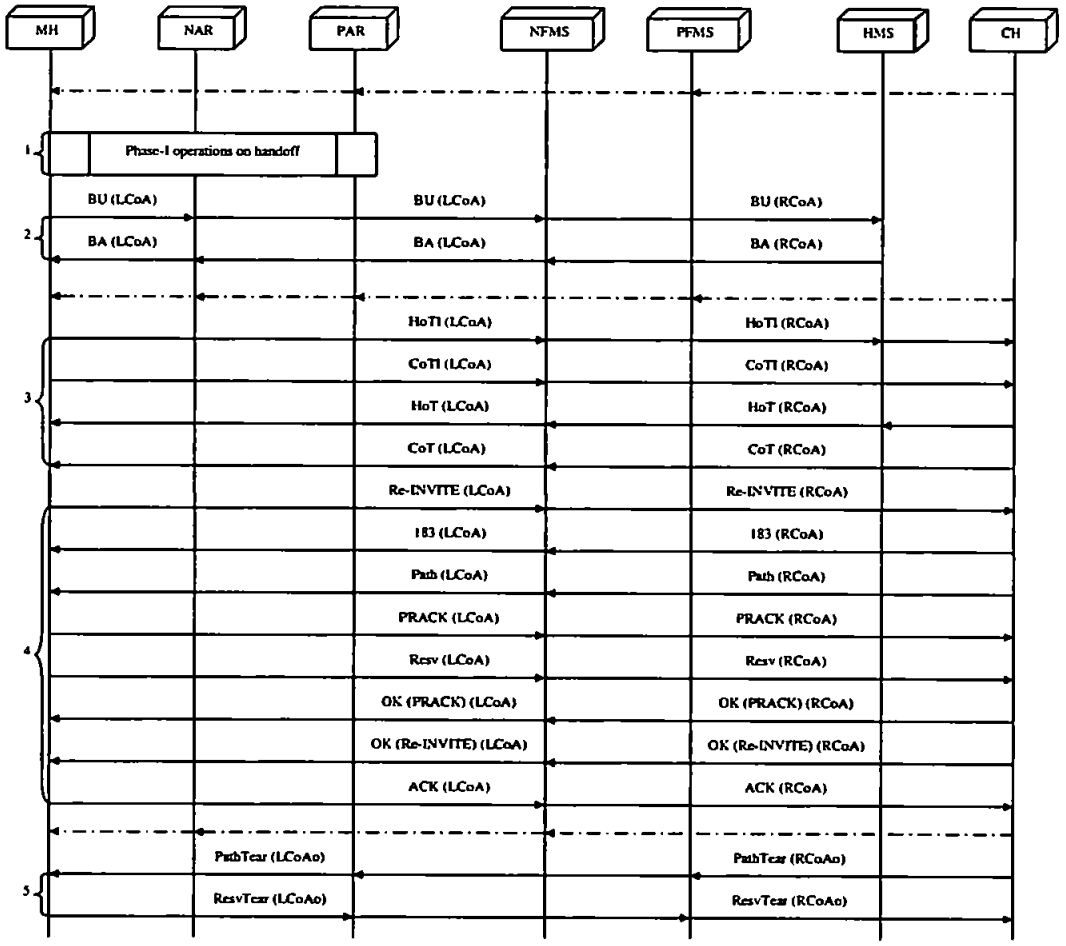


Figure 6.8 Macro-mobility handoff in the presence of the micro-mobility architecture

Step 1: The Phase-I operations are carried out if both the previous and the new domains support the proposed micro-mobility architecture, as assumed here, though an enhanced AAA operations facilitated by the context transfer may be involved.

Step 2: The MH sends a BU towards the new FMS (NFMS). In the BU, the LCoA is the source address; the proposed RCoA, the MIP HoA, and the NAI (or the SIP AOR) are set in the Alternate Care-of-Address, Home Address and NAI options, respectively. The NAR intercepts the BU, creates a binding record in its local host database, and then

forwards the BU to NFMS. The NFMS verifies the proposed RCoA using PAVER and binds the LCoA and a valid RCoA of the MH, and then sends a BU to the MH's HMS, notifying it of the MH's new RCoA for home registration. The binding record maintained in the FMS for each MH is (SIP AOR/MIP NAI, MIP HoA, HMIP RCoA, HMIP LCoA), whereas the HMS is unaware of LCoA. On a successful home registration, the HMS replies with a BA. The NFMS sends a BA (with the valid RCoA set in the Type-2 routing header) to NAR, which in turn forwards it to the MH. Note that this step is based on the local and home registration procedures described in HMIPv6, whereas the message exchanges between the MH and the GW-FMS and thus the round trips are reduced compared with HMIPv6 thanks to a similar usage of PAVER in GW-FMS.

Step 3: The MH imitates the MIPv6 Return Routability tests to authenticate itself to the CH. This step is optional and only performed when such authentication cannot be achieved by the following SIP INVITE message because certain reasons, e.g., no security association was established in the SIP session setup stage or the INVITE cannot make use of any other AAA information that can be embedded in it.

Step 4: The end-to-end session renegotiation and QoS route optimisation is then initiated between an MH and its CH. In this process, SIP and RSVP messages are dovetailed as specified in [RFC3312]. To adjust [RFC3312] for this micro-mobility scenario, operations are needed to convert addresses for SIP messages including the enclosed SDP session descriptions. In the uplink direction, the NFMS swap the LCoA with the RCoA as the packet's source address, and modifies the Contact header and the SDP 'c' session description if included by replacing the LCoA with the RCoA. The inverse operations are conducted in the downlink direction. The MH sends a SIP Re-INVITE message to the CH, notifying its new CoA for the binding update in the CH and the session's preconditions (constrains). Having been assured of this new RCoA through a

successful Step 4, and the CH returns a SIP 183 message to indicate the progress of the ongoing session. As the sender of the session, meantime the CH sends a PATH towards the new RCoA. When the PATH reaches the MH via NFMS, the MH replies a Resv for resource reservation. Similar as the SIP messages, address conversion in both packets' headers and internal state blocks are performed at the NFMS for Path and Resv messages. Notably, the CH continues to send packets to the MH's old RCoA until it receives the Resv from the MH. Then it starts to send packets to the MH's new RCoA. The MH receives packets through the extended route from the old domain first and then packets from the optimised route. For bi-directional communications between the MH and the CH, the SIP UPDATE message is needed to notify the CH that the QoS route of the other direction is ready (not shown here).

Step 5: Though the old QoS route can expire without further refresh, it is recommended that the route is torn down and the previously reserved resources associated with the route are released through the explicit RSVP PathTear and ResvTear messages as soon as appropriate. This can be initiated by a PathTear towards the MH's old RCoA (RCoAo) when the CH starts to send packets to the new RCoA. The ARs along the old route also deregister the MH's old CoAs.

6.5 Performance Evaluation

In this section, we evaluate the performance of the proposed protocol, compared with their counterparts, respectively. We first carry out a theoretical analysis in Section 6.5.1 and then provide the analytical and simulation results in Sections 6.5.2 and 6.5.3, respectively.

6.5.1. Performance Analyses

The performances of the proposed micro-mobility architecture is evaluated in terms of the following metrics, and compared with the standard HMIPv6 and FMIPv6 wherever appropriate.

6.5.1.1. Binding Update Delay

The binding update delay refers to the elapsed time between the epoch when an MH starts L2 handoff and the epoch when a valid binding update is performed in the appropriate network entity (mobility server or ARs) or the CH, which then can confirm that the MH has regained its IP connectivity with the new valid CoA or LCoA. In those schemes where no fast handoff (FH) is implemented, this delay is proportional to the handoff packet loss generated since the mobility server or the CH now stops sending packets to the previous CoA or LCoA and will resume the sending when the QoS route is repaired. On the other hand, in the FH-enabled schemes the new AR can now send IP packets (the Path message in the concerned schemes) to the MH as the MH has configured a unique topologically correct IPv6 address.

For presentation purpose, let define $T(message\_i, A-B) \equiv T_i^h$, where h represents the distance (in hops) between entity A and entity B and T_i^h has been given by (5.13). Thus, $T(message\_i, A-B)$ represents the end-to-end delay for *message_i* between A and B.

In HMIPv6, the binding update delay is given by

$$T_{HMIP}^{BU} = T_{L2} + T_{RA} + T_{LCoA} + T(BU, MH-MAP), \tag{6.14}$$

where T_{L2} is the typical L2 handoff delay including the L2 switch delay $T_{L2-switch}$ and the delays before the L2 switch for scanning the new L2 attachment etc, T_{RA} is the average minimum delay to detect the NAR, and T_{LCoA} is the minimum delay to for an MH to obtain a valid LCoA using the DAD process.

In FH-enabled schemes, let

$$\tau_1 = \tau_0 + T_{L2\text{-}switch} + T(FNA, MH - NAR), \tag{6.15}$$

where τ_0 is the delay between an MH sends the FBU to PAR and the L2 switch is initiated. Normally, the smaller τ_0 is, the higher the probability that the L2 switch is actually happening. Note that $T_{L2\text{-}switch}$ should be replaced with T_{L2} in (6.15), if the MH cannot scan for NAR(s) before losing the connectivity with the PAR.

In FMIPv6, let

$$\tau_{2\text{-}1} = T(FBU, MH - PAR) + T(HI, PAR - NAR) + T_{LCoA}, \text{ and} \tag{6.16}$$

$$\tau_{3\text{-}1} = \max(\tau_1, \tau_{2\text{-}1}) - \tau_0 + T(FBA, NAR - MH) + T(BU, MH - NAR). \tag{6.17}$$

In the proposed HMIP-FHs, let

$$\tau_{2\text{-}2} = T(FBU, MH - PAR) + T(HI, PAR - NAR) + T'_{LCoA}, \text{ and} \tag{6.18}$$

$$\tau_{3\text{-}2} = \max(\tau_1, \tau_{2\text{-}2}) - \tau_0 + T(FBA, NAR - MH) + T(BU, MH - NAR), \tag{6.19}$$

where T'_{LCoA} is the delay to generate a valid LCoA using the PAVER scheme at the NAR.

Then, their binding update delay is given by

$$T^{BU}_{FMIP} = \tau_{3\text{-}1}, \text{ and} \tag{6.20}$$

$$T^{BU}_{HMIP\text{-}FHs} = \tau_{3\text{-}2}, \text{ respectively.} \tag{6.21}$$

6.5.1.2. Handoff Delay

The handoff delay is defined here as the elapsed time between the epoch when an MH starts an L2 handoff (or L2 switch in FH-enabled schemes) and the epoch when the QoS route towards the MH's new location is repaired.

In HMIPv6, the handoff delay is computed by

$$T^{HO}_{HMIP} = T^{BU}_{HMIP} + \max[T(BA, MAP - MH), T(Path, MAP - MH)] + T(Resv, MH - MAP). \tag{6.22}$$

In the FH-enabled schemes, let

$$\tau_{4-m} = \tau_{2-m} - \tau_0 + T(HAck, NAR - PAR) + T(Path, PAR - NAR), \text{ and} \quad (6.23)$$

$$\tau_{5-m} = \max(\tau_{3-m}, \tau_{4-m}), \quad (6.24)$$

where m = 1 for FMIPv6; m = 2 for HMIP-FHs.

Thus, their handoff delay is respectively expressed as

$$T^{HO}_{FMIP} = \tau_{5-1} + T(Path, NAR - MH) + T(Resv, MH - NAR) + T(Resv, NAR - PAR), \text{ and} \quad (6.25)$$

$$T^{HO}_{HMIP-FHs} = \tau_{5-2} + T(Path, NAR - MH) + D(Resv, MH - NAR) + T(Resv, NAR - PAR). \quad (6.26)$$

6.5.1.3. Handoff Packet Loss

The handoff packet loss denotes the number of lost packets due to a micro handoff. Let λ_d be the packet arrival rate of the ongoing session and assume that the network starts to buffer packets for the MH on the notification (through an FBU or a BU) from the MH.

In HMIPv6, since packets arriving at the PAR are not buffered at the ARs during the handoff, packets are simply dropped. Assume that packets are started to be buffered at the CH or the MAP when a BU reaches it, the handoff packet loss are respectively given by

$$L^{HO}_{HMIP} = \lambda_d \cdot T^{BU}_{HMIP} \quad (6.27)$$

In the FH-enabled schemes, on the contrary, thanks to the fast smooth handoff the on-the-fly packets are buffered throughout the handoff delay time and thus packets loss could be eliminated given the buffer size in an AR is large enough and the FBU reaches the PAR early enough. Let RTT_1 denote the round-trip time between an MH sends the FBU before the L2 switch and the PAR sends a data packet, and it is given by

$$RTT_1 = T(FBU, MH - PAR) + T(data, PAR - MH), \quad (6.28)$$

and their handoff packet loss is given by

$$L^{HO}_{HMIP-FHs} = L^{HO}_{FMIP} = \begin{cases} \lambda_d \cdot (RTT_1 - \tau_0) & \text{If } RTT_1 > \tau_0 \\ 0 & \text{If } RTT_1 \le \tau_0. \end{cases} \quad (6.29)$$

Thus, to eliminate the handoff packet loss, an MH should start the L2 switch at least RTT_1 time after it sends an FBU to PAR, though this requirement may not always be met.

6.5.1.4. Out-of-Sequence Packets

The out-of-sequence packets (OSP) are generated at an MH when more than one packet stream of a same session is sent towards it simultaneously. In HMIPv6, HMIP-FH-neverRO and the first phase of the HMIP-FH-optimsedRO, packets are delivered in a single sequence, and thus no out-of-order packets are produced. Therefore,

$$OSP_{HMIP} = OSP_{HMIP-FH-neverRO} = 0. \quad (6.30)$$

On the other hand, in FMIPv6, HMIP-FH-alwaysRO and the second phase of HMIP-FH-optimisedRO when the QoS route from the CH or the central mobility server (GW-MAP or GW-FMS) is ready, the CH or the central mobility server stops sending packets to the previous LCoA (intercepted by the PAR), and starts sending packets to the new LCoA (via the NAR). Therefore, two packet streams are sent to the MH in parallel: one is from the buffer of the PAR, and the other is from the CH directly or via the GW-FMS. The travel delay difference between the two streams corresponds to the packets delivered in order from the PAR, whereas the remaining packets of the buffered packets in PAR are interleaved with those from the other source and thus the actual OSP is doubled assuming both streams arrive at the rate of λ_d.

In FMIPv6, the OSP are estimated by

$$OSP_{FMIP} = \{\lambda_d \cdot T^{HO}_{FMIP} - \lambda_d \cdot [T(data, CH - NAR) - T(data, PAR - NAR)]\} \cdot 2. \quad (6.31)$$

In HMIP-FHs without the REED scheme, the OSP are estimated similarly by

$$OSP_{HMIP-FHs\_w/o\_REED} = \{\lambda_d \cdot T^{HO}_{HMIP-FHs} - \lambda_d \cdot [T(data, CH - NAR)$$

$$-T(data, PAR-NAR)]\}\cdot 2. \tag{6.32}$$

To deal with this problem, we introduced REED to guarantee the data packet sequence. Thus,

$$OSP_{HMIP-FH-optimizedRO} = OSP_{HMIP-FH-alwaysRO} = 0. \tag{6.33}$$

This is achieved at the price of higher buffer size requirements as computed later.

6.5.1.5. Handoff Buffer Size Requirements

Buffers are needed to reduce handoff packet loss or holding outgoing packets until the QoS route is repaired after a handoff. Different handoff schemes impose different buffer size requirements on the involved entities. In the FH-enabled schemes, the handoff delay is proportional to the buffer size required in the ARs; whereas in the non-FH-enabled schemes, the difference of the handoff delay and the binding update delay corresponds to the buffer size required in the mobility server or the CH.

Therefore, in HMIPv6 the required buffer size for K handoffs is calculated by

$$B_{HMIP}^{HO} = \lambda_d \cdot \sum_{i=1}^{K} (T_{HMIP}^{HO} - T_{HMIP}^{BU})_i \cdot P(K). \tag{6.34}$$

The buffer size requirements in FMIPv6 and HMIP-FH-alwaysRO are represented, respectively, by

$$B_{FMIP}^{HO} = \lambda_d \cdot \sum_{i=1}^{K} (T_{FMIP}^{HO})_i \cdot P(K), \text{ and} \tag{6.35}$$

$$B_{HMIP-FH-neverRO}^{HO} = \lambda_d \cdot \sum_{i=1}^{K} (T_{HMIP-FHs}^{HO})_i \cdot P(K) \tag{6.36}$$

In HMIP-FHs except HMIP-FH-neverRO, Phase I has a similar buffer requirements in PAR as FMIPv6; yet Phase II has an additional buffer requirement for storing incoming packets at the MH during the REED process. Similar with the OSP analysis, the total expected buffer size requirements are expressed by

$$B_{HMIP-FH-optimisedRO}^{HO} = B_{HMIP-FH-alwaysRO}^{HO} = \lambda_d \cdot \sum_{i=1}^{K} (T_{HMIP-FHs}^{HO})_i \cdot P(K) + \lambda_d \cdot \sum_{i=1}^{K'} \{(T_{HMIP-FHs}^{HO}) - [T(data, FMS-NAR) - T(data, PAR-NAR)]\}_j \tag{6.37}$$

$\cdot P(K)$

where K' is the actual times when the Phase II is triggered, and j corresponds to the subnet where a Phase II is performed. When the MH's itinerary does not form any loops, and given the optimised trigger threshold L_{opt} for Phase II, we have

$$K' = \lfloor K/L_{opt} \rfloor \text{ and } j = L_{opt} \cdot i, \tag{6.38}$$

where $\lfloor \cdot \rfloor$ is the function to round the element to the nearest integer smaller than the element.

6.5.1.6. Expected Total Handoff Costs

The expected total handoff (ETHO) costs metric takes into account the expected signalling and data delivery costs incurred for the K handoffs during a session's lifetime. The expected total handoff costs in HMIP-FHs are collectively expressed by (6.12), though their individual costs are different due to their different actions.

In HMIPv6, the ETHO costs are computed by

$$C_{HMIP}^{ETHO} = \sum_{i=1}^{K} [(C_{HMIP}^{HO})_i + C_{data} \cdot (x_i'+1) \cdot (T_i - (T_{HMIP}^{HO} - T_{HMIP}^{BU})_i)] \cdot P(K), \tag{6.39}$$

where

$$\begin{aligned} C_{HMIP}^{HO} &= C_{LCoA} + C(BU, MH - MAP) + C(BA, MAP - MH) + \\ &C(Path, MAP - MH) + C(Resv, MH - MAP). \end{aligned} \tag{6.40}$$

In FMIPv6, the costs are

$$C_{FMIP}^{ETHO} = \sum_{i=1}^{K} [(C_{FMIP}^{HO})_i + C_{data} \cdot (x_i'+1+X) \cdot T_i] \cdot P(K), \tag{6.41}$$

where

$$C_{FMIP}^{HO} = C_{FMIP-I}^{HO} + C_{FMIP-II}^{HO}, \tag{6.42}$$

$$\begin{aligned} C_{FMIP-I}^{HO} &= C(RtSolPr, MH - AR_{i-1}) + C(PrRtAdv, AR_{i-1} - MH) \\ &+ C(FBU, MH - AR_{i-1}) + C(FNA, MH - AR_i) \\ &+ C(HI, AR_{i-1} - AR_i) + C(HAck, AR_i - AR_{i-1}) \\ &+ C(FBA, AR_i - MH) + C(BU, MH - AR_i) \end{aligned} \tag{6.43}$$

$$+C(Path, AR_i - AR_{i-1}) + C(Path, AR_{i-1} - MH)$$
$$+C(Resv, MH - AR_i) + C(Resv, AR_i - AR_{i-1}), \text{ and}$$

$$C^{HO}_{FMIP-II} = C(BU, AR_i - CH) + C(BU, CH - MH)$$
$$+C(Path, CH - MH) + C(Resv, MH - CH). \quad (6.44)$$

6.5.1.7. Expected Signalling Costs for Location Update at the Central Mobility Server

The signalling costs for location update include the costs for binding update and acknowledgement upon each handoff. This metric serves as an indicator of the costs incurred at the central mobility server.

In FMIPv6 (running over MIPv6 by default), upon each handoff an MH updates its location at the HA, and the signalling costs for K handoffs are given by

$$C^{LU}_{FMIP} = \sum_{i=1}^{K} [C(BU, MH - HA) + C(BA, HA - MH)]_i \cdot P(K)$$
$$= \sum_{i=1}^{K} (BU + BA) \cdot (x_i' + 1 + X_i) \cdot P(K), \quad (6.45)$$

where X_i denotes the number of hops between the HA and the GW when the ith handoff happens.

In HMIPv6 and HMIP-FH-alwaysRO, an MH updates its location at the GW-MAP or GW-FMS, and the costs are given by

$$C^{LU}_{HMIP} = C^{LU}_{HMIP-FH-alwaysRO} = \sum_{i=1}^{K} (BU + BA) \cdot (x_i' + 1) \cdot P(K). \quad (6.46)$$

In HMIP-FH-optimisedRO, an MH only updates its location at the GW-FMS when Phase II is triggered after every L_{opt} handoffs, and the costs are

$$C^{LU}_{HMIP-FH-optimisedRO} = \sum_{i=1}^{K'} (BU + BA) \cdot (x_j' + 1) \cdot P(K), \quad (6.47)$$

where K' and j are the same parameters defined in Section 6.5.1.6 and are given by (6.41) when no loops are formed during an MH's trajectory. At other times, this overhead is distributed among the ARs.

In HMIP-FH-neverRO, an MH only registers with ARs and never conducts location

updates in a central mobility sever. Thus, these costs are

$$C^{LU}_{HMIP-FH-neverRO} = 0. \tag{6.48}$$

6.5.2. Analytical Results

In this subsection, we present the numerical results based on the above analyses. The parameter configuration is given and the results are illustrated and analysed.

6.5.2.1. Input Parameters

To obtain numerical results, we assign typical values to the involved input parameters, as listed in Figure 6.9 and Table 6.1

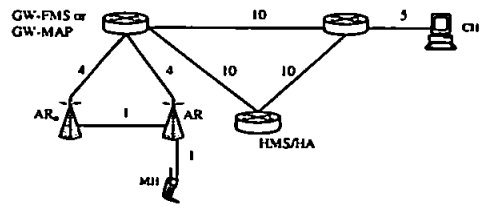


Figure 6.9 Default distance (in hops) between entities

Table 6.1 Parameter setting for evaluating micro-mobility protocols

(b) Message size

Protocol	Message Size (bytes)	Protocol	Message Size (bytes)
MIPv6	100	SIP	400
HMIPv6	100	RSVP	200
FMIPv6	100	Other	100

(b) Other parameters

Parameter	Value	Parameter	Value
Mean service rate of an RSVP message at a router	10,000 messages/s	Mean service rate of an RSVP message at an end host	800 messages/s
λ_d	30 packets/sec	*Data packet size*	200 bytes
τ_0	5 ms	T_{LCoA}	1000 ms
T_{L2}	300 ms	$T_{L2-switch}$	30 ms
T_{RA}	50 ms	T_{LCoA}	1 ms

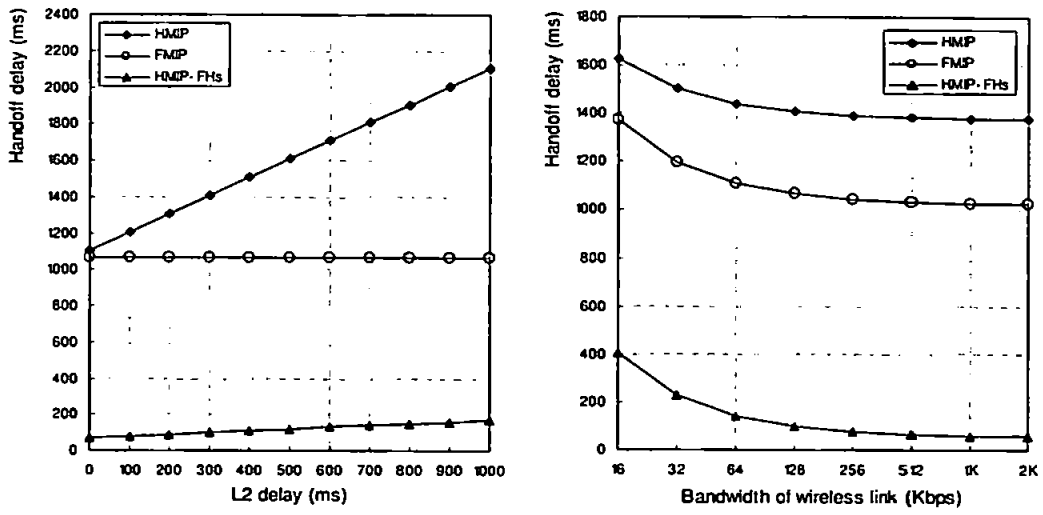
Most of the values are adopted from the literature [Lo etc 2004, Hwang etc 2004, Nakajima etc 2003, McNair etc 2001 and Mishra etc 2003]. Other involved parameters

have been assigned in Table 5.2 of Chapter 5. In addition, we assume that no loop is formed during the MH's itinerary. This assumption is reasonable since it is common that a fast-moving user travels along a road and does not revisit a subnet during a session.

6.5.2.2. Handoff Delay

The micro-handoff delay largely determines the service disruption time in a given micro-mobility support protocol. Figure 6.10 (a) and (b) show the handoff delays of the proposed HMIP-FHs in contrast with other schemes under two varying conditions, respectively. Firstly, the delays increase obviously as the L2 delay increases in all the schemes except in FMIPv6 (a). The reason is that in FMIPv6 the L2 switch delay (10% of the L2 delay assumed) in the MH side is always too small to compensate for the lengthy DAD process in the new AR side within the concerned L2 delay range. Secondly, all the delays decrease when the wireless bandwidth increases since the transmission delay of a message is reduced. In all these situations, the proposed HMIP-FHs have the lowest handoff delays consistently thanks to the introduction of the PAVER scheme. On the other hand, HMIPv6 has the highest handoff delays since no FH mechanism is applied. As to the others, FMIPv6 is much better than HMIPv6. However, the use of DAD results in long delays in FMIPv6, unacceptable for most real-time applications.

When all the involved parameters are set to be their default values, the default handoff delays in HMIPv6, FMIPv6 and the proposed HMIP-FHs are 1407 ms, 1068 ms, and 100 ms, respectively. Clearly, HMIP-FHs dramatically reduce the handoff delays when compared with both standard HMIPv6 and FMIPv6. When the wireless bandwidth is larger than 128 Kbps, the handoff delays in HMIP-FHs are less than 100 ms, which would not noticeably damage the user perceptual QoS of real-time applications.



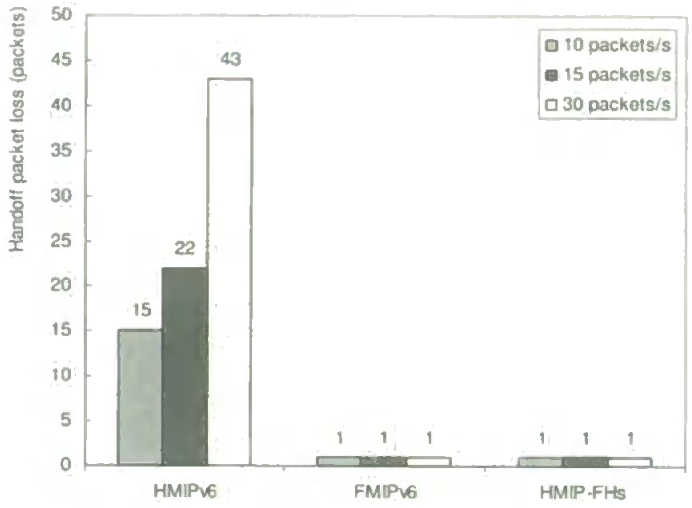
(a) vs. L2 delay (b) vs. bandwidth of wireless link

Figure 6.10 Handoff delay

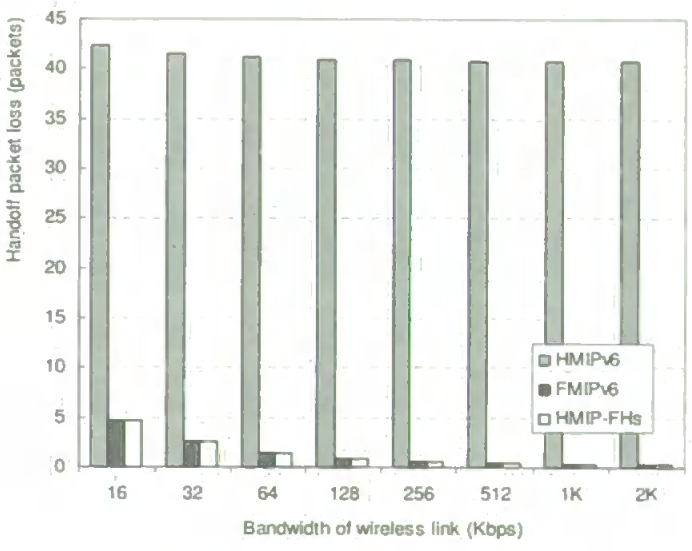
Regarding to the binding delays they contribute a major part to the handoff delays and thus their changing trends resemble those of the handoff delays. For conciseness, figures of binding delays are omitted.

6.5.2.3. Handoff Packet Loss

The handoff packet loss is another important metric influencing the user QoS during handoffs. Firstly, this metric is proportional to the packet arrival rate of an ongoing session as shown in Figure 6.11 (a). Secondly, with a given constant packet arrival rate, this metric is linearly decided by the binding delays in non-FH-enabled schemes, whereas it is jointly determined by τ_0 as well as the round-trip time (*RTT1*) between the MH and the PAR in FH-enabled schemes. Only when *RTT1* is larger than τ_0 will any packet be lost in FH-enabled handoffs. The maximum possible packet loss in the FH-enabled schemes happens when $\tau_0 = 0$ ms. Figure 6.11 (b) presents such an example while the wireless bandwidth varies. Clearly, as depicted in both (a) and (b), only minimum packets are lost in FMIPv6 and the proposed schemes even in this worst scenario, in contrast with the rather large loss in HMIPv6, where FH is not available.



(a) vs. session packet arrival rate



(b) vs. vs. bandwidth of wireless link

Figure 6.11 Handoff packet loss

6.5.2.4. Out-of-Sequence Packets

Figure 6.12 demonstrates the number of the out-of-sequence packets when the packet arrival rate varies. This metric is mainly decided by the handoff delays if no additional mechanism is applied. Again, not surprisingly, HMIP-FHs show the best performance compared with other FMIPv6 even without the proposed REED process. Surely, with

REED the out-of-sequence packets can be eliminated. HMIPv6 does not have this problem simply because no FH is used.

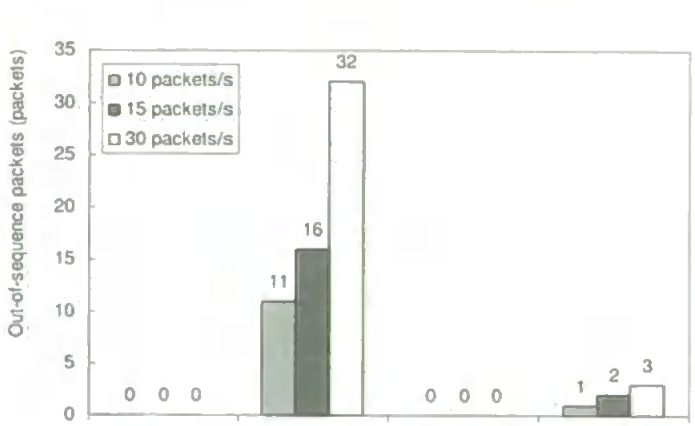


Figure 6.12 Out-of-sequence packets

6.5.2.5. Buffer Size Requirement

The downside of the REED process is to impose additional buffer requirements on HMIP-FHs (except HMIP-FH-neverRO). Fortunately, the additional requirements are insignificant at all. Figure 6.13 illustrates such a result when the subnet resident time of an MH obeys a Gamma distribution with the mean value 20 sec and the variance 20 sec. and up to 30 consecutive handoffs are considered. With the increase of the packet arrival rate (λ_d), the expected buffer size in each protocol tends to be larger. Compared with HMIPv6, the expected additional buffer requirements turn out to be zero when λ_d = 10 or 15 packets/s or just one packet when λ_d = 30 packets/s. This one-packet more buffer size is fairly reasonable and affordable. FMIPv6 requires the largest buffer size due to the much larger handoff delay compared with HMIP-FHs and this is the price paid for minimum handoff loss in FMIPv6. The buffer requirements in HMIPv6 are low since a large number of packets have been simply dropped before the buffering.

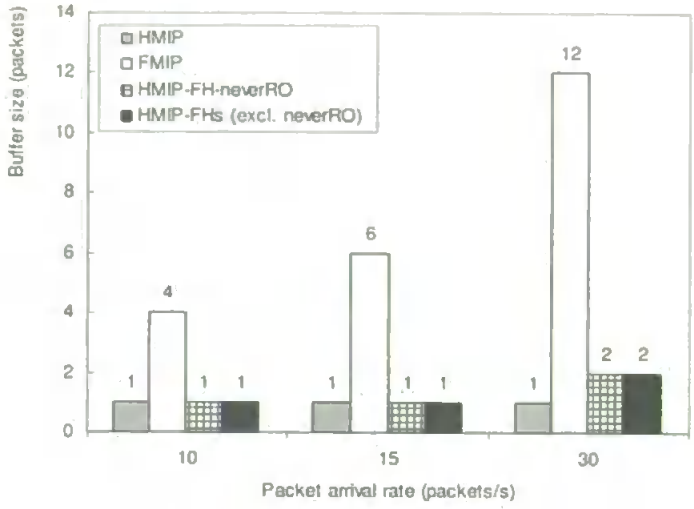


Figure 6.13 Buffer size requirements

6.5.3. Simulation Results

In the above analyses, we have focused on the handoff delay and related metrics. To complement the analytical results, the following simulations are performed to evaluate the signalling costs in the proposed architecture, compared with alternative and related ones. The simulations are developed with Microsoft Visual C++ 7.0.

6.5.3.1. Simulation Configuration

The simulations have the following configurations. The subnet resident time of an MH obeys a Gamma distribution with the mean value 20 sec and the variance 20 sec. The session holding time is exponentially distributed with varying mean values. In each simulation, an MH randomly selects a targeted subnet, which is l hops away from its current subnet in the same domain. The variable l is uniformly distributed over [5, 30]. As each hop between two neighbouring subnets corresponds to one movement, l accumulative micro handoffs occur during the MH's journey. The distance factors (in terms of hops) between the GW-MAP and an AR are 5 or 10 in each simulation, and the distance factors between two neighbouring ARs are 1. The signalling cost generated by a message is

calculated by the product of the message's size and the corresponding distance factor. The data delivery costs are homogeneous to the signalling costs with the mean unit value 100 bytes/sec/hop. The simulations in each of the following scenario are repeated and the averaged results are collected.

6.5.3.2. Signalling Costs for Location Updates at the Central Mobility Server

Firstly, we investigate the signalling costs at a central mobility server since these costs directly affect the scalability of the server, and thus the corresponding mobility support protocol. Figure 6.14 demonstrates the expected location updates costs as an indicator of the signalling costs.

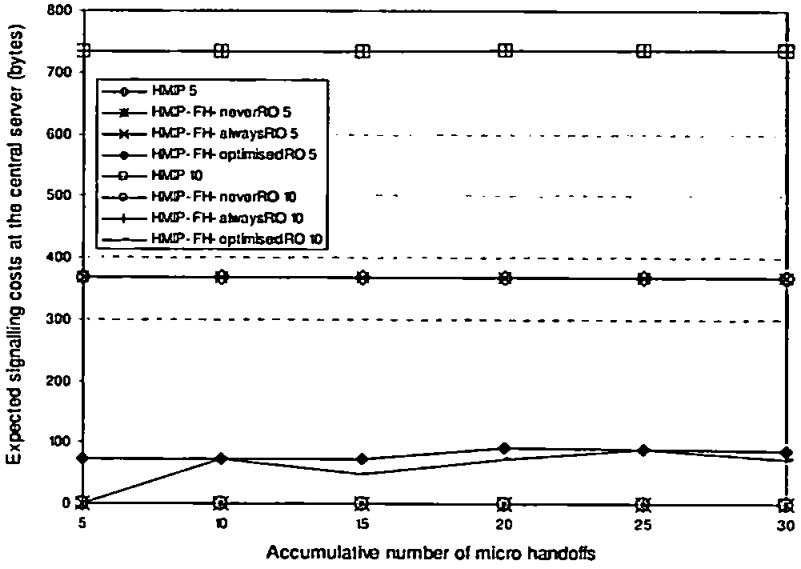


Figure 6.14 Expected signalling costs at the central server

Among the proposed schemes, HMIP-FH-neverRO only updates the ARs and thus the concerned costs are zero; HMIP-FH-alwaysRO performs such location update in each micro handoff and thus generates similar costs as HMIPv6; HMIP-FH-optimisedRO makes a trade-off of these two extremes: it performs this update only when the RO threshold, L_{opt} in this case, is triggered. The hops between MH and GW/MAP, affecting the value of L_{opt}, are shown following the schemes' names. Generally, L_{opt} determines the location updates

frequency in HMIP-FH-optimisedRO and hence the reduction percentage. By applying the cost-driven algorithm, we have L_{opt} = 4 using the default parameter values and When the hops between an MH and the GW-FMS or GW-MAP is changed from 5 to 10, L_{opt} = 8. That is why the reduction percentages against HMIPv6 approach 75% and 87.5% when L_{opt} is 4 and 8, respectively. Of all the schemes, FMIPv6 (over MIPv6) generates the highest costs since the signalling travels globally between the MH and the HA each time on a micro handoff.

6.5.3.3. Accumulative Costs

Next, Figure 6.15 illustrates the accumulative costs for handoff signalling and data delivery during a session; and Figure 6.16 shows the corresponding expected accumulative costs, which are computed by the product of the accumulative costs and the probability that the corresponding number of accumulative handoffs (denoted by K) could happen during the session (this probability is denoted by $P(K)$). As indicated in Figure 6.15, the accumulative costs increase in all the schemes with the growth of K, and the increase of the HMIP-FH-neverRO scheme is most sharp. Regarding the expected accumulative costs, the costs in each scheme except the HMIP-FH-neverRO tend to be much more stable and only vary within a limited range, because $P(K)$ decreases with the increase of K.

As far as the cost reduction is concerned, both figures demonstrate the same degree of improvements in the proposed HMIP-FH-optimisedRO scheme compared with the others. Firstly, HMIP-FH-optimisedRO is consistently more cost-effective than the other two HMIP-FH combinations. When the distance factor between the GW-MAP and an AR is 5, the cost reductions are up to 62% and 10% compared with HMIP-FH-neverRO and HMIP-FH-alwaysRO, respectively. When that distance factor becomes 10, these reductions are up to 44% and 19%, respectively. These changes happen because Phase II becomes more expensive when the domain distance factor increases. Secondly, HMIP-FH-optimisedRO

generates comparable costs as HMIPv6 does, though the proposed scheme tends to outperform HMIPv6 when the domain distance factor is larger. Moreover, low cost as it is, HMIPv6 incurs the largest handoff delays as discussed. Thirdly, the expected costs in neverRO grow sharply with the increase of micro handoff numbers. This is because that the accumulative data delivery costs in HMIP-FH-neverRO soon outweigh the cost saving in signalling via fully distributed operations. On the other hand, due to the opposite reason HMIP-FH-neverRO provokes lower costs than the other schemes except HMIP-FH-optimisedRO when only a few micro handoffs occur during a session. This also explains why HMIP-FH-optimisedRO outperforms the other two combination schemes constantly thanks to its dynamic trade-off between the signalling and data delivery costs.

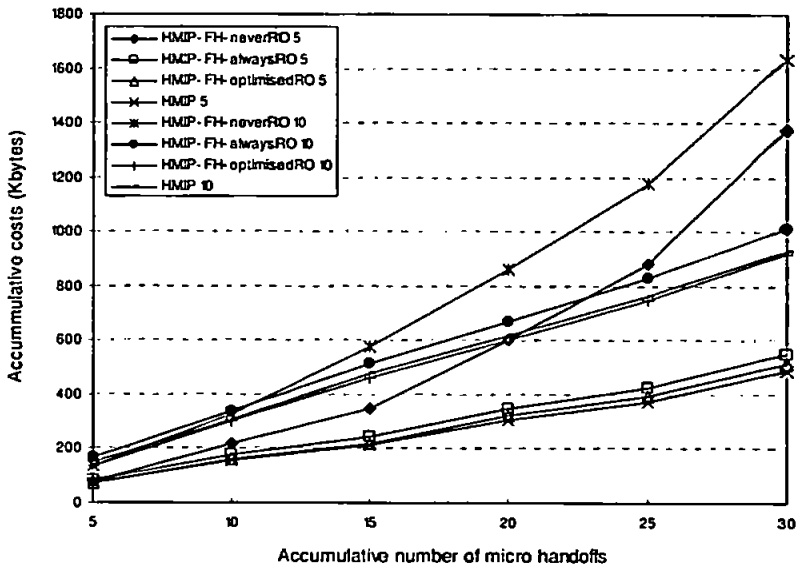


Figure 6.15 Accumulative costs

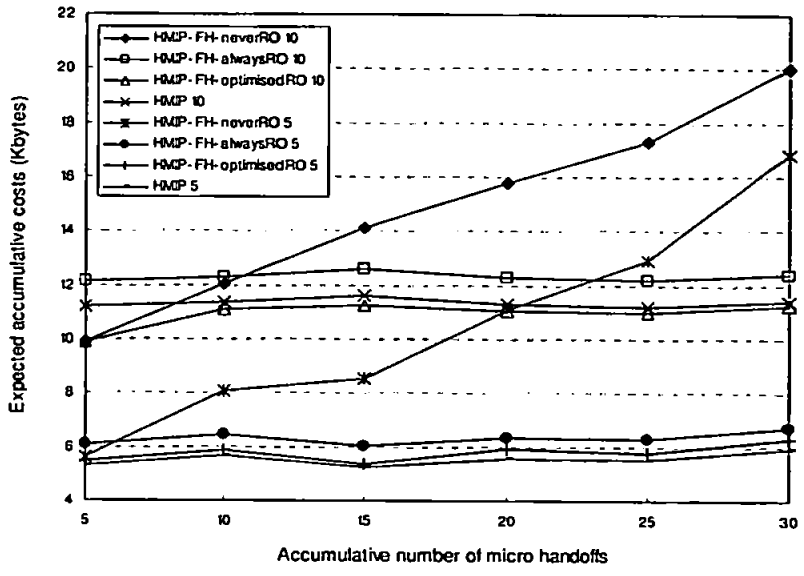


Figure 6.16 Expected accumulative costs

6.6 Concluding Remarks

Combing HMIPv6 and FMIPv6 is an attractive solution to achieving improved micro-mobility management, though in-depth investigations are entailed. Our technical contributions in this chapter are multifold.

Firstly, we proposed a cost-effective micro-mobility architecture, HMIP-FH-optimisedRO, which optimisedly integrates FMIPv6 and HMIPv6. A couple of other combination scenarios (HMIP-FH-neverRO and HMIP-FH-alwaysRO) were also explored. Secondly, we devised the interaction schemes between mobility protocols and QoS protocols. Thirdly, we designed a prompt IPv6 address verification and complementary address replacement scheme, PAVER. The PAVER scheme greatly reduces the handoff delays by replacing the bottleneck in the standard address auto-configuration procedure. Fourthly, we introduced into the architecture another scheme named REED to eliminate out-of-order packets found when FMIPv6 and HMIPv6 are jointly applied.

The overall analytical and simulation results demonstrate that the HMIP-FH-optimisedRO architecture is the most cost-effective one in the three combined HMIPv6 and FMIPv6 architectures (HMIP-FHs). Furthermore, the proposed architecture minimises handoff delays and eliminates out-of-sequence packets at insignificant additional buffer requirements.

Chapter 7

Conclusions and Perspectives

In this chapter, we conclude this thesis with a summary of the main results and achievements obtained from this project, and present the perspectives for future work. Furthermore, the contributions to knowledge are highlighted and the limitations of the work are identified.

7.1 Summary of the Project

In this thesis, we have systematically reported our work on architectures and protocols that support next-generation (Beyond 3G or B3G) mobility in all IP networks. In the following, we summarise the thesis by recalling the main points.

The rapid penetration of both mobile and Internet technologies has led to the new converged communication paradigm over a unified all-IP-based platform. Such a paradigm shift offers great opportunities as well as huge challenges to both industry and academia. One of the fundamental problems that the research community faces is advanced mobility management that supports mobility of different dimensions envisioned for the new paradigm. The solution should support both macro and micro mobility, both real-time and non-real-time applications, both terminal and personal mobility (and potentially other mobility types) in an effective yet efficient way. Numerous architectures and protocols have been proposed in the literature during the last few years to handle part of this problem. However, no existing solution appears to meet all or even most of the requirements

imposed by the advanced mobility support, though some of the promising proposals are being standardised and should be exploited wherever appropriate.

Therefore, the main aim of this project is to explore mobility support architectures and protocols that are capable to satisfy the diverse requirements, and preferably established on top of standards. We cast a cross-layer perspective on the topic that is too demanding to be tackled by a conventional single-layer approach. In general, each protocol layer is affected by mobility whilst in turn it may be convenient for a specific layer to contribute to one or more aspects of mobility handling. In particular, the network layer is able to deal with most of terminal mobility and the application layer is more ready for personal mobility and advanced terminal mobility. In addition, the link-layer can accelerate the handoffs of upper layers. Specifically, Mobile IP (MIP, together with its variants such as HMIP and FMIP) and the Session Initiation Protocol (SIP) have been chosen to deliver the contributions from the network layer and the application layer, respectively. To facilitate the information exchanges related to mobility events, cross-layer signalling methods are needed. A multi-layer framework towards a complete mobility support can thus be envisioned.

The thesis focuses on one of the most essential building blocks of the framework: a macro-mobility solution (complemented by a micro-mobility solution) that supports both terminal and personal mobility with real-time and non-real-time applications. Two novel macro-mobility architectures are proposed based on the integration of MIP and SIP, which dynamically combines and leverages the advantages of both protocols in a cost-effective way. The design motivation and methodology is to make full use of standardised work from both protocols, select composite processes that are more efficient for common functions, integrate or coordinate similar entities and procedures to reduce redundancies, and avoid further duplicate standardisation from each protocol's own perspective. By these means, the system efficiency is greatly improved and the mobility functionalities are

significantly enhanced. In the first architecture, the tightly integrated MIP-SIP or TI-MIP-SIP architecture, the functionalities and entities of both MIP and SIP are converged, yielding the maximum cost-effectiveness. Alternatively, interactions are introduced in the loosely integrated MIP-SIP or LI-MIP-SIP architecture between MIP and SIP entities to achieve a trade-off between efficiency and deployment conveniences. Despite the differences in design details, both architectures are optimised integrations of MIP and SIP infrastructure and protocols combining the best of them, and hence both proposals are able to meet the design challenges with significant cost savings and performance improvements compared with the emerging Hybrid MIP-SIP architectures. Each of the two proposed macro-mobility architectures can operate as standalone solutions to all kinds of mobility scenarios (UDP and TCP mobility, terminal and personal mobility, and even more mobility types), or collaborate with a well-interfaced micro-mobility scheme to improve the handoff performances further. In addition, the design philosophies and methodologies are applicable to both IPv4 and IPv6 contexts.

After the architectural and protocol signalling designs of the proposed TI-MIP-SIP and LI-MIP-SIP architectures, a set of analytical and simulation models and methodologies were then developed to evaluate the proposals and to compare them with other approaches. Both the analytical and simulation results show that the proposed architecture outperforms the Hybrid MIP-SIP mobility approaches. Firstly, the two integrated architectures yield a clear-cut consistent reduction in mobility signalling costs: more than 60% in most cases in TI-MIP-SIP, and up to over 50% in LI-MIP-SIP respectively. Therefore, the system cost-efficiency is dramatically improved. Secondly, the proposed architectures are superior in supporting both UDP and TCP mobility in terms of significantly reduced handoff delays, handoff packet loss, and flexibly enhanced session setup and handoff capabilities etc. with corresponding standard-message-based options proposed. Therefore, the effectiveness of

supporting real-time and non-real-time applications in mobile environments is advanced by the proposed architectures.

Naturally, the next complementary step is to explore a micro-mobility solution that can efficiently support both faster handoffs of real-time applications for high-mobility users and restrict global home registrations at the same time. The standard FMIPv6 and HMIPv6 were chosen as the basis for our design since they appear to be the most promising candidates for micro-mobility support. We approached our design by identifying the shortcomings of the standard FMIPv6 and HMIPv6, and the problems in the existing approaches that combine both protocols. The proposed solution, HMIP-FH-optimisedRO, is built on a cost-driven dynamic combination of FMIPv6 and HMIPv6 with a set of optimisations and enhancements introduced including efficient interworking with QoS signalling and acceleration of IPv6 address auto-configurations etc. The overall analytical and simulation results demonstrate that the HMIP-FH-optimisedRO architecture is the most cost-effective one in the three combined HMIPv6 and FMIPv6 architectures (HMIP-FHs) by reducing up to about 20% and 60% total costs compared with the other two respectively. Furthermore, in contrast to the standard FMIPv6 and HMIPv6, the proposed architecture minimises handoff delays and eliminates out-of-sequence packets at insignificant additional buffer requirements.

In summary, these proposed architectures and protocols can support diverse mobility scenarios such as macro and micro mobility, UDP and TCP mobility, terminal and personal mobility (and potentially additional mobility types) effectively and efficiently. They outperform the competing approaches in terms of significantly higher cost-efficiency and superior handoff performances, evaluated through extensive theoretical analyses and software simulations.

7.2 Contributions to Knowledge

7.2.1. Technical Contributions

The work performed in this project has made distinct contributions to the current knowledge of mobility support and related areas in the following aspects.

1. **An original generic multi-layer approach for comprehensive mobility support**

   Most of the previous proposals for mobility management are based on a single protocol layer and thus they can hardly fulfil the complex collective requirements of next-generation (B3G) mobility imposed on almost each layer. The prospected multi-layer framework attempts to exploit the mobility-related contributions from each layer and combine each layer's powerfulness in a coordinated way as a joint effort. The framework defines potential interactions between multiple layers for a cooperative mobility support. Although the thesis emphasises a combined work of the network and the application layers together with L2 triggers, other combinations are possible. Thus, this cross-layer perspective opens up alternative approaches to meeting the next-generation mobility challenges.

2. **A new generic and efficient cross-layer signalling method**

   Cross-layer signalling methods are essential to achieve information exchanges across a protocol stack for many cross-layer designs and optimisations. Previously, several methods were outlined sparsely in the literature and their pros and cons were unknown to the research community. We have filled the gap with an investigation and comparison. More importantly, the existing methods do not appear suitable to serve as a generic and efficient mechanism. In contrast, the proposed method seems more promising to support time-stringent and

complicated upwards and downwards interlayer messaging. Notably, the studies on cross-layer signalling can benefit a broad range of areas where cross-layer designs are desired and thus the contributions are not limited to the mobility support subject.

3. **A novel macro-mobility architecture that tightly integrates MIP and SIP**

   In light of the complementary powerfulness of MIP and SIP in advanced mobility support, an integration of both protocols is entailed. While this idea is also being explored in some emerging proposals, MIP and SIP were typically utilised independently even on a same platform. Such a hybrid approach simplifies the deployment whereas it invokes enormous costs from redundant entities, functionalities and parallel large-scale signalling, which would seriously deteriorate the system performances. In contrast, the proposed tightly integrated architecture dramatically decreases the costs by minimising the redundancies in a unified architecture and maximises the efficiency in a long run. The design philosophies are to make full use of standardised work from both protocols, select composite processes that are more efficient for common functions, integrate similar entities and procedures to reduce redundancies, and avoid further duplicate standardisation. The originality of the work is reflected by not only these design philosophies but also design details such as the decomposition of similar entities and the reconstruction of integrated mobility servers, the unified address management, the selective reuse of MIP and SIP messages and proposed options for session setup and route optimisation etc. These design philosophies and methodologies are applicable to both IPv4 (MIPv4) and IPv6 (MIPv6) cases though the IPv6 context is focused on. Moreover, the design principles could be applied to other integration scenarios of similar areas.

4. **A novel macro-mobility architecture that loosely integrates MIP and SIP**

   The major novelty of this architecture lies in an alternative approach to a highly cost-efficient integration of MIP and SIP without merging their physical entities. This is achieved by establishing necessary interactions between MIP and SIP servers. A couple of schemes are devised to provide such interactions based on MIP and/or SIP messages though the reuse of MIP messages is demonstrated. Similar to the tightly integrated architecture, this architecture supports both terminal and personal mobility and both UDP (real-time applications) and TCP (non-real-time applications) mobility at a price slightly higher than the tightly integrated architecture yet far lower than the hybrid ones. Therefore, the loosely integrated architecture may act as an intermediate step towards a full integration of MIP and SIP in the shorter term.

5. **A novel micro-mobility architecture that optimises the combination of HMIPv6 and FMIPv6 with performance improvement enhancements**

   Although HMIPv6 and FMIPv6 are being developed independently, a combination of both protocols would share their mutual strengths in a unified architecture for micro-mobility support. Nevertheless, existing combination approaches either simply superimpose the two protocols or only catered for domains with simple hierarchy. The former approach is not cost-effective and the latter does not consider complex hierarchy or high-mobility users. The proposed architecture resolves these problems by dynamically combining enhanced HMIPv6 and FMIPv6 with a set of optimisations to achieve improved performances such as faster handoffs compared with standard FMIPv6 and lower accumulative costs in contrast to two other typical combined approaches. The

associated optimisation techniques, especially the mechanism to accelerate IPv6 address auto-configurations, would be applicable to other architectures.

6. **Useful methodologies and models for analyses and evaluations on this subject**

   During the project, a set of methodologies and models for analyses and evaluations are developed. Based on the literature, especially those presented in premier journals, the analytical models are refined more or less to evaluate the performances such as costs and delays of the proposals and existing ones from more angles (with more metrics), in more details (with more parameters), or more accurately (with more typical configurations). Meanwhile, reusable simulation models are developed to complement and/or verify the analytical results. Both the analytical and simulation models along with the evaluation methodologies can serve as useful tools for future research on this subject.

7.2.2. Contributions to Literature

Part of the work has been disseminated to the research community through nine publications (except [Lopez etc QoS04], whose contents are not included in this thesis), which have enriched the literature on the subject of mobility support and related areas. Each of these publications is (or will be) indexed or abstracted by one or more leading bibliographic databases such as SCI, EI, ISTP, INSPEC etc., and the full texts of the electronic versions of the papers are (or will be) available in IEE/IEEE digital libraries or the publishers' on-line services.

Furthermore, according to a non-exhaustive search via on-line search engines including Google, Yahoo etc., the six earlier publications ([Wang and Abu-Rgheff LCS02, WCNC03, EPMCC03, 3G2003, ICC04, CE]) alone had been referenced by peer

researchers worldwide for more than 50 times by the end of January 2006. Part of the citations are found in premier journals such as IEEE Transactions on Vehicular Technology, Proceedings of IEEE and IEEE Communications Magazine, and top conferences such as Globecom'04 and WCNC'05. The other references are in other journals, conference proceedings, deliverables of IST (Information Society Technologies) projects such as 4MORE and FLOWS, technical reports, postgraduate theses and proposals, workshop tutorials, seminar presentations, on-line articles, and teaching or research reading lists. Briefly, the above evidence indicates that quite a few peer researchers of relevant areas have benefited from our contributions to knowledge.

Finally, it is understandable that it takes some time to circulate the three latest publications ([Wang and Abu-Rgheff 3G2004, 3G2005 and IJCS]) before they can obtain any citations to enlarge the non-self citation list, which is growing in size. In addition, a couple of more papers are in preparation for publication to report more results discussed in the thesis.

7.3 Limitations of the Current Work

Although every effort has been taken to ensure a comprehensive work, we are aware of the following limitations, some of which may be addressed in the future work.

7.3.1. Limited Considerations on System Diversity

The diversity in networks and terminals are not explicitly addressed under the all IP umbrella. The protocol designs have followed the IETF all-IP paradigm, and thus some system-specific mobility-related operations such as UMTS PDP management and their interworking with pure IP mobility protocols are not addressed. Regarding terminals, the capabilities of different type of terminals are not considered.

7.3.2. Limited Analyses on Negative Effects of Cross-Layer Design

The thesis has advocated a cross-layer design on mobility support and demonstrated the performance improvements through such as a methodology. The involved cross-layer signalling is limited and well controlled in the specific designs in Chapters 4, 5 and 6. Nevertheless, more considerations should be given to achieve the envisioned multi-layer mobility support framework, where complex cross-layer signalling takes place, and the possible negative effects of cross-layer design need to be further investigated.

7.3.3. Limited Validation of the Work

We have proposed a number of novel architectures and protocols in the thesis. Although most of the proposals have been evaluated through theoretical analyses and/or simulations, the mobility models used are limited and more experiments, especially large-scale and cross-proposal ones, may still be needed. Preferably, the proposals are implemented and validated in a real-world testbed.

7.4 Future Work

In light of the limitations of the current work and the possible extensions, the following selected projects may be undertaken as future work.

7.4.1. Support for Additional Mobility Types

This thesis is focused on terminal and personal mobility with an emphasis on the former because it is still the dominant mobility types in the near future. Other mobility types are briefly discussed and may be investigated in more details. For example, handoffs between terminals of different types, e.g., from cellular phone to laptop, in the session mobility support may deserve a further design with the diversity in their capabilities

considered. Furthermore, a costs analysis on the introduction of additional mobility types may also be conducted to understand the prices the system has to pay for these additional services.

7.4.2. Comprehensive Cross-Layer Design

Another natural extension of the current work is to further explore the intricate multi-layer framework for comprehensive mobility support including advanced QoS considerations especially QoS adaptation to mobility. A policy-based global controller to the whole protocol stack may be designed to trigger cross-layer signalling, coordinate cross-layer optimisation behaviours, detect and avoid potential conflicts when complicated cross-layer operations are running simultaneously. Both the positive and negative effects of cross-layer designs should be evaluated. In addition, it is worthy of investigating the interactions with the next-generation QoS signalling protocol NSIS and QoS routing algorithms [Friderikos etc 2004].

7.4.3. Interactions with AAA Protocols

In the IETF, the interactions of AAA protocols such as Diameter [RFC3588] with either MIP [RFC4004] or SIP [Garcia-Martin etc 2005] are under investigations, respectively, in a separate way. However, it is desirable to devise a unified architecture for efficient AAA interactions with both MIP and SIP simultaneously or alternately, especially when the integrated MIP-SIP architectures proposed in the thesis are considered. In addition, the AAA interactions with HMIPv6 and FMIPv6, e.g., through context transfer, in the micro-mobility scenario also deserve a further study.

7.5 Conclusions

Finally, we reach our conclusions. In this thesis, we have shown a systematic study on mobility support for next-generation all IP networks.

We introduced this project by presenting the motivations, the aim and objectives, etc. in Chapter 1. We then surveyed the current literature on mobility support and criticised the existing work in Chapter 2. In Chapter 3, we outlined the requirements of next-generation (B3G) mobility support and the corresponding design challenges, and the rationale of cross-layer design to meet the requirements and challenges. We proposed a new method for generic and efficient cross-layer signalling. Furthermore, we prospected a multi-layer framework as the direction to achieve a comprehensive mobility support. Lastly, we identified the major building blocks of the framework and specified the focus of the project.

In the subsequent three chapters (i.e., Chapters 4 - 6), we proposed and evaluated two macro-mobility architectures and a micro-mobility architecture, respectively. These architectures are mainly built atop of standardised protocols including MIPv4, MIPv6, SIP, HMIPv6 and FMIPv6 with a set of integration and optimisation strategies. The detailed architectural and protocol signalling designs are expounded and their performances are evaluated through theoretical analyses and software simulations.

The numerical results and analyses have indicated that the proposed architectures are promising candidates to support various mobility scenarios expected for the next-generation mobility support both efficiently and effectively. Recalling the research questions posted in the motivation section in Chapter 1, we expect that our work through this project has advanced the knowledge to answer these questions by making multifold contributions to this subject.

Bibliography

1-9

[3GPP-25215] 3GPP TS 25.215 V5.1.0, "Physical Layer - Measurements (FDD)," September 2002.

[3GPP-23119] 3GPP TS 23.119 V5.0.0, "Gateway Location Register (GLR) - Stage 2," June 2002.

[3GPP2-PR0001] 3GPP2 P.R0001 V1.0.0, "Wireless IP Network Architecture based on IETF protocols," July 2000.

[3GPP-24229] 3GPP TS 24.229 V6.2.0, "IP Multimedia Call Control Protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP)," March 2004.

[3GPP2-XP0013] 3GPP2 X.P0013-004-0 V2.0, "All-IP Core Network Multimedia Domain: IP Multimedia Call Control Protocol Based on SIP and SDP; Stage 3," January 2005.

[3GPP2-PS0001] 3GPP2 P.S0001-B V1.0.0, "Wireless IP Network Standard," October 2002.

[3GPP-23923] 3GPP TR 23.923 V3.0.0, "Combined GSM and Mobile IP Mobility Handling in UMTS IP CN," May 2000.

[3GPP-GP508] 3GPP TSG GERAN, "A Comparison between Packet-Switched Call Setup Using SIP and GSM Circuit-Switched Call Setup Using RIL3-CC, RIL3-MM, RIL3-RR, and DTAP", GP-000508, Nortel Networks, Nov 2000.

A

[Abolhasan etc 2004] Abolhasan M, T. Wysocki, and E. Dutkiewicz, "A Review of Routing Protocols for Mobile Ad Hoc Networks," *Ad Hoc Networks*, vol.2, no. 1, pp. 1-22, January 2004.

[Akyildiz etc 1996] Akyildiz I. F., J. S. M. Ho, and Y.-B. Lin, "Movement-Based Location Update and Selective Paging for PCS Networks," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 629-638, August 1996.

[Akyildiz etc 1998] Akyildiz I. F., J. McNair, J. Ho, H. Uzunalioglu, and W. Wang, "Mobility Management in Current and Future Communications Networks," *IEEE Network*, vol. 12, no. 4, pp. 39-49, July/August 1998.

[Akyildiz and Wang 2002] Akyildiz I. F. and W. Wang, "A Dynamic Location Management Scheme for Next Generation Multi-Tier PCS Systems," *IEEE Transactions on Wireless Communications*, vol. 1, no. 1, pp. 178-189, January 2002.

[Alam etc 2001] Alam M., R. Prasad, and J. Farserotu, "Quality of Service among IP-Based Heterogeneous Networks," *IEEE Personal Communications*, vol. 8, no. 6, pp. 18-24, December 2001.

[Aretz etc 2001] Aretz K., M. Haardt, W. Konhäuser, and W. Mohr, "The Future of Wireless Communications Beyond the Third Generation," *Computer Networks*, vol. 37, no. 1, pp. 83-92, September 2001.

[Aust etc 2003] Aust S., D. Proetel, N. A. Fikouras, C. Pampu, and C. Görg, "Policy based Mobile IP Handoff Decision (POLIMAND) using Generic Link Layer Information," in *Proc. of the 5th IEEE International Conference on Mobile and Wireless Communication Networks (MWCN'03)*, Singapore, October

2003.

B

[Bafutto etc 1994] Bafutto M., P. J. Khn, and G. Willmann, "Capacity and Performance Analysis of Signalling Networks in Multivendor Environments," *IEEE Journal on Selected Areas in Communication*, vol. 12, no. 3, pp. 490-500, April 1994.

[Balakrishnan etc 1997] Balakrishnan H., V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 756-769, December 1997.

[Balakrishnan 1998] Balakrishnan H., "Challenges to Reliable Data Transport over Heterogeneous Wireless Networks," *Ph.D. Thesis*, the University of California at Berkeley, USA, 1998.

[Banerjee etc 2004] Banerjee N., W. Wu, K. Basu, and S. K. Das, "Analysis of SIP-based Mobility Management in 4G Wireless Networks," *Computer Communications*, vol. 27, no. 8, pp. 697-707, 2004.

[Berezdivin etc 2002] Berezdivin R., R. Breinig, and R. Topp, "Next-Generation Wireless Communications Concepts and Technologies," *IEEE Communications Magazine*, vol. 40, no. 3, pp. 108-16, March 2002.

[Bernardos etc 2005] Bernardos C. J., I. Soto, J. I. Moreno, T. Melia, M. Liebsch, and R. Schmitz, "Experimental Evaluation of A Handover Optimization Solution for Multimedia Applications in a Mobile IPv6 Network," *European Transactions on Telecommunications*, vol. 16, no. 4, pp. 317-328, 2005.

[Bernaschi etc 2004] Bernaschi M., F. Cacace, and G. Iannello, "Vertical Handoff Performance in Heterogeneous Networks," in *Proc. of IEEE International Conference on Parallel Processing Workshops (ICPPW'04)*, Montreal, Canada, August 2004.

[Bhagwat etc 2002] Bhagwat P., D. A. Maltz, and A. Segall, "MSOCKS+: An Architecture for Transport Layer Mobility," *Computer Networks*, vol. 39, no. 4, pp. 385-403, July 2002.

[Bos and Leroy 2001] Bos L. and S. Leroy, "Toward an All-IP-Based UMTS System Architecture," *IEEE Network*, vol. 15, no. 1, pp. 36–45, January/February 2001.

C

[Campbell and Castellanos 2000] Campbell A. T. and J. Castellanos, "IP Micro-Mobility Protocols," *ACM Mobile Computing and Communications Review*, vol. 4, no. 4, pp. 45-53, October 2000.

[Campbell etc 2000] Campbell A. T., J. Gomez, S. Kim, A. G. Valkó, C.-Y. Wan, and Z. R. Turányi, "Design, Implementation, and Evaluation of Cellular IP," *IEEE Personal Communications*, vol. 7, no. 4, pp. 42-49, August 2000.

[Campbell etc 2002] Campbell A. T., J. Gomez, S. Kim, C.-Y. Wan, Z. R. Turanyi, and A. G. Valko, "Comparison of IP Micromobility Protocols," *IEEE Wireless Communications*, vol. 9, no. 1, pp. 72-82, February 2002.

[CE2005] "Mobile Subs on the Up... for Now," Cellular News, *IEE Communications*

Engineer, vol. 3, no. 3, pp.5, June/July 2005.

[Chen etc 2002] Chen K., S. H. Shan, and K. Nahrstedt, "Cross-Layer Design For Data Accessibility in Mobile Ad Hoc Networks," *Wireless Personal Communications*, vol. 21, no. 1, pp. 49-76, April 2002.

[Chiussi etc 2002] Chiussi F. M., D. A. Khotimsky, S. Krishnan, "Mobility Management in Third-Generation All-IP Networks," *IEEE Communications Magazine*, vol. 40, no. 9, pp. 124-135, September 2002.

[Curran and Parr 2002] Curran K. and G. Parr, "A Framework for the Transmission of Streaming Media to Mobile Devices," *International Journal of Network Management*, vol. 12, no. 1, pp. 41-59, January/February 2002.

D

[Das etc 2000] Das S., A. Misra, P. Agrawal, and S. K. Das, "TeleMIP: Telecommunications-Enhanced Mobile IP Architecture for Fast Intradomain Mobility," *IEEE Personal Communications*, vol. 7, no. 4, pp. 50-58, August 2000.

[Das etc 2002] Das S., A. Mcauley, A. Dutta, A. Misra, and S. K. Das, "IDMP: An Intra-Domain Mobility Management Protocol for Next Generation Wireless Networks," *IEEE Wireless Communications*, vol. 9, no. 3, pp. 38-45, June 2002.

[Dimopoulou etc 2004] Dimopoulou L., G. Leolieis, and I. S. Venieris, "Introducing a Hybrid Fast and Hierarchical MIPv6 scheme in a UMTS-IP converged Architecture," in *Proc. of the 29th IEEE International Conference on Local Computer Networks (LCN'04)*, Sydney, Australia, December 2004.

[Dutta etc 2001A] Dutta A., R. Jain, D. Wong, J. Burns, K. Young and H. Schulzrinne, "Multilayered Mobility Management for Survivable Network," in *Proc. of IEEE MILCOM'01*, Virginia, USA, October 2001.

[Dutta etc 2001B] Dutta A., F. Vakil, J. C. Chen, M. Tauil, S. Baba, and H. Schulzrinne, "Application Layer Mobility Management Scheme for Wireless Internet," in *Proc. of 3Gwireless'01*, San Francisco, USA, May 2001.

E

[Eom etc 2002] Eom D. S., H. Lee, M. Sugano, M. Murata and H. Miyahara, "Improving TCP Handoff Performance in Mobile IP Based Networks," *Computer Communications*, vol. 25, no. 7, pp. 635-646, May 2002.

[Evans and McLaughlin 2000] Evans G. and S. McLaughlin, "Visions of 4G," *IEE Electronics & Communication Engineering Journal*, vol. 12, no. 6, pp. 293-303, December 2000.

[EVOLUTE] The EVOLUTE project, http://evolute.intranet.gr/

[E-consultancy 2005] E-consultancy, "Internet Statistics Compendium," March 2005.

F

[Faccin etc 2004] Faccin S. M., P. Lalwaney, and B. Patil, "IP Multimedia Services: Analysis of Mobile IP and SIP Interactions in 3G Networks," *IEEE*

Communications Magazine, vol. 42, pp. 113- 120, January 2004.

[Fasbender etc 1999] Fasbender A., F. Reichert, E. Geulen, J. Hjelm, and T. Wierlemann, "Any Network, Any Terminal, Anywhere," *IEEE Communications Magazine*, vol. 6, no. 2, pp. 22-30, April 1999.

[Festag 2002] A. Festag, "Optimisation of Handover Performance by Link Layer Triggers in IP-Based Networks: Parameters, Protocol Extensions and APIs for Implementation," *Technical Report, Version 1.0, TKN-02-014*, Technical University Berlin, August 2002.

[Fikouras etc 2001] Fikouras N. A., A. J. Könsgen, and C. Görg, "Accelerating Mobile IP Handoffs through Link-Layer Information," in *Proc. of the International Multiconference on Measurement, Model-ling, and Evaluation of Computer-Communication Systems (MMB'01)*, Aachen, Germany, September 2001.

[Friderikos etc 2004] Friderikos V., A. Mihailovic, and A. H. Aghvami, "Analysis of Cross Issues between QoS Routing and μ-Mobility Protocols," *IEE Proceedings - Communications*, vol. 151, no. 3, pp. 258-262, June 2004.

G

[Gao etc 2004] Gao X., G. Wu, and T. Miki, "End-to-End QoS Provisioning in Mobile Heterogeneous Networks," *IEEE Wireless Communications*, vol. 11, no. 3, pp. 24-34, June 2004.

[Garcia-Martin etc 2005] Garcia-Martin M., M. Belinchon, M. Pallares-Lopez, C. Canales, K. Tammi, "Diameter Session Initiation Protocol (SIP) Application," *IETF Internet Draft*, draft-ietf-aaa-diameter-sip-app-10, work in progress, October 2005.

[Goff and Phatak 2004] Goff T. and D. S. Phatak, "Unified Transport Layer Support for Data Striping and Host Mobility," *IEEE Journal On Selected Areas In Communications*, vol. 22, no. 4, pp. 737-746, May 2004.

[GSMA34] GSM Association, "Inter-PLMN Backbone Guidelines," *IR.34*, v. 3.5.2, August 2004.

[Gustafsson etc 2004] Gustafsson E., A. Jonsson, and C. Perkins, "Mobile IPv4 Regional Registration," *IETF Internet Draft*, draft-ietf-mip4-reg-tunnel-00, work in progress, November 2004.

[Gustafsson and Jonsson 2003] Gustafsson E. and A. Jonsson, "Always Best Connected," *IEEE Wireless Communications*, vol. 10, no. 1, pp. 49-55, Feb. 2003.

H

[Haas 2001] Haas Z. H., "Design Methodologies for Adaptive and Multimedia Networks," Guest Editorial, *IEEE Communications Magazine*, vol. 39, no. 11, pp. 106-107, November 2001.

[Haratcherev etc 2005] Haratcherev I., J. Taal, K. Langendoen, R. Lagendijk, and H. Sips, "Fast 802.11 Link Adaptation for Real-Time Video Streaming by Cross-Layer Signalling," in *Proc. of IEEE ISCAS'05*, Kobe, Japan, May 2005.

[Haverinen and Malinen 2000] Haverinen H. and J. Malinen, "Mobile IP Regional Paging", *IETF Internet draft*, draft-haverinen-mobileip-reg-paging-00.txt, work in progress, June 2000.

[Hsieh etc 2002] Hsieh R., A. Seneviratne, H. Soliman, and K. El-Malki, "Performance Analysis on Hierarchical Mobile IPv6 with Fast-handoff over End-to-End TCP," in *Proc. of IEEE GLOBECOM'02*, Taipei, China, 2002.

[Hsieh etc 2003] Hsieh R., Z. Zhou, and A. Seneviratne, "S-MIP: A Seamless Handoff Architecture for Mobile IP," in *Proc of IEEE Infocom'03*, San Francisco, USA, 2003.

[Hwang etc 2004] Hwang S.-H., Y.-H. Han, C.-S. Hwang, and S.-G. Min, "An Address Configuration and Confirmation Scheme for Seamless Mobility Support in IPv6 Network," *LNCS*, vol. 2957, pp.74-86, February 2004.

I

[IEEE802.21] IEEE P802.21™/D00.01, Draft IEEE Standard for Local and Metropolitan Area Networks: Media Independent Handover Services, July 2005.

[ITU114] ITU-T Recommendation G.114, "One-Way Transmission Time," May 2000.

[ITU2002] ITU-T, "Vision, Framework and Overall Objectives of the Future Development of IMT-2000 and Systems Beyond IMT-2000", 2002.

J

[Jain etc 1999] Jain R., T. Raleigh, D. Yang, L.-F. Chang, C. Graff, M. Bereschinsky, and M. Patel "Enhancing Survivability of Mobile Internet Access using Mobile IP with Location Registers," in *Proc. of IEEE INFOCOM'99*, New York, USA, March 1999.

[Jönsson etc 2000] Jönsson U., F. Alriksson, T. Larsson, P. Johansson, and G. Q. Maguire Jr., "MIPMANET – Mobile IP for Mobile Ad Hoc Networks," in *Proc. of IEEE/ACM MobiHoc'00*, pp. 75-85, 2000.

[Jung and Koh 2004] Jung H.-Y. and S.-J. Koh, "Fast Handover Support in Hierarchical Mobile IPv6," in *Proc. of 6th International Conference on Advanced Communication Technology*, Phoenix Park, Korea, February 2004.

[Jung etc 2003] Jung J.-W., H.-K. Kahng, R. Mudumbai, and D. Montgomery, "Performance Evaluation of Two Layered Mobility Management using Mobile IP and Session Initiation Protocol," in *Proc. Of IEEE GLOBECOM'03*, San Francisco, USA, December 2003.

K

[Kaaranen etc 2001] Kaaranen H., S. Naghian, L. Laitinen, A. Ahtiainen, and V. Niemi, *UMTS Networks: Architecture, Mobility and Services*, John Wiley & Sons Ltd., 2001.

[Kari 1999] Kari H. H., "On Wireless Future", Presentation at MediaPoli seminar, November 1999.

[Kawadia and Kumar 2005] Kawadia V. and P.R. Kumar, "A Cautionary Perspective on Cross-Layer Design," *IEEE Wireless Communications*, vol. 12, no. 1, pp. 3-11, February 2005.

[Kempf and Mutaf 2003] Kempf J. and P. Mutaf, "IP Paging considered unnecessary: Mobile IPv6 and IP paging for dormant mode location update in macrocellular and hotspot networks," in *Proc. of IEEE WCNC'03*, New Orleans, USA,

March 2003.

[Keszei etc 2001] Keszei, C., N. Georganopoulos, Z. Turanyi, and A. Valko, "Evaluation of the BRAIN Candidate Mobility Management Protocol," in *Proc. of IST Mobile Summit'01* , Barcelona, Spain, September 2001.

[Kim 2001] Kim B.-J. 'J', "A Network Service Providing Wireless Channel Information for Adaptive Mobile Applications: Part I: Proposal," in *Proc. of IEEE ICC'01*, pp. 1345-1351, Helsinki, Finland, June 2001.

[Kim and Kim 2003] Kim T.-S. and S. Kim, "A Fast Rerouting Scheme using Reservation in Wireless ATM," *IEEE Transactions on Vehicular Technology*, vol. 52, no.4, pp. 1125-1142, July 2003.

[Kleinrock 1976] Kleinrock L., *Queuing Systems*, vol.2, Wiley, 1976.

[Koh etc 2004] Koh S. J., M. J. Chang, and M. Lee, "mSCTP for Soft Handover in Transport Layer," *IEEE Communications Letters*, vol. 8, no. 3, pp. 189-191, March 2004.

[Kohler etc 2005] Kohler E., M. Handley, and S. Floyd, "Datagram Congestion Control Protocol," *IETF Internet draft*, draft-ietf-dccp-spec-11.txt, work in progress, March 2005.

[Korhonen 2001] Korhonen J., *Introduction to 3G Mobile Communications*, London: Artech House, 2001.

[Kwon etc 2002] Kwon T. T., 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.

L

[La Porta etc 1996] La Porta T. F., M. Veeraraghavan, and R. W. Buskens, "Comparison of Signalling Loads for PCS Systems," *IEEE/ACM Transactions on Networking*, vol. 4, no. 6, pp. 840-856, December 1996.

[Larzon etc 2002] Larzon L-Å, U. Bodin, and O. Schelen, "Hints and Notifications," in *Proc. Of IEEE WCNC'02*, pp. 559-565, Orlando, Florida, USA, March 2002.

[Lee etc 2001] Lee G. C., T. P. Wang, and C. C. Tseng, "Resource Reservation with Pointer Forwarding Schemes for the Mobile RSVP," *IEEE Communications Letters*, vol. 5, no. 7, pp. 298–300, July 2001.

[Lee etc 2003] Lee H., S. W. Lee, and D.-H. Cho, "Mobility Management Based on the Integration of Mobile IP and Session Initiation Protocol in Next Generation Mobile Data Networks," in *Proc. of IEEE VTC'03-Fall*, Orlando, Florida, USA, October 2003.

[Lee and Akyildiz 2003] Lee Y. J. and I. F. Akyildiz, "A New Scheme for Reducing Link and Signaling Costs in Mobile IP," *IEEE Transactions on Computers*, vol. 52, no. 6, pp. 706-713, June 2003.

[Li etc 1997] Li B., S. Jiang, and D. Tsang, "Subscriber-Assisted Handoff Support in Multimedia PCS," *ACM Mobile Computing and Communications Review*, vol. 1, no. 3, pp. 29-36, September 1997.

[Lin and Chlamtac 2001] Lin Y.-B. and I. Chlamtac, "Wireless and Mobile Network Architecture," John Wiley & Sons, Inc., 2001.

[Lin etc 1994] Lin Y.-B., S. Mohan, and A. Noerpel, "Queuing Priority Channel Assignment Strategies for Handoff and Initial Access for a PCS Network,"

IEEE Transactions on Vehicular Technology, vol. 43, no. 3, pp. 704-712, August 1994.

[Lin etc 2001] Lin Y.-B., Y.R. Haung, Y.K. Chen and I. Chlamtac, "Mobility Management: from GPRS to UMTS," *Wireless Communications and Mobile Computing*, vol. 1, no. 4, pp. 339-359, 2001.

[Lo etc 2004] Lo S.-C., 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.

M

[Mahonen etc 2004] Mahonen P., J. Riihijarvi, M. Petrova, and Z. Shelby, "Hop-by-Hop toward Future Mobile Broadband IP," *IEEE. Communications Magazine*, vol. 42, no. 3, pp. 138-146, March 2004.

[Malinen and Perkins 2001] Malinen J. and C. Perkins, "Mobile IPv6 Regional Registrations," *IETF Internet Draft*, draft-malinen-mobileip-regreg6-01.txt, work in progress, March 2001.

[Maniatis etc 1999] Maniatis P., M. Roussopoulos, E. Swierk, K. Lai, G. Appenzeller, X. Zhao, and M. Baker, "The Mobile People Architecture," *ACM Mobile Computing and Communications Review*, vol. 3, no. 3, pp. 36-42, July 1999.

[Manzoni etc 1995] Manzoni P., D. Ghosal, and G. Serazzi, "Impact of Mobility on TCP/IP: An Integrated Performance Study," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 5, pp. 858-867, June 1995.

[McAuley etc 2000] McAuley A., S. Das, S. Madhani, S. Baba, and Y. Shobatake, "Dynamic Registration and Configuration Protocol," *IETF Internet Draft*, draft-itsumo-drcp-01.txt, work in progress, July 2000.

[McNair etc 2000] McNair J., I.F. Akyildiz, and M. Bender, "An Inter-System Handoff Technique for the IMT-2000 System," in *Proc. of INFOCOM'00*, pp. 208-216, Tel Aviv, Israel, March 2000.

[McNair etc 2001] McNair J., I. F. Akyildiz, and M. Bender, "Handoffs for Real-Time Traffic in Mobile IP Version 6 Networks," in *Proc. of IEEE Globecom'01*, pp. 3463-3467, San Antonio, Texas, USA, November 2001.

[Mishra etc 2003] Mishra W. A. A., M. Shin, and W. Arbaugh, "An Empirical Analysis of the IEEE 802.11 MAC Layer Handoff Process," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 2, pp. 93-102, April 2003.

[MobyDick] EU IST Moby Dick project: Mobility and Differentiated Services in a Future IP Network, http://www.ist-mobydick.org.

[Moon and Aghvami 2001] Moon B. and A. H. Aghvami, "RSVP Extensions for Real-Time Services in Wireless Mobile Networks," *IEEE Communications Magazine*, vol. 39, no. 12, pp.52–59, December 2001.

[Moon and Aghvami 2003] Moon B. and H. Aghvami, "DiffServ Extensions for QoS Provisioning in IP Mobility Environments," *IEEE Wireless Communications*, vol. 10, no. 5, pp. 38-44, October 2003.

[Moon and Aghvami 2004] Moon B. and A. H. Aghvami, "Quality of Service Mechanisms in All-IP Wireless Access Networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp.873-888, June 2004.

[Moore 2005] Moore N., "Optimistic Duplicate Address Detection for IPv6," *IETF Internet draft*, <draft-ietf-ipv6-optimistic-dad-05.txt>, work in progress, February 2005.

[Moskowitz 2005] Moskowitz R., P. Nikander, P. Jokela and T. Henderson, "Host Identity Protocol," *IETF Internet Draft*, draft-ietf-hip-base-03, Work in Progress, June 2005.

[MPEG] The MPEG official website, http://www.chiariglione.org/mpeg/

[Murakami etc 2004] Murakami K., O. Haase, J. Shin, and T. F. La Porta, "Mobility Management Alternatives for the Migration to Mobile Internet Session-based Services," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 818-832, Jun 2004.

N

[Nakajima etc 2003] Nakajima N., A. Dutta, S. Das, and H. Schulzrinne, "Handoff Delay Analysis and Measurement for SIP Based Mobility in IPv6," in *Proc. of IEEE ICC'03*, pp. 1085-1089, Anchorage, Alaska, USA, May 2003.

O

[Oouchi etc 2002] Oouchi H., T. Takenaga, H. Sugawara, and M. Masugi, "Study on Appropriate Voice Data Length of IP Packets for VoIP Network Adjustment," in *Proc. of IEEE GLOBECOM'02*, pp. 1628-1632, Taipei, China, November 2002.

[OPNET] OPNET, http://www.opnet.com

P

[Pack and Choi 2004] Pack S. 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.

[Paskalis etc 2003] Paskalis S., A. Kaloxylos, E. Zervas, and L. Merakos, "An Efficient RSVP-Mobile IP Interworking Scheme," *ACM Mobile Networks and Applications*, vol. 8, no. 3, pp.197-207, June 2003.

[Patel etc 2004] Patel A., K. Leung, H. Akhtar, M. Khalil and K. Chowdhury, "Network Access Identifier Option for Mobile IPv6," *IETF Internet Draft*, draft-ietf-mip6-nai-option-00.txt, work in progress, July 2004.

[Patel and Dennett 2000] Patel G. and S. Dennett, "The 3GPP and 3GPP2 Movements toward an All-IP Mobile Network," *IEEE Personal Communications*, vol. 7, no. 4, pp. 62-64, August 2000.

[Perkins and Johnson 2001] Perkins C. and D. B. Johnson, "Route Optimisation in Mobile IP," *IETF Internet Draft*, draft-ietf-mobileip-optim-11.txt, work in progress, September 2001.

[Politis etc 2003] Politis C., K. Chew, and R. Tafazolli, "Multilayer Mobility Management for All-IP Networks: Pure SIP vs. Hybrid SIP/Mobile IP," in *Proc. of IEEE VTC2003-Spring*, Jeju, Korea, April 2003.

[Politis etc 2004] Politis C., K. A. Chew, N. Akhtar, M. Georgiades, R. Tafazolli, and T. Dagiuklas, "Hybrid Multilayer Mobility Management with AAA Context Transfer Capabilities for All-IP Networks," *IEEE Wireless Communications*, vol. 11, no. 4, pp. 76-88, August 2004.

[Pollini etc 1995] Pollini G. P., K. S. Meier-Heustern, and D. J. Goodman, "Signalling Traffic Volume Generated by Mobile and Personal Communications,"

IEEE Communications Magazine, vol. 33, no. 6, pp. 60-65, June 1995.

[Pous etc 2004] Pous M. I., D. Pesch, and G. Foster, "SIP-Based Applications in UMTS: A Performance Analysis," in *Proc. of EW'04*, Barcelona, Spain, April 2004.

R

[Ramjee etc 2002A] Ramjee R., K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang, and T. F. La Porta, "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-Area Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396-410, June 2002.

[Ramjee etc 2002B] Ramjee R., L. Li, T. La Porta, and S. Kasera, "IP Paging Service for Mobile Hosts," *ACM Wireless Networks (WINET)*, vol. 8, no. 5, pp. 427–441, September 2002.

[RFC768] Postel J., "User Datagram Protocol," *IETF RFC 768*, September 1981.

[RFC791] Postel J. (ed.), "Internet Protocol," *IETF RFC 791*, September 1981.

[RFC792] Postel J., "Internet Control Message Protocol," *IETF RFC 792*, September 1981.

[RFC793] Postel J. (ed.), "Transmission Control Protocol," *IETF RFC 793*, September 1981.

[RFC821] Postel J., "Simple Mail Transfer Protocol," *IETF RFC 821*, August 1982.

[RFC854] Postel J. and J. Reynolds, "Telnet Protocol Specification," *IETF RFC 854*, October 1983.

[RFC959] Postel J. and J. Reynolds, "File Transfer Protocol (FTP)," *IETF RFC 959*, October 1985.

[RFC1663] Braden R., D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: An Overview," *IETF RFC 1633*, June 1994.

[RFC1885] Conta A. and S. Deering, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6)," *IETF RFC 1885*, December 1995.

[RFC2131] Droms R., "Dynamic Host Configuration Protocol," *IETF RFC 2131*, 1997.

[RFC2205] Braden, R., L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) - Version 1, Functional Specification", *IETF RFC 2205*, September 1997.

[RFC2210] Wroclawski J., "The Use of RSVP with IETF Integrated Services," *IETF RFC 2210*, September 1997.

[RFC2327] Handley M.and V. Jacobson, "SDP: Session Description Protocol," *IETF RFC 2327*, April 1998.

[RFC2406] Kent S. and R. Atkinson, "IP Encapsulating Security Payload (ESP)," *IETF RFC 2406*, November 1998.

[RFC2460] Deering S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification," *IETF RFC 2460*, December 1998.

[RFC2461] Narten T., E. Nordmark, and W. Simpson, "Neighbour Discovery for IP Version 6 (IPv6)," *IETF RFC 2461*, December 1998.

[RFC2462] Thomson S. and T. Narten, "IPv6 Stateless Address Autoconfiguration," *IETF RFC 2462*, December 1998.

[RFC2474] Nichols K., S. Blake, F. Baker, and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers," *IEFT RFC 2474*, 1998.

[RFC2475] Blake S., D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *IETF RFC 2475*, December 1998.

[RFC2486] Aboba B., and M. Beadles, "The Network Access Identifier," *IETF RFC 2486*, January 1999.

[RFC2616] Fielding R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1," *IETF RFC 2616*, June 1999.

[RFC2794] Calhoun P. and C. Perkins, "Mobile IP Network Access Identifier Extension for IPv4," *IETF RFC 2794*, March 2000.

[RFC2865] Rigney C., S. Willens, A. Rubens, and W. Simpson, "Remote Authentication Dial in User Service (RADIUS)," *IETF RFC 2865*, June 2000.

[RFC2960] Stewart R., Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson, "Stream Control Transmission Protocol," *IETF RFC 2960*, October 2000.

[RFC2976] Donovan S., "The SIP INFO Method," *IETF RFC 2976*, October 2000.

[RFC2998] Bernet Y., P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, and et al., "A Framework for Integrated Services Operation over DiffServ Networks," *IETF RFC 2998*, November 2000.

[RFC3113] Rosenbrock K., R. Sanmugam, S. Bradner, and J. Klensin, "3GPP-IETF Standardisation Collaboration," *IETF RFC 3113*, June 2001.

[RFC3141] Hiller T., P. Walsh, X. Chen, M. Munson, G. Dommety, S. Sivalingham, and et al, "CDMA2000 Wireless Data Requirements for AAA," *IETF RFC 3141*, June 2001.

[RFC3175] Baker F., C. Iturralde, F. Le Faucheur, and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 Reservations," *IETF RFC 3175*, September 2001.

[RFC3261] Rosenberg J., H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks and et al. "SIP: Session Initiation Protocol," *IETF RFC 3261*, June 2002.

[RFC3263] Rosenberg J. and H. Schulzrinne, "SIP: Locating SIP Servers," *IETF RFC 3263*, June 2002.

[RFC3312] Camarillo G., W. Marchshall, and J. Rosenberg, "Integration of Resource Management and Session Initiation Protocol (SIP)," *IETF RFC 3312*, October 2002.

[RFC3315] Droms R., J. Bound, B. Volz, T. Lemon, C. Perkins, and M.Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)," *IETF RFC 3315*, July 2003.

[RFC3319] Schulzrinne H. and B. Volz, "Dynamic Host Configuration Protocol (DHCPv6) Options for Session Initiation Protocol (SIP) Servers," *IETF RFC 3319*, July 2003.

[RFC3344] Perkins C., "IP Mobility Support for IPv4," *IETF RFC 3344*, August 2002.

[RFC3486] Camarillo G., "Compressing the Session Initiation Protocol (SIP)," *IETF RFC 3486*, February 2003.

[RFC3515] Sparks R., "The Session Initiation Protocol (SIP) Refer Method," *IETF RFC 3515*, April 2003.

[RFC3550] Schulzrinne H., S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," *IETF RFC 3550*, July 2003.

[RFC3583] Chaskar H. (ed.), "Requirements of a Quality of Service (QoS) Solution

for Mobile IP," *IETF RFC 3583*, September 2003.

[RFC3588] Calhoun P., J. Loughney, E. Guttman, G. Zorn, and J. Arkko, "Diameter Base Protocol," *IETF RFC 3588*, September 2003.

[RFC3775] Johnson D., C. Perkins, and J. Arkko, "Mobility Support in IPv6," *IETF RFC 3775*, June 2004.

[RFC3776] Arkko J., V. Devarapalli, and F. Dupont, "Using IPsec to Protect Mobile IPv6 Signalling Between Mobile Nodes and Home Agents," *IETF RFC 3776*, June 2004.

[RFC3819] Karn P., C. Bormann, G. Fairhurst, D. Grossman, R Ludwig, J. Mahdavi, and et al, "Advice for Internet Subnetwork Designers," *IETF RFC 3819*, July 2004.

[RFC3828] Larzon L-Å, M. Degermark, S. Pink, L.-E. Jonsson, and G. Fairhurst, "The Lightweight User Datagram Protocol (UDP-Lite)," *IETF RFC 3828*, July 2004.

[RFC3963] Devarapalli V., R. Wakikawa, A. Petrescu, and P. Thubert, "Network Mobility (NEMO) Basic Support Protocol," *IETF RFC 3963*, January 2005.

[RFC4004] Calhoun P., T. Johansson, C. Perkins, T. Hiller, and P. McCann, "Diameter Mobile IPv4 Application," *IETF RFC 4004*, August 2005.

[RFC4068] Koodli R. (ed.), "Fast Handovers for Mobile IPv6," *IETF RFC 4068*, July 2005.

[RFC4140] Soliman H., C. Catelluccia,K. E. Malki, and Ludovic Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)," *IETF RFC 4140*, August 2005.

[RFC4301] Kent S. and K. Seo, "Security Architecture for the Internet Protocol," *IETF RFC 4301*, December 2005.

[Royer and Toh 1999] Royer E. M. and C.-K. Toh, "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," *IEEE Personal Communications*, vol. 6, no. 2, pp. 46-55, April 1999.

S

[Sami 2003] Sami U., "Key Concepts for Evolution toward Beyond 3G Networks," *IEEE Wireless Communications*, vol. 10, no. 1, pp. 43-48, February 2003.

[Schulzrinne and Wedlund 2000] Schulzrinne H. and E. Wedlund, "Application-Layer Mobility using SIP," *ACM Mobile Computing and Communications Review*, vol. 4, no. 3, July 2000, pp. 47-57.

[Schulzrinne and Arabshian 2002] Schulzrinne H. and K. Arabshian, "Providing emergency services in Internet telephony," IEEE Internet Computing, vol. 6, no. 3, pp. 39-47, May/June 2002.

[Schulzrinne 2006] H. Schulzrinne, "Emergency Services URI for the Session Initiation Protocol," *IETF Internet Draft*, <draft-ietf-sipping-sos-02.txt>, work in progress, January 2006.

[Sen etc 1999] Sen S. K., A. Bhattacharya, and S. K. Das, "A Selective Location Update Strategy For PCS Users," *Wireless Networks*, vol. 5, no.5, pp. 313–326, September 1999.

[Seshan 1995] Seshan S., "Low-Latency Handoff for Cellular Data Networks," *Ph.D. Thesis*, the University of California at Berkeley, USA, 1995.

[Shelby 2001] Shelby Z., D. Gatzounas, A. Campbell, and C.-Y. Wan, "Cellular IPv6," *IETF Internet Draft*, <draft-shelby-cellularipv6-01.txt>, work in progress, July 2001.

[Snoeren and Balakrishnan 2000] Snoeren A. C. and H. Balakrishnan, "An End-to-End Approach to Host Mobility," in *Proc. of* ACM/IEEE MobiCom'00, Boston, USA, August 2000.

[Stewart etc 2005] Stewart R., M. Ramalho, Q. Xie, M. Tuexen, and P. Conrad, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration," *IETF Internet Draft*, draft-ietf-tsvwg-addip-sctp-12.txt, work in progress, June 2005.

[Sudame and B. Badrinath 2001] Sudame P. and B. R. Badrinath, "On Providing Support For Protocol Adaptation in Mobile Wireless Networks," *ACM Mobile Networks and Applications*, vol. 6, no. 1, pp. 43-55, January/February 2001.

[Swami etc 2005] Swami Y.P., K. Le and W. Eddy, "Lightweight Mobility Detection and Response (LMDR) Algorithm for TCP," *IETF Internet Draft*, draft-swami-tcp-lmdr-06.txt, work in progress, August 2005.

T

[Tabbane 1998] Tabbane S., "Modelling the MSC/VLR Processing Load due to Mobility Management," in *Proc. of IEEE ICUPC'98*, pp. 741-744, Florence, Italy, October 1998.

[Taha etc 2005] Taha A. M., H. S. Hassanein and H. T. Mouftah, "Extensions for Internet QoS Paradigms to Mobile IP: A Survey," *IEEE Communications Magazine*, vol. 43, no. 5, pp. 132-139, May 2005.

[Tanenbaum 1996] Tanenbaum A. S., *Computer Networks*, Third Edition, London: Prentice Hall, 1996.

[Tripathi etc 1998] Tripathi N. D., J. H. Reed, and H. F. VanLandinoham, "Handoff in Cellular Systems," *IEEE Personal Communications*, vol.5, no.6, pp. 26-37, December 1998.

[Tseng and Shen 2003] Tseng Y.-C. and C.-C. Shen, "Integrating Mobile IP with Ad Hoc Networks," *IEEE Computer*, vol. 36, no. 5, pp. 48-55, May 2003.

V

[Vakil etc 2001] Vakil F., A. Dutta, J.-C. Chen, M. Tauil, S. Baba, N. Nakajima, and H. Schulzrinne, "Supporting Mobility for TCP with SIP," *IETF Internet Draft*, draf-itsumo-sipping-mobility-tcp-00.txt, work in progress, June 2001.

[Vatn and Maguire 1998] Vatn J.-O. and G. Q. Jr. Maguire, "The Effect of Using Co-Located Care-of Addresses on Macro Handover Delay," in *Proc. of the 14th Nordic Teletraffic Seminar*, Lyngby, Denmark, August 1998.

[Vidales etc 2004A] Vidales P., R. Chakravorty, and C. Policroniades, "PROTON: A Policy-Based Solution for Future 4G devices," In *Proc. of 5th International Workshop on Policies for Distributed Systems and Networks (IEEE POLICY2004)*, New York, USA, June 2004.

[Vidales etc 2004B] Vidales P., L. Patanapongpibul, G. Mapp, and A. Hopper, "Experiences with Heterogeneous Wireless Networks – Unveiling the Challenges," in *Proc. of 2nd International Working Conference on Performance*

Modelling and Evaluation of Heterogeneous Networks (HET-NETs), Hilkley, UK, July 2004.

W

[Walke 2002] Walke B. H., *Mobile Radio Networks: Networking, Protocols and Traffic Performance*, 2nd Ed., John Wiley & Sons, Inc., Chichester, UK, 2002.

[Wang and Katz 2001] Wang H. J. and R. H. Katz, "Mobility Support in Unified Communication Networks," in *Proc. of Workshop on Wireless Mobile Multimedia*, Rome, Italy, July 2001.

[Wang etc 2000] Wang H. J., B. Raman, C-N. Chuah, R. Biswas, R. Gummadi, B. Hohlt, X. Hong, E. Kiciman, Z. Mao, J. S. Shih, L. Subramanian, B. Y. Zhao, A. D. Joseph, and R. H. Katz, "ICEBERG: An Internet Core Network Architecture for Integrated Communications," *IEEE Personal Communications*, vol. 7, no. 4, pp. 10-19, August 2000.

[Wang etc 1999] Wang H. J., R. H. Katz, and J. Giese, "Policy-Enabled Handoffs Across Heterogeneous Wireless Networks," in *Proc. of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA '99)*, New Orleans, LA, USA, February 1999.

[Wang and Akyildiz 2001] Wang W., I. F. Akyildiz, "A New Signalling Protocol for Intersystem Roaming in Next-Generation Wireless Systems," *IEEE Journal on Selected Areas in Communications,* vol. 19, no. 10, pp. 2040-2052, October 2001.

[Wedlund and Schulzrinne 1999] Wedlund E. and H. Schulzrinne, "Mobility Support using SIP," in *Proc. of 2nd ACM/IEEE WoWMoM'99*, Seattle, USA, August 1999.

[Willmann and Kuhn 1990] Willmann G. and P. J. Kuhn, "Performance Modelling of Signalling System No.7," *IEEE Communications Magazine*, vol. 28, no. 8, pp. 44-56, July 1990.

[Wong and Lim 1997] Wong D. and T. J. Lim, "Soft Handoffs in CDMA Mobile Systems," *IEEE Personal Communications*, vol. 4, no. 6, pp. 6–17, December 1997.

[Wong etc 2003] Wong K. D., A. Dutta, J. Burns, R. Jain, K. Young, and H. Schulzrinne, "A Multilayered Mobility Management Scheme for Auto-Configured Wireless IP Networks," *IEEE Wireless Communications*, vol. 10, no. 5, pp. 62-69, October 2003.

[Wong etc 2001] Wong W. S. V., M. Lewis, and V. Leung, "Stochastic Control of Path Optimization for Inter-Switch Handoffs in Wireless ATM Networks," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, pp. 336-350, June 2001.

[Wong etc 2000A] Wong W. S. V., H. Chan, and V. Leung, "Performance Evaluations of Path Optimization Schemes for Inter-Switch Handoff in Wireless ATM Networks," *Wireless Networks*, vol. 6, no. 4, pp. 251-262, July 2000.

[Wong etc 2000B] Wong W. S. V. and V. C. M. Leung, "Location Management for Next-Generation Personal Communications Networks," *IEEE Network*, vol. 14, no. 5, pp. 18-24, September/October 2000.

[Wisley etc 2002] Wisley D., P. Eardley, and L. Burness, *IP for 3G: Networking Technologies for Mobile Communications*, John Wiley & Sons, Chichester, UK, 2002.

[Wu etc 1999] Wu G., Y. Bai, J. Lai, and A. Ogielski, "Interactions between TCP and RLP in Wireless Internet," in *Proc. of IEEE GLOBECOM'99*, pp. 661-666, Rio de Janeiro, Brazil, December 1999.

[Wu etc 2002] Wu W., S. K. Das, A. Misra, and S. Das, "Performance Evaluation of IDMP's QoS Framework," in *Proc. of IEEE GLOBECOM'02*, pp. 2515-2519, Taiwan, China, November 2002.

Z

[Zhang etc 2002] Zhang X., J. G. Castellanos, and A. T. Campbell, "P-MIP: Paging Extensions for Mobile IP," *ACM Mobile Networks and Applications (MONET)*, vol. 7, No. 2, pp. 127-141, April 2002.